

MacDonald, Connor Rees Murray (2026) *Developing quantitative structure-property relationships for predicting properties of self-assembled materials.* PhD thesis.

https://theses.gla.ac.uk/86347/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk
research-enlighten@glasgow.ac.uk



University of Glasgow

Developing Quantitative Structure-Property Relationships for Predicting Properties of Self-Assembled Materials

Connor Rees Murray MacDonald

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

School of Chemistry
College of Science and Engineering
University of Glasgow

May 2026

Declaration of Authorship

This Thesis has been written and compiled by the author, Connor MacDonald, who carried out the research at the University of Glasgow between 2022 and 2026 under the supervision of Professor Emily Draper. I declare that, except where explicit reference is made to the contributions of others, this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Connor MacDonald

Abstract

Smart materials are increasingly utilised across many innovative technologies. As these applications become increasingly complex, so too must the materials that facilitate them. Supramolecular materials are of particular interest due to their functional, tuneable nature and their derivability from sustainable feedstocks. While this tunability affords significant opportunity for highly tailored material design from a small library of molecules, the variability in resulting bulk properties makes predictability an ongoing challenge. In this Thesis, we report the development of quantitative structure-property relationships (QSPRs) designed to bridge the gap in predictive ability, enabling a more targeted investigation of material candidates.

First, we introduce the concepts central to developing our quantitative structure-property relationship framework and show how it complies with the Organisation for Economic Co-operation and Development (OECD) guidelines for effective QSPR development. By benchmarking the predictive performance of a model trained on molecular structure descriptors and an IC_{50} dataset, we show how our system can target a defined endpoint with an unambiguous algorithm, as assessed using appropriate metrics. Predictions can be made by identifying those with sufficient chemical similarity to afford reliable predictions. Lastly, we show how the model can be probed to identify the key chemistries that govern the models it yields.

We go on to demonstrate how our QSPR framework can be adapted to investigate the solubility of amino acid functionalised naphthalene and rylene derivatives in organic solvents. We show how investigating physicochemical and structural descriptors alone is insufficient to elucidate the complexity governing solubility. We demonstrate that QSPRs are an effective strategy for establishing useful relationships between molecular descriptors and this molecular-scale property. Notably, we can highlight the limitations of the current model iteration, providing insight into how future models could be significantly improved.

We then demonstrate how such systems can provide insight into more complex interactions at the nano- and microscale. Using a synchrotron small angle X-ray scattering instrument, a high-quality dataset of structural information was obtained. We discuss the complexity of data fitting and instead show how a simpler quantification of scattering quality can provide insight into the presence of well-defined micellar structures in each solution. The resulting QSPR models demonstrate the ability of QSPR modelling to utilise descriptors of both the

solute and the solvent to afford a system capable of incorporating descriptors of chemical constitution and material preparation.

We highlight how the systems developed across this project can be used in tandem to reduce the pool of all candidate materials to those most likely to afford useful aggregated structures.

We conclude with a summary of the achievements of this thesis, limitations that need to be addressed in future systems, and current limitations of literature reporting in the supramolecular community, discussing the need for increased standardisation across the community if more insightful QSPRs are to be developed.

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Emily Draper, who has been a constant source of support and encouragement throughout my PhD. Your supervision and insight across these projects, as well as the care you have for the group, have made the last three and a half years a truly joyous experience. I am so very grateful for all the opportunities you have given me throughout this project, and for everything you do, both inside and outside the lab, to make the Draper group an honour to be a part of.

I'd like to thank everyone in the Draper group, past and present. My time here would not have been the same without the love and support you all provide, which makes every day at the university a privilege. To Rebecca, thank you for being a constant source of joy and support. You are truly one of the biggest inspirations in my life. I am so lucky to continue to see you shine, and to have had you here to cheer me on every step of the way. To Nick, there are few people who can spread laughter the way you do. Every conversation was always funnier than the last. You make every room you're in a brighter place. To Ziyao, I'm so glad to have shared this journey with you. You are a constant beacon of positivity, and I will miss sharing what comes next with you. It's been amazing to see you thrive and live your best life! I hope you've enjoyed your time in Scotland as much as we have been blessed to have you. To Lineta, I wish I could bring at least a fraction of the positive energy you bring to the group. Thank you for being the source of zen I needed during scattering and while I wrote this thesis. To Ellen, please never let your infectious excitement end. It makes everything you're a part of "P E R F E C T!!!". To Matthew, thank you for teaching me about the world. I don't know if I'd have finished my thesis without our daily games to keep me going, or if I'd have known where Poland is.

I would also like to thank the Adams group for making the Raphael lab such a welcoming place. It truly is a blessing to be surrounded by such welcoming and supportive people each day. To Emma, I'm so lucky to have had you by my side for this journey. Your kindness and joy are unparalleled, and this would not have been the same without you. I am so excited to see you go on to inspire others in chemistry, as you have for me.

I would also like to thank Dr Bart Dietrich, Professor Neil Berry, and Dr Christopher Woodley for their support across this project. I'd like to thank Chris for the significant guidance he offered at the outset of this project. Without your keen knowledge of QSPR and willingness to share your time, this project would not be what it has become.

I'd like to thank everyone at Charles River who supported my growth as a scientist. I miss OA09 greatly, and I'm so lucky to have been part of a team filled with such lovely people that make me still miss my time there several years on. I'd like to thank my study directors, Josh Kennedy and Heather Wicksted, whose constant support and guidance during my time there gave me the confidence to take on this PhD. Whether it's waylaying stresses with dog photos and traybakes, or *via* a cracking penalty to the face, you both made it a pleasure to arrive on site each day. To Anna, I miss getting to share our crippling laughs at nothing in particular each day.

To my friends who have supported me throughout, it is amazing to have had you all to share the excitement (and stress) of this part of my life. Every big moment and every little laugh we have shared during my PhD have made every day so much easier.

To Ryan, Shona, Heather, and Douglas, my life here in Glasgow would not be the same without you all. While we never know what life has in store for us, I know I'll always have your love and support. Thank you to Zac, Sammi, Jamie, Jess, John, Ivanna, Sean, Findlay, and Felix; you have been some of my biggest cheerleaders. Words can't say how lucky I am to call you all friends. Your support has carried me further than you know. And thank you to Andrew, Mitchell, Gordon, Matthew, Alex, Alasdair, and Ryan (again), whom I can always trust to be weird little goblins that let me forget the stresses of the week.

Finally, thank you to my family for your constant support during every tough moment and every celebration for each little win. I can't express how much I love you all. To my parents, Jacqueline and Raymond, thank you for the constant encouragement throughout my PhD, every ounce of support you gave me to get me there, and the love you give every day. I would not be where I am today without you both; everything I've achieved is thanks to you.

List of Publications from This Degree

1. **C. R. M. MacDonald** and E. R. Draper, Applications of microscopy and small-angle scattering techniques for the characterisation of supramolecular gels, *Beilstein Journal of Organic Chemistry*, 2024, **20**, 2608-2634.

2. **C. R. M. MacDonald**, Alex S. Loch, Neil G.Berry, and E. R. Draper, An accessible property classification framework to predict the solubility of functionalised naphthalenes and rylenes in organic solvents. Digital Discovery, 2026, https://doi.org/10.1039/d6dd00275g

Table of Contents

Prelims

Declaration of Authorship ii
Abstract iii
Acknowledgements iv
List of Publications from This Degree vii
Table of Contents viii
List of Abbreviations xiv

Chapter 1: Introduction 1
1.1 Smart Materials 2
1.1.1 Applications of Smart Materials 2
1.1.2 Existing Smart Material Systems 5
1.1.2.1 Inorganic Materials 5
1.1.2.2 Polymers 7
1.2 Small Molecules 10
1.2.1 Small Molecule Self-Assembly 11
1.2.2 Supramolecular Interactions 13
1.2.3 Tuning Self-Assembly 17
1.2.4 Rylene Self-Assembly 22
1.2.5 Tunability of Rylene Materials 24
1.3 Predicting Self-Assembly 28
1.3.1 Existing Techniques 28
1.3.1.1 Critical Packing Parameter 28
1.3.1.2 Quantum Mechanical Methods 29
1.3.1.3 Molecular Dynamics Methods 31
1.3.1.4 Statistical Thermodynamics Methods 33
1.3.1.5 Limitations of Current Methods 34
1.3.2 Machine Learning 35
1.3.2.1 Machine Learning in Material Chemistry 38
1.4 Quantitative Structure Property Relationships 43
1.4.1 Fundamental Principles of QSPRs 43
1.4.2 QSPRs in Supramolecular Applications 45
1.5 Aims of this Thesis 49

1.6 References 50

Chapter 2: Developing a quantitative structure-property relationship framework in KNIME 57

2.1 Introduction 58
2.2 Dataset Acquisition 60
2.2.1 Experimental Data 60
2.2.2 Literature Mining 61
2.3 Machine Learning Platforms 64
2.3.1 Python 64
2.3.2 Visual Programming 64
2.4 Dataset Pre-processing 67
2.4.1 Data Curation 68
2.4.2 Molecular Descriptors 70
2.4.2.1 Fingerprint Descriptors 71
2.4.3 Modelability Index 72
2.4.4 Applicability Domain 74
2.4.5 Dataset Preparation 75
2.4.6 Class Balancing 78
2.4.6.1 Undersampling 79
2.4.6.2 Oversampling 80
2.5 Machine Learning Model Training 81
2.5.1 K-Nearest Neighbours 82
2.5.2 Decision Trees 83
2.5.2.1 Ensemble Learning 84
2.5.2.2 Gradient Boosted Learning 85
2.5.3 Support Vector Machines 86
2.5.4 Naïve Bayes 87
2.5.5 Logistic Regression 88
2.5.6 Hyperparameter Optimisation 89
2.5.7 Cross Validation 91
2.6 Model Validation 92
2.6.1 Performance Metrics 92
2.6.2 Model Robustness 93
2.6.3 External Validation 95

2.7 Model Visualisation 96
2.7.1 Descriptor Importance 96
2.7.2 Substructure Identification 98
2.8 PLA2-IIA Model Performance 100
2.9 Conclusion 104
2.10 References 105

Chapter 3: Predictive classification of organic solubility of functionalised naphthalenes and rylenes 110
3.1 Introduction 111
3.2 Results and Discussion 113
3.2.1 Existing Library 113
3.2.2 Literature Datasets 114
3.2.3 Expanding the Dataset 116
3.2.3.1 Preparing Candidate Structures 116
3.2.3.2 Optimising the Training Library 118
3.2.3.3 Synthesis of Target Molecules 120
3.2.3.4 Introducing Similar Molecular Classes 121
3.2.4 Characterising Solubility 123
3.2.5 QSPR Data Preprocessing 128
3.2.6 Machine Learning Model Training 132
3.2.7 Applicability Domain 135
3.2.8 Model Robustness 137
3.2.9 Model Validation 139
3.2.10 Feature Importance 140
3.3 Conclusions 142
3.4 Experimental 144
3.4.1 Materials and Synthetic Procedures 144
3.4.2 Representative Synthesis of Protected Dipeptides 144
3.4.3 Representative Deprotection of Peptides 146
3.4.3.1 Representative Trifluoroacetic Acid Deprotection 146
3.4.3.2 Representative Hydrochloric Acid Deprotection 148
3.4.4 Representative Synthesis of Imide Functionalised Rylenes 149
3.4.5 Synthesis of Validation Molecules 151

3.4.6 Determination of Molar Absorption Coefficient 154
3.4.7 Determination of Organic Solubility 155
3.5 References 156

Chapter 4: Predictive identification of self-assembled aggregates in naphthalene and rylene solutions 159
4.1 Introduction 160
4.1.1 Characterising Self-assembly 164
4.1.2 Microscopy Techniques 164
4.1.3 Scattering Techniques 165
4.1.4 Literature Data 167
4.1.5 Characterisation Techniques 169
4.1.6 Small Angle X-ray Scattering 170
4.1.7 SAXS Model Complexity 171
4.2 Results and Discussion 174
4.2.1 Data Collection 174
4.2.2 Scattering Quality 177
4.2.3 Data Fitting 181
4.2.4 QSPR Data Preprocessing 185
4.2.5 Machine Learning Model Training 190
4.2.6 Applicability Domain 192
4.2.7 Model Robustness 194
4.2.8 Model Validation 196
4.2.9 Feature Importance 199
4.3 Conclusions 203
4.4 Experimental 204
4.4.1 Materials 204
4.4.2 Solution Preparation 204
4.4.3 Small Angle X-ray Scattering 204
4.5 References 206

Chapter 5: Conclusions and Future Work 209
5.1 Conclusions 210
5.2 Future Work 213

Appendices **I**

Chapter 3 Appendices **II**

A 3.1 Synthetic Characterisations II

A 3.1.1 Dipeptide Synthesis Boc Reagents II

A 3.1.2 Dipeptide Synthesis OtBu Reagents III

A 3.1.3 Protected Dipeptide Products IV

A 3.1.4 Deprotected Dipeptide Products XIII

A 3.1.5 Rylene Precursor Reagents XX

A 3.1.6 Rylene Products XXI

A 3.1.7 Validation Products and Precursors XXII

A 3.2 NMR and HRMS Spectra XXVI

A 3.2.1 Protected Dipeptide Products XXVI

A 3.2.2 Deprotected Dipeptide Products LXXVIII

A 3.2.3 Rylene Products CXXIV

A 3.2.4 Validation Products and Precursors CXXX

A 3.3 Model Compounds CXL

A 3.4 Calibration Data CXLII

A 3.5 Solubility Data CLI

A 3.6 Classification Model Data Summary CLXIII

A 3.7 Physicochemical Descriptor Distribution CLXIV

A 3.8 Fingerprint Descriptor Distribution CLXIX

A 3.9 Partitioning Impact Assessment CLXXIV

A 3.10 Model Performance Metrics CLXXV

A 3.11 Receiver Operator Characteristic Curves CLXXVIII

A 3.12 Feature Importance CLXXIX

Chapter 4 Appendices **CLXXXIV**

A 4.1 Model Molecules CLXXXIV

A 4.2 Scattering Classification CLXXXV

A 4.2.1 5 mgmL-1 Aqueous Solution Classifications CLXXXV

A 4.2.2 10 mgmL-1 Aqueous Solution Classifications CXC

A 4.2.3 Acetone Solution Classifications CXCIV

A 4.2.4 DMF Solution Classifications CXCVI

A 4.2.5 DMSO Solution Classifications CC

A 4.2.6 Ethanol Solution Classifications CCIV

A 4.2.7 THF Solution Classifications CCVI

A 4.2.8 Deuterated Solution Classifications CCIX

A 4.2.8.1 5 mgmL-1 Aqueous Solution Classifications CCIX

A 4.2.8.2 10 mgmL-1 Aqueous Solution Classifications CCXI

A 4.2.8.3 d6-DMSO Solution Classifications CCXIII

A 4.2.8.4 d6-Ethanol Solution Classifications CCXIV

A 4.2.9 Multicomponent Solution Classifications CCXV

A 4.3 Classification Summaries CCLIV

A 4.3.1 Single Component Summary CCLIV

A 4.3.2 Multicomponent Scattering Classification CCLVI

A 4.4 Solvent Isotope Scattering Comparisons CCLVIII

A 4.4.1 Example across Solvents CCLVIII

A 4.4.2 Examples in H2O and D2O CCLXII

A 4.4.3 Examples in DMSO and d6-DMSO CCLXIV

A 4.4.4 Examples in Ethanol and d6-Ethanol CCLXVI

A 4.5 Physicochemical Descriptor Distribution CCLXVII

A 4.6 Solvent Descriptor Distribution CCLXX

A 4.7 Fingerprint Descriptor Distribution CCLXXI

A 4.8 Model Performance Metrics CCLXXVII

A 4.9 Partitioning Impact Assessment CCLXXX

A 4.10 Receiver Operator Characteristic Curves CCLXXXI

A 4.11 Validation Predictions CCLXXXII

A 4.12 Feature Importance CCLXXXVI

List of Abbreviations

μ	Dipole moment
0D	Zero-dimensional
1D	One-dimensional
2D	Two-dimensional
2θ	Scattering angle
3D	Three-dimensional
A	Absorption
a_0	Head group area
AD	Applicability domain
ADASYN	Adaptive synthetic sampling
AI	Artaficial intelligence
AMS	Amsterdam modelling suite
AP	Aggregation propensity
AUC	Area under the curve
B	Bad
C	Chemical constitution
c	Concentration
C4.5	C4.5 ensemble decision trees
CMC	Critical micelle concentration
COSMO-RS	Conductor-like screening model for liquid systems
CPP	Critical packing parameter
CSV	Comma-separated value
d	Average distance
DAWN	Data analysis workbench
DMF	Dimethyl formamide
DMSO	Dimethyl sulfoxide
DPD	Dissipative particle dynamics
DSC	Differential scanning calorimetry
dsz	d+(s×0.5)
EM	Electron microscopy
f	Function
FN	False negative
FP	False positive
G	Good
G′	Storage modulus
G″	Loss modulus
GHG	Greenhouse gas
GNN	Graph neural network
HPLC	High-performance liquid chromatography
HSP	Hansen solubility parameters
I(q)	Scattering intensity
IC_{50}	Inhibitory concentration

J_C	Intermolecular coulombic coupling
k	Number of nearest neighbours
KNIME	Konstanz information miner
kNN	k-nearest neighbours
l	Pathlength
l_0	Chain length
LLM	Large language model
LOOCV	Leave-one-out cross-validation
LR	Logistic regression
maxdist	Maximum distance
MCC	Matthew's correlation coefficient
MD	Molecular dynamics
MDP	Markov decision process
ML	Machine learning
MODI	Modelability index
MOLT	Molecular theory
Nap	Naphthoyloxyacetyl
NB	Naive Bayes
NDI	Napthalene diimide
NMR	Nuclear magnetic resonance
OECD	Organisation for economic co-operation and development
p95	95th percentile
PBI	Perylene bisimide
PCA	Principle component analysis
PEDOT:PSS	Poly(3,4-ethylenedioxythiophene):poly(styrenesulfonate)
PLA2-IIA	Phospholipase A2 group IIA
PyFAI	Fast azimuthal integration python library
PZT	Lead zirconate titanate
q	Scattering vector
QM	Quantum mechanical
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
R	Intermolecular distance
R^2	Coefficient of determination
RBF	Radial basis function
RF	Random forest
RL	Reinforcement learning
ROC	Receiver operating characteristic
s	Standard deviation
SANS	Small angle neutron scattering
SASA	Solvent accessible surface area
SAXS	Small angle X-ray scattering
SDF	Structure data file
SEM	Scanning electron microscopy
SHAP	Shapley additive explanations

SL	Supervised learning
SLD	Scattering length density
SMILES	Simplified molecular input line entry system
SMOTE	Synthetic minority over-sampling technique
SNR	Signal-to-noise ratio
SVM	Support vector machine
T_C	Tanimoto coefficient
TEM	Transmission electron microscopy
THF	Tetrahydrofuran
TMAP	Tree MAP
TN	True negative
TP	True positive
UL	Unsupervised learning
UV-vis	Ultraviolet-visible
V	Volume of a surfactant
XG	XGBoost
Zc	Critical Z-score
α	Level of significance
ε	Molar absorption coefficient
ε_r	Dielectric constant
θ	Intermolecular angle
θ_M	Magic Angle
κ	Cohen's kappa
λ	Wavelength
λ_{max}	Wavelength of maximum absorption
σ	Complexity parameter
ϕ	Physiological activity

Chapter 1: Introduction

This Chapter is adapted, in part, from the following publication:

"Applications of microscopy and small-angle scattering techniques for the characterisation of supramolecular gels"

C. R. M. MacDonald and E. R. Draper, *Beilstein J. Org. Chem,* 2024, **20**, 2608 – 2634

CRMM was responsible for researching, creating figures, and writing the publication. CRMM wrote the initial draft of the manuscript, to which both authors contributed for the final publication

1.1 Smart Materials

1.1.1 Applications of Smart Materials

As new technologies emerge, there is a growing need to design materials that will improve or enable their application. Over the past two decades, significant advances have been made in synthesising new materials, which are primarily categorised into four groups: polymers, ceramics, metals, and smart materials. Smart materials are gaining increasing popularity due to their broader range of applications compared to conventional materials (Figure 1.1).[1]

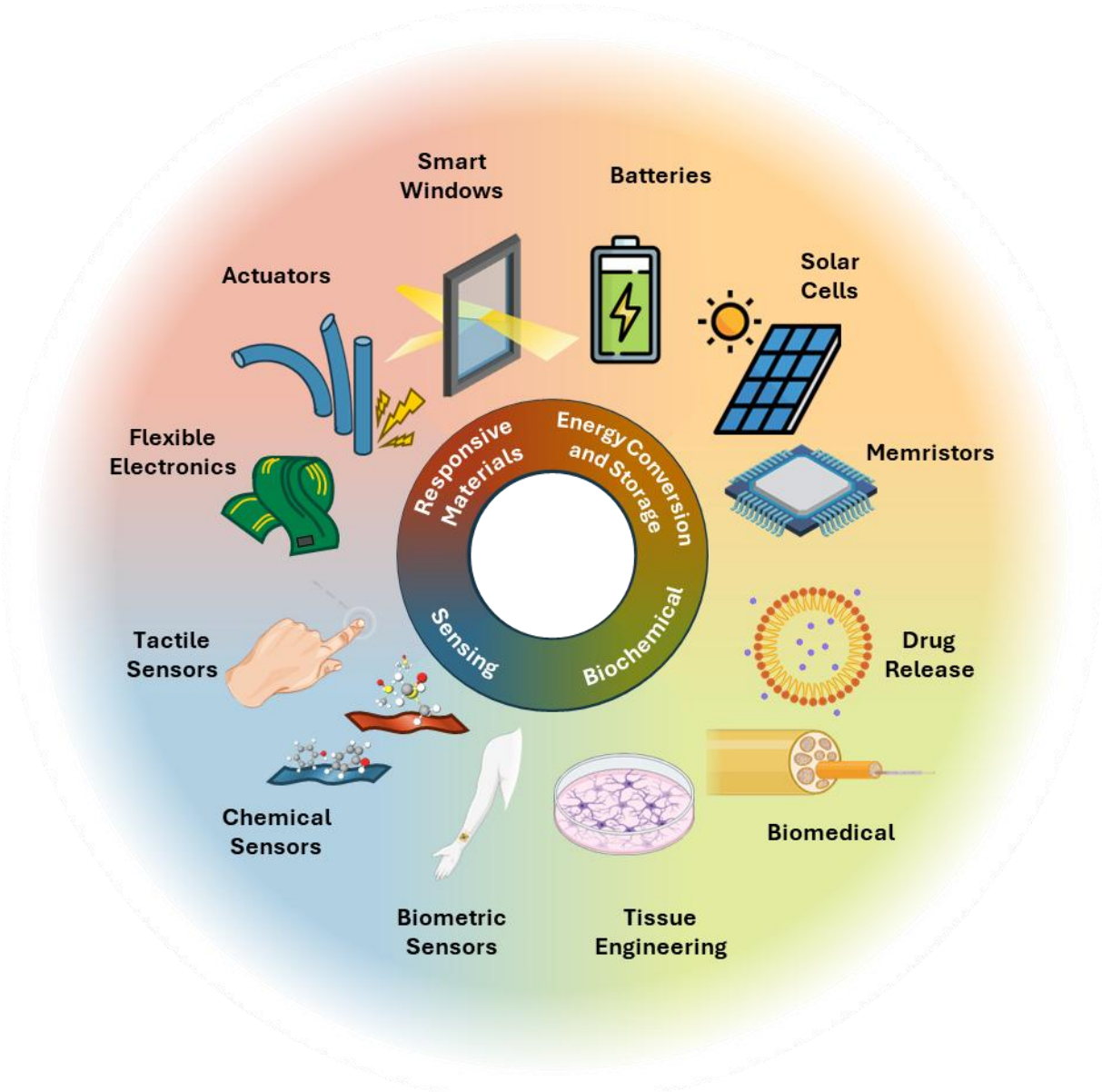


Figure 1.1. Cartoon of examples of reported uses of smart materials across a range of applications including energy conversion and storage, biochemical applications, sensing, and as responsive materials. [2-13]

Smart materials respond to external stimuli such as temperature, pH, electromagnetic fields, light, and other chemical compounds. These responses can include changes in shape,

electrical charge, or colour. These reactive properties can be tailored for a given application, affording functionalities that a traditional material cannot. Smart materials have many features that differentiate them from their traditional counterparts, including:[14]

- Transience: adopting various states allows a response to different stimuli.
- Immediacy: response to stimulus occurs without delay
- Self-actuation: the material can change its own conformation without external force
- Selectivity: the response to stimuli is distinguished and predictable
- Directness: the response occurs at the place of stimulation

Smart materials can be further classified as passive or active based on their properties. Passive smart materials can react to environmental changes while transferring energy in a single form.[15] Fibre-optic materials are a common example of a passive smart material capable of detecting changes in pressure through the internal refractive properties, which cause measurable changes in the incident light within the fibre.[16] Active smart materials can similarly respond to a range of stimuli, with the additional capability to transduce energy. This ability takes many forms, such as the capacity of the material to change shape under the influence of electric, thermal, or magnetic fields.[15] The advent of such materials has revolutionised the field of material science, giving rise to many exciting applications.

Mechanoresponsive materials were among the first examples of smart materials, with the discovery of piezoelectricity by Pierre and Jacques Curie in 1880. Piezoelectric materials generate an electric charge on their surface in response to the application of mechanical stress.[17] Conversely, when they are under the effect of an electric charge, a mechanical deformation can be induced. Since their initial reporting, smart materials have been sought for countless applications, particularly in the fields of sensing,[2, 3] biochemical engineering,[4-7] responsive materials,[8-11] and energy conversion and storage.[12, 13] For many applications, there is a need to engineer materials that combine a breadth of functions to enable their integration into these novel problems. For example, aerospace applications that demand miniaturisation for greater efficiency require materials that integrate the functions of sensors, actuators, and control circuit elements into a single system that can respond adaptively to changes in its surroundings.[18] Smart wearable textiles are also gaining interest; they need to be flexible, conductive, biocompatible, and comfortable, while providing highly sensitive physiological monitoring.[19] An extensive review of multifunctional material systems by Ferreira *et al*. demonstrates the demand and development of such materials for various applications.[20]

One area that has seen rapid advancement due to the emergence of multifunctional smart materials is the development of e-skins. Such materials aim to go beyond the functionality of skin-attachable healthcare devices, wearable technologies, and prosthetics. Human skin has many critical functions, ranging from environmental sensing to biological regulation and chemical release.[2] For e-skins to mimic or augments such functions, highly responsive and biocompatible materials are required. If such materials are to mimic the sensitivity of human skin, traditional rigid materials with higher deformation pressures are insufficient. However, the advent of responsive soft materials has seen the development of sensors capable of detecting low pressures with high sensitivity.[2] Beyond this, smart materials have proven capable of thermal sensing for monitoring a wearer's temperature,[21] chemical sensing for biohazard detection,[22] and electrical potential monitoring for diagnostics and human-machine interfacing,[23, 24] amongst other properties. Such multifunctionality would historically require invasive, bulky, and uncomfortable multicomponent constructions with traditional materials. However, the integration of soft, polymer-based hybrid electronics has enabled the development of unobtrusive and multifunctional tissue interfaces.[25]

Below the surface of the skin, smart materials have been designed to adapt and respond to physiological conditions, mimic tissue structure, support cell adhesion, and enable targeted molecular release.[4] To minimise the invasiveness of therapeutic techniques, there is a need for injectable, biocompatible materials that can respond to and adapt to various stimuli. Injectable materials capable of altering both chemical and self-assembled structures via enzymatic-, temperature-, and pH-initiated processes have been reported.[4] Such materials can be loaded with biomolecular agents to reduce reactive oxygen species to tackle health risks such as myocardial infarctions and osteoarthritis;[4, 26] enable the glucose-modulated release of insulin to non-invasively reduce the risk of hypoglycaemia and cognitive decline in Type 1 diabetes patients on prolonged and aggressive regimens of insulin,[6] or form conductive hydrogels to modulate the proliferation, migration, and differentiation of nerve cells to enable and accelerate nerve damage repair.[5]

Conductive smart materials are also key to developing sustainable materials that enable net-zero technologies. In developed countries, building energy consumption accounts for approximately a third of total energy consumption, with ventilation, heating, and air conditioning accounting for roughly half of that.[10] Therefore, developing efficient temperature regulation systems could afford a massive reduction in energy consumption and its associated emissions. Electrochromic smart windows have proven more effective than traditional operable blinds, able to decrease energy consumption by up to 30%.[27] Despite the

significant advantages, the necessity of an external power source has limited their adoption. To overcome this photochromic and thermochromic smart windows are under development to take advantage of incident light or temperature gradients, respectively, to passively power these technologies. Such materials could see greater adoption due to the low complexity installation in the form of a simple coating, as opposed to the highly complex, multilayer device structure required by electrochromics.[9]

To further diversify where electronic smart material can be integrated, there is a need for flexible yet functional materials. Materials which combine exceptional mechanical strength with a high degree of tunability allow them to be integrated with conductive polymers to yield multifunctional composites that are compatible with non-traditional processing techniques. This has given rise to the development of flexible electrochemical energy storage systems to enable sustainable wearable electronics.[28] By developing flexible materials with self-healing capabilities, these electronics can withstand a tremendous amount of strain, allowing them to be integrated into touch-sensitive displays.[12] With access to new modes of fabrication, smart materials can be woven directly into textile fibres. Such advances allow electronics to be integrated into clothing for passive light harvesting to power these emerging wearable technologies.[8]

1.1.2 Existing Smart Material Systems

The wide range of material applications each has unique needs. The increasing interest in smart materials has led to the development of many new functional materials. While many of these have facilitated the development of new technologies, there are serious limitations to address if the integration of smart materials in these promising applications is to continue advancing.

1.1.2.1 Inorganic Materials

Traditionally, metals such as copper, silver, and gold have been used in electronic applications due to their impressive electrical conductivity. In the case of the noble metals, their resistance to oxidation and corrosion arises from passivation by forming thin oxide surfaces and stable d-electron configurations. This has made noble metals particularly essential where long-term stability and dependability are required.[29] However, metals such as these are passive conductors with no ability to respond to external stimuli. In addition, while their conductivity makes them thermally resistant, they are mechanically brittle, making them incompatible with the flexible electronic applications that are emerging.[30] To

enable the development of truly smart systems, more complex composites and inorganic alternatives have been increasingly investigated. For example, inorganic smart materials have been widely used in the design of tactile sensors, high-precision actuators, and energy-harvesting applications, among others.[31]

Smart ceramic materials have found applications in actuation. For example, lead zirconate titanate (PZT) piezoelectrics are widely used due to their synthetic accessibility and cost-effectiveness.[32] However, the dynamic range of a piezoelectric device is limited to tens of micrometres, much smaller than the millimetre scale required for many applications.[33, 34] Not only do the mechanical properties of ceramics make them problematic for application in certain devices, but the most common materials of this class are lead-based. The toxic nature of lead to human health and the environment is well known.[35] As a result, significant regulations have been enacted to limit the use of lead-based technologies.[36] Subsequently, there has been significant interest and investigation into lead-free piezoceramic materials in the last two decades. Of the many papers and patents on this field, 61% are based on bismuth materials.[17] Despite bismuth being deemed a "green metal", its environmental fate is poorly understood.[37] However, due to its benefits over lead-based materials, the use of bismuth in smart material technologies is expected to increase. Given the unknown environmental consequences of its widespread use, environmentally compatible yet functional smart materials must be designed.

Light-harvesting materials are an important example of smart materials that require maximised efficiency. The high-performance methylammonium lead iodide is the most common light-absorbing material used in perovskite devices.[38] However, the state of the art in high-efficiency perovskite solar cells requires the tandem stacking of mixed bromide-iodide perovskites with lead-tin perovskites.[39] The highest-performing solar cells currently reported have shown much greater efficiency than their perovskite alternatives (Figure 1.2); however, they share similar environmental concerns. Similarly, the best lead-free III-V multijunction cell reported utilises inorganic materials composed of gallium, indium, arsenic, and phosphorus.[40]

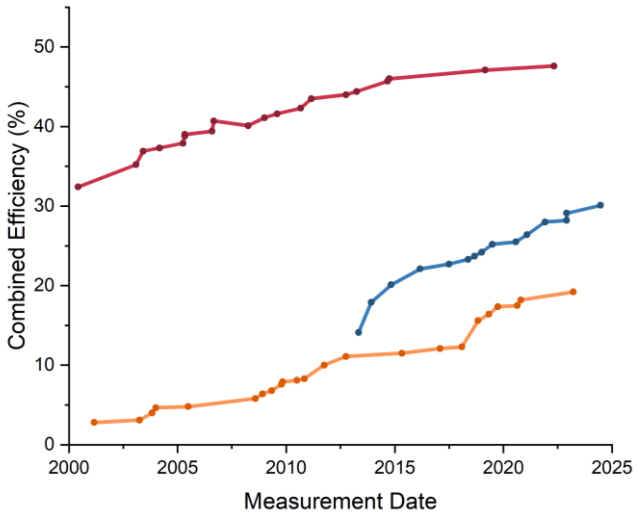


Figure 1.2. A chart of the highest confirmed conversion efficiencies for research cells. Efficiencies for III-V multijunction concentrator cells are shown in red, perovskite cells are shown in blue, and organic cells are shown in orange.[41]

Inorganic electrochromic films are at the forefront of the smart window market. Research has focused on cathodic electrochromic oxides that switch from a bleached state to a coloured state upon reduction. The most studied cathodic electrochromic oxide material is tungsten oxide, owing to its high optical contrast and stability in electrochromic applications.[42] However, both the high cost and poor conductivity of large-area metal-oxide smart windows limit their marketable applications.[43] Furthermore, oxide films are more stable with lithium electrolytes than with other electrolytes; thus, large-area applications are dominated by lithium devices.[42, 43] With no viable alternative to lithium battery technologies, serious shortages of lithium could soon emerge.[44] Given that cost is currently one of the main limitations to product viability, relying on lithium-ion electrolyte systems could significantly increase the cost of these technologies in the future. Therefore, systems developed from sustainable feedstocks could have significant marketable advantages when designing smart material systems.

1.1.2.2 Polymers

Unlike traditional, rigid, and brittle electronics; future electronics must be lighter, more portable, biocompatible, wearable, and more mechanically stable.[1] While many smart materials currently at the forefront of new technologies outperform traditional materials, this has come with high environmental, ecological, and economic costs. To overcome this, organic-based smart materials have been of increasing interest in recent years.

Poly(3,4-ethylenedioxythiophene):poly(styrenesulfonate) (PEDOT:PSS) is one such material which has attracted much attention (Figure 1.3). The excellent colour-switching

response speed, thermal stability, biocompatibility, stretchability, and electrical conductivity have made PEDOT:PSS a candidate for many smart material applications, particularly in electrochromic systems.[45]

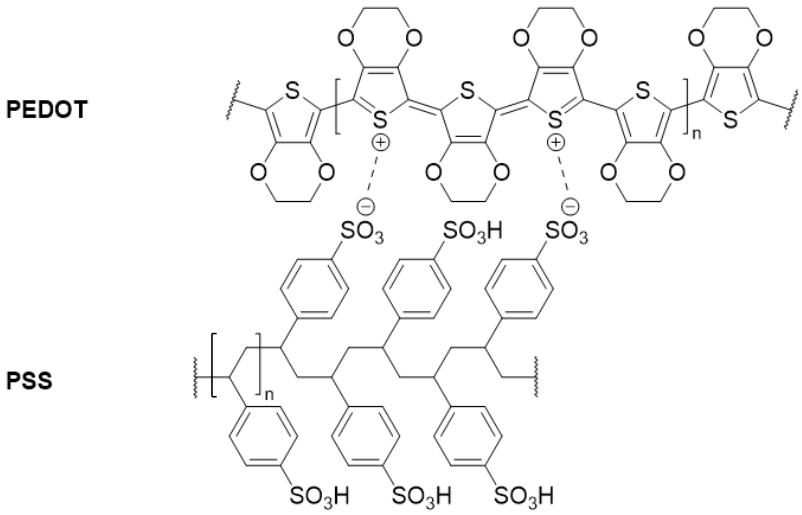


Figure 1.3. Chemical structures of PEDOT and PSS with non-covalent interactions between the polymers visualised with dashed bonds. Oxidative polymerisation of the EDOT monomer results in an electron deficient PEDOT polymer. The charge-charge attraction between PEDOT and PSS stabilises the PEDOT by achieving charge equilibrium.[46]

A major advantage of PEDOT:PSS is its robust processability. The compatibility of PEDOT:PSS with a range of application techniques, including casting, coating, and deposition, makes it a versatile material that can meet the needs of highly varied applications.[47] However, PEDOT:PSS films suffer from poor water stability due to their dispersibility in water, leading to the delamination of films. This has limited the application of such materials in aqueous media, particularly in bioelectronics.[48] To overcome this, chemical cross-linking is typically required to produce water-stable films at the cost of significant electrical performance. However, a recent investigation by Doshi *et al.* indicates that water-stable PEDOT:PSS films can be achieved solely through thermal processing, which may make PEDOT:PSS a viable material for future bioelectronic applications.[48] While advancements in PEDOT:PSS-based materials continue to be made, several challenges remain to be overcome if they are to be integrated into real-world applications. The dense morphology of PEDOT:PSS reduces ion pathways and hinders ion migration, resulting in poor cycling stability and a narrow optical tuning range.[49] Even with property enhancement with the addition of organic polymers or inorganic compounds, PEDOT:PSS materials are still not viable in many applications.[45] Furthermore, while PEDOT:PSS-based

"flexible" materials have been reported, most exhibit only small-angle bending and stretching, making them incompatible with flexible device assembly.[45]

In contrast, the preparation of truly flexible smart materials prepared from conductive polymers has been well documented. Nanofibre synthesis can yield flexible and functional materials with unique properties, but it also has limitations, including limited scalability in manufacturing, reduced uniformity and reproducibility, and reduced durability.[50] A variety of techniques can be used to make functional nanofibres, each with unique advantages and disadvantages. Viscoelastic polymers can be drawn from a micropipette during solvent evaporation to form long fibres. While such fibres are typically uniform, diameters smaller than 200 nm can be unachievable. The low and discontinuous production rate makes them a poor candidate for application.[50, 51] Template synthesis enables aqueous polymer solutions to be mixed with a coagulant to afford conductive nanofibres with controlled dimensions. Still, again, the discontinuous nature of template synthesis limits its application.[50, 52] Phase separation-mediated nanofibre growth can yield uniquely porous nanofibres; however, the few polymers compatible with this technique form nanofibres with highly variable dimensions and morphology, making their preparation difficult to control.[50] Electrospinning is one of the most widely researched methods for preparing polymer nanofibres. A strong electric field accelerates a jet of polymer solution, allowing it to quickly dry into fibres with high tunability in microstructure and dimensions. While the fibres can be prepared in a continuous process, electrospinning methods typically require the use of toxic solvents, which makes large-scale preparation hazardous and again limits their application.[50, 53] If nanofibres are to become a viable candidate material for device applications, preparations that produce robust, reproducible, and scalable fibres are required, without the need for toxic substrates.

There are a range of synthetic and natural polymer materials which provide various advantages when fashioned into smart materials. Such polymers are often biocompatible, easily synthesisable, and highly tunable.[54] Integrating functional chemical motifs into the polymer structure can combine the mechanical strengths of the polymer that can't be obtained by the functional substructure alone, while achieving unique functions that are impossible for the functional molecule or polymer alone.[55] However, the overall performance of smart polymer materials can suffer significantly, limiting their applicability. The high activation voltages associated with electro-responsive polymers can cause heating, increased battery consumption, and material degradation, impacting the durability of the polymers.[56] Compared to their metal-based alternative, shape memory polymers, which can

undo mechanical deformation upon heating, suffer from either slow transitions or high activation temperatures. More generally, thermoresponsive polymer systems are often limited by their narrow operating range and poor repeatability.[57] The limited number of precursor monomers provides a significant challenge for tailoring design.[58] While there is scope for the tailoring of smart polymers by synthesising new monomers, or tuning properties by copolymerising different monomers, the limited understanding of design rules for such systems makes this an ongoing challenge.[55, 58] While the synthesis of such polymers may be simple, they can be difficult to control. Even with a well-controlled synthetic procedure, polymers can suffer from inconsistent molecular weight distributions, topologies, and polymerisation sequences (in the case of co-polymerisation). This can alter properties at both the molecular and bulk scales, making the reproducible synthesis of smart polymers problematic.

1.2 Small Molecules

In contrast to polymers, self-assembling small molecules can form materials from discrete molecular components with well-defined chemical structure.[59] Changes to the self-assembled structure will directly affect the macroscopic properties of the material. A consequence of being held together by forces weaker than covalent bonding means the tuning of bulk properties is enabled at the molecular level.[59, 60] Many different classes of self-assembling small molecules have been investigated, each with unique properties and applications (Figure 1.4).

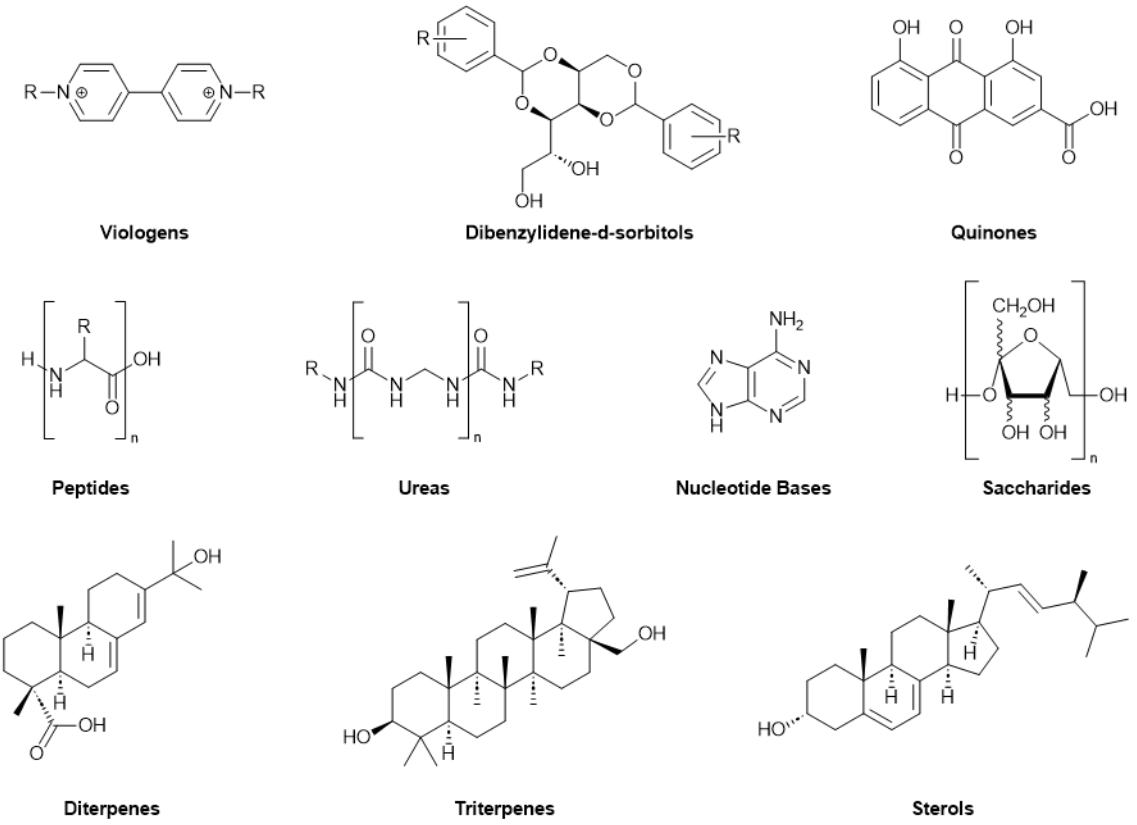


Figure 1.4. Examples of self-assembling small molecules. General chemical structures are given for viologens, dibenzylidene-d-sorbitols, peptides, ureas, saccharides, naphthalene diimides, and perylene bisimides. Chemical structures of rhein, dehydroabietic acid, butelin, and ergosterol are given as examples of quinones, diterpenes, triterpenes, and sterols, respectively. [61-69]

1.2.1 Small Molecule Self-Assembly

While polymer materials rely on the entanglement and cross-linking of covalently linked polymeric chains, supramolecular materials are instead built from small molecules which assemble in one dimension, which can then go on to assemble and entangle in a hierarchical manner on the micro- and mesoscales.[70] Such molecules are typically amphiphilic, utilising the incompatibility between the polar/non-polar regions of the molecule and a given solvent to drive self-assembly.[71] In the case of a polar solvent, this hydrophobic effect causes the self-assembling molecules to preferentially sequester their nonpolar structures to form micelles. Other non-covalent interactions such as hydrogen bonding and π-stacking have a significant effect on a molecules propensity to self-assemble, as well as controlling the topology of the resulting self-assemblies.[72] The use of weak non-covalent interactions to facilitate the assembly of the hierarchical structures makes the formation of the material much easier to reverse. This enhanced reversibility confers stimuli-responsive properties, enabling their use as smart materials.[73]

Given the ordered structural nature of supramolecular self-assembly, crystal structure has historically been investigated to understand the self-assembly occurring in supramolecular materials. Crystal structure characterisation is becoming increasingly routine and is rarely incorrect.[74] With an abundance of large organic crystal structure libraries compiled across chemistries, this may appear to be an effective approach for understanding supramolecular systems.[75] However, as understanding of supramolecular systems increased, it became clear that the crystal structure is often independent of the supramolecular structure of the resulting material, even when grown directly from it.[72] For example, Houton *et al*. have shown that, in a dipeptide-functionalised naphthalene, the lattice packing shows a subtle relationship between the crystal phase and the supramolecular fibres they can form; however, the lattice packing of the fibre was less condensed. It was also highlighted that the interactions and separations of the fibre molecules were aligned to promote 1D growth, while the crystals were aligned to multiple axes to enable 3D growth.[76] Supramolecular self-assembly is governed by multiple interactions between solute molecules and the chemical environment, rendering them thermodynamically independent of the crystal state. Furthermore, changes in material formation enable access to different thermodynamic states, yielding materials with varying nanostructures and properties.[77] While this thermodynamic complexity makes understanding supramolecular materials much more difficult, it also offers an opportunity to develop and optimise a range of materials from a single molecule (Figure 1.5).

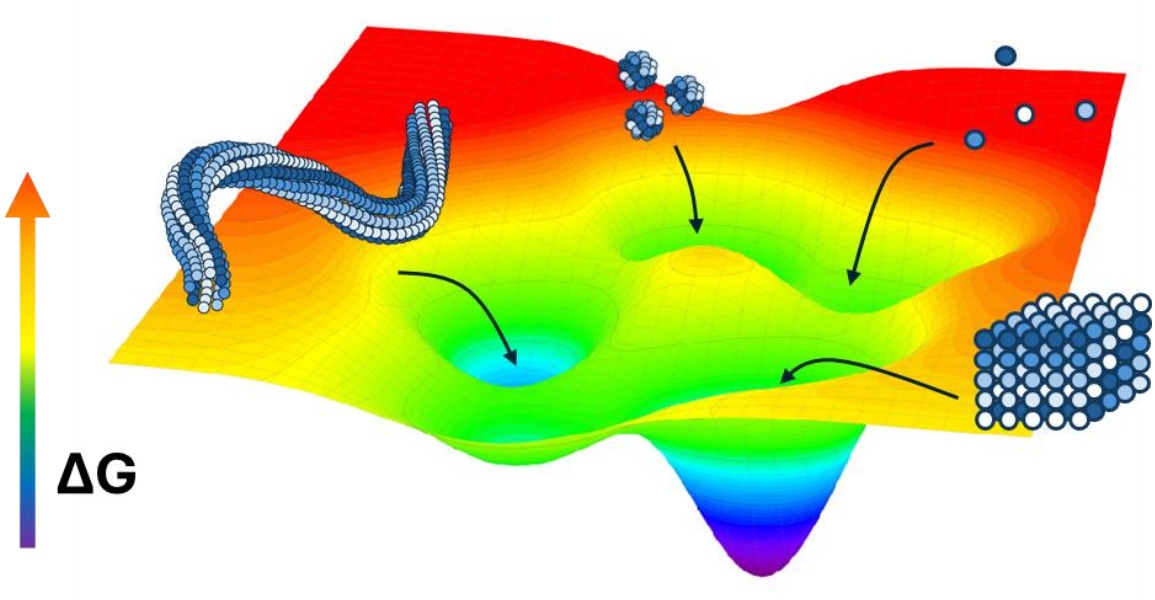


Figure 1.5. Cartoon of the Gibbs free energy landscape illustrating the thermodynamic complexity of self-assembled systems. Thermodynamically stable (or metastable) states are determined by global and local minima of the Gibbs free energy (ΔG). These self-assembled states are independent of the crystal state, and can be accessed though changes to the chemical environment represented as the X and Y plane of the surface. Adapted from Sasselli *et al*.[77]

1.2.2 Supramolecular Interactions

Supramolecular chemistry is driven by a spontaneous decrease in the Gibbs free energy (ΔG), governed by the change in enthalpy (ΔH) and the change in entropy (ΔS) of the self-assembly process, with the importance of entropy in determining the free energy of self-assembly increasing with increasing temperature (T). At higher temperatures, the greater thermal motion of molecules amplifies the thermodynamic influence of entropy, making favourable entropy changes increasingly important in driving self-assembly (Equation 1.1).

$$\Delta G = \Delta H - T\Delta S \qquad \text{(Equation 1.1)}$$

Chemical processes are spontaneous when the Gibbs free energy is less than zero. Given that absolute temperature is positive, ΔG and ΔH are proportional, and ΔG and ΔS are inversely proportional, there are four possible regards to the parity of the enthalpy and entropy changes (Table 1.1).

Table 1.1. Summary of the impact of changes in enthalpy and entropy on Gibbs free energy.[78]

	ΔH > 0 (endothermic)	ΔH < 0 (exothermic)
ΔS > 0 (increase in entropy)	ΔG < 0 at high temperature ΔG > 0 at low temperature Process is spontaneous at high temperature	ΔG < 0 at any temperature Process is spontaneous at any temperature
ΔS < 0 (decrease in entropy)	ΔG > 0 at any temperature Process is not spontaneous at any temperature	ΔG > 0 at high temperature ΔG < 0 at low temperature Process is spontaneous at low temperature

Entropy, as defined by the Boltzmann's statistical definition of entropy (Equation 1.2), is directly proportional to the natural logarithm of microstates available to a system (Ω).[79] A self-assembling chemical system is achieved when molecules spontaneously associate, thereby restricting their translational, rotational, and conformational freedom. Subsequently the number of microstates available when the molecules are dispersed decreases dramatically as they self-assemble, leading to a ΔS with a large negative magnitude.

$$S = k_B \ln \Omega \qquad \text{(Equation 1.2)}$$

Therefore, the self-assembly of the solute molecules provides a substantial thermodynamic barrier to be overcome. If self-assembly is to occur spontaneously there must either be some other process by which the entropy of the overall chemical system is increased, or made enthalpically favourable by releasing energy. Therefore, the formation of attractive non-covalent interactions is necessary to enable the thermodynamic accessibility of self-assembled states. While non-covalent interactions are typically weaker than covalent interactions, a variety of these interactions can be cooperatively utilised to achieve a stable supramolecular system which is necessary to overcome the significant decrease in entropy afforded by the solute molecules.[80] Several key non-covalent interactions have been summarised in Table 1.2.

Table 1.2. Comparison of strength and distance dependence or typical range of covalent and supramolecular interactions.[80-83]

Interaction	Strength (kJ mol^{-1})	Distance Dependence	Range (Å)
Covalent	150 – 450		1 – 2
Charge-charge	200 – 300	$\propto r^{-1}$	
Charge-dipole	50 – 200	$\propto r^{-2} – r^{-4}$	
Dipole-dipole	5 – 50	$\propto r^{-3} – r^{-6}$	
Hydrogen bonding	4 – 120		1 – 3
π–π	0 – 50		3 – 4
Cation–π	5 – 80		4 – 6
Van der Waals	< 5 kJ mol^{-1}, variable depending on surface area	$\propto r^{-6}$	
Hydrophobic	Related to solvent–solvent interaction energy		< 100

Coulombic interactions including charge-charge, charge-dipole, and dipole-dipole interactions vary substantially in strength and range. While charge-charge interactions are of a similar strength to covalent bonds, they are not orientation dependent, lacking the directionality required to guide the assembly of complex architectures.[84] Subsequently the highly-specific orientation necessitated by the weaker dipole-dipole interactions often makes them more impactful, driving the alignment and orientation of the self-assembly.[85]

The benefits of including highly directional interactions in supramolecular systems have made hydrogen bonds one of the most important non-covalent interactions in supramolecular design. Hydrogen bonds are a specific dipole-dipole interaction between a proton donor (D), a moiety with a hydrogen covalently bound to an electronegative atom; and a proton acceptor (A) is an electron-withdrawing moiety with which the positively charged hydrogen can

interact. The resulting hydrogen bond is weak, yet multiple hydrogen bonds together can impart substantial stability to molecular interactions.[84] Molecules can be designed to make use of these interactions to balance both attractive and repulsive forces to moderate the overall strength and orientation of the self-assembly (Figure 1.6).[80]

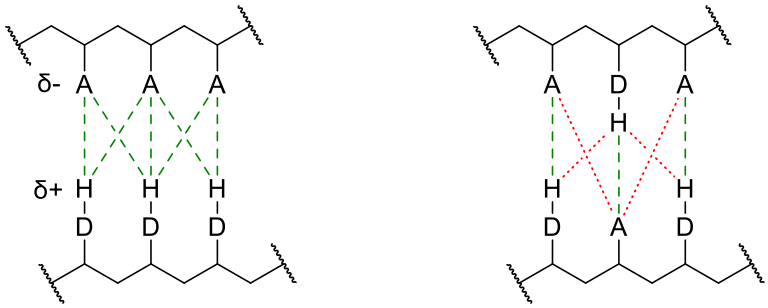


Figure 1.6. Degree of supramolecular attraction is moderated by the design of donor acceptor systems. AAA-DDD system promotes attractive forces between each acceptor and donor (green dash). The mixed ADA-DAD system exhibits both attractive AD interactions as well as repulsive AA and DD interactions (red dots).

Interactions between typically electron-rich π-electron cloud and electron-poor σ-hydrogens give rise to attractive π-π interactions between aromatic molecules. Unlike the localized and highly directional nature of hydrogen bonds, π-π stacking involves extended, delocalized electron clouds distributed across planar aromatic rings. Small unsubstituted aromatics typically favour T-shaped (edge-to-face), while larger polyaromatic systems and those with electron withdrawing functionalities more commonly adopt parallel-displaced geometries. These arrangements maximise attraction between the electron rich π-cloud and electron poor σ-system, while minimising repulsion between both π-systems and both σ-systems. Co-facial assemblies observed only rarely but can arise when electron density from the π-cloud is sufficiently lowered to promote attraction between electron-rich and electron-poor π-systems, among other scenarios (Figure 1.7).[86, 87] These interactions operate over broad molecular surfaces, promoting the formation of two- and three-dimensional supramolecular architectures and contribute significantly to the stability of the extended frameworks. In addition to π-π stacking, the electron-rich π-cloud of an aromatic system can interact favourably with cationic species through cation-π interactions. The strength of these interactions varies considerably depending on both the nature of the cation and the electronic properties of the aromatic π system. Owing to their strong regioselectivity, cation-π interactions provide an effective means of directing and controlling supramolecular self-assembly processes.[88]

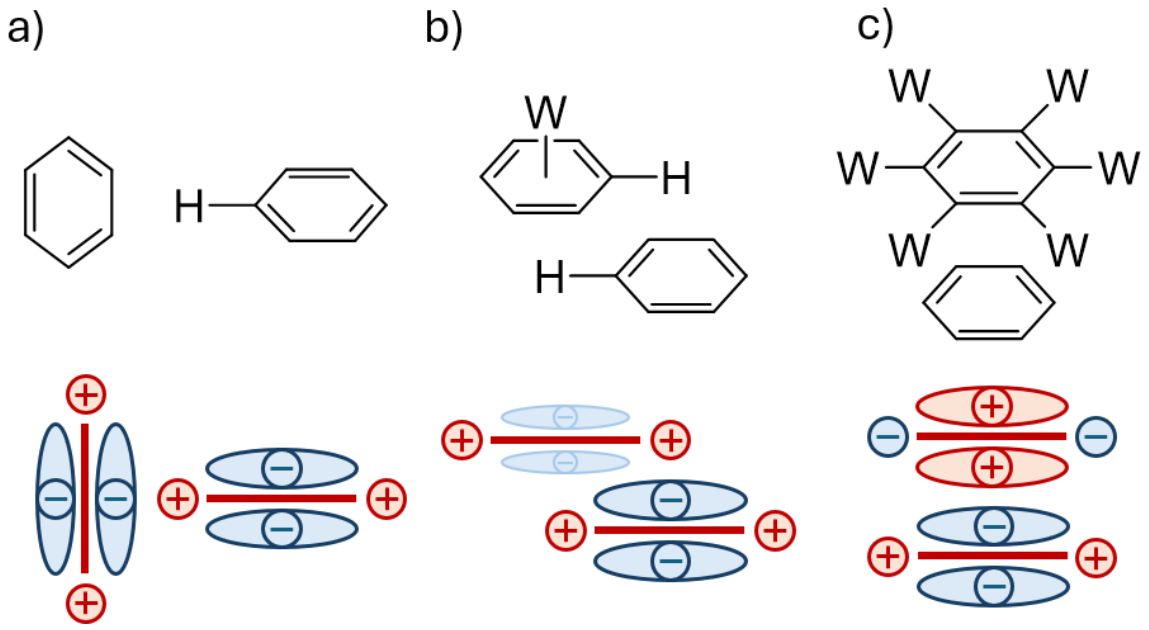


Figure 1.7. Schematic depiction of different geometric configurations of interactions between atomic rings. a) Small unsubstituted aromatics stack edge-to-face to maximise interactions between the electron rich π-cloud and electron poor σ-hydrogens, while minimising repulsion between both π-systems and both σ-hydrogen systems. b) Electron withdrawing group (W) depletes the π-electron density and relieves the electrostatic repulsion with the π-electron cloud of the unsubstituted ring. Attractive interactions between the electron deficient σ-hydrogens and π-electron clouds favours offset stacking. c) Complete substitution with electron withdrawing groups results in both an electron deficient core and electron rich substituents that can favourably interact with the electron rich π-system and electron poor σ-hydrogens of the unsubstituted ring, respectively, leading to face-to-face stacking. [86, 87]

Van der Waals forces encompass several interactions between uncharged molecules including those between instantaneously induced dipoles (London dispersion forces), and between permanent and induced dipoles (Debye forces). Although they are typically weaker than electrostatic interactions or hydrogen bonding, their ubiquity in all molecular systems means they can always play an important role in non-covalent binding. While they provide fewer opportunity for tailoring design, as they scale with system size their importance can grow for larger molecules and molecular assemblies.[89] This can provide a significant enthalpic contribution to achieving spontaneous self-assembly across systems.

Hydrophobic effects arise from the tendency of non-polar molecules or molecular regions to minimize their exposure to aqueous environments. This behaviour is driven by the preference of water molecules to interact with one another or with other polar species through favourable hydrogen-bonding and dipole interaction. Hydrophobic effects can be divided into enthalpic and entropic contributions.[90] Self-assembly can be enthalpically favoured through attractive interactions between the molecules, however, the resulting structure may form a hydrophobic cavity containing energetically unfavourable water molecules which increases the system's enthalpy. Displacement of these confined water

molecules into the bulk solvent reduces the unfavourable interactions while also allowing them to participate in the solvent's hydrogen-bonding network, increasing the number of hydrogen bonding interactions, thereby further lowering the enthalpy and increasing the thermodynamic favourability of the overall process. In contrast, self-assembly is intrinsically associated with an entropic penalty due to the loss of translational, rotational, and conformational degrees of freedom of the assembling components.[80] However, this penalty is often offset by favourable solvent entropy changes. Water molecules confined within hydrophobic cavities or ordered around exposed hydrophobic surfaces become more disordered upon release into the bulk solvent.[91] Furthermore, because hydrophobic molecules cannot participate in water's dynamic hydrogen-bonding network, water molecules at the solvent-solute interface adopt a more ordered arrangement. Self-assembly reduces the total exposed hydrophobic surface area and consequently decreases the extent of this ordered interfacial water. The resulting increase in solvent entropy can provide a significant driving force for assembly. Consequently, the hydrophobic effect is often a major contributor to supramolecular assembly in aqueous solution, arising from a combination of favourable enthalpic interactions among water molecules and increases in solvent entropy upon the reduction of hydrophobic surface area.[80]

The diversity of non-covalent interactions exhibited by small molecules provides extensive opportunities for the design of functional supramolecular materials. The balance between attractive and repulsive intermolecular forces governs the resulting thermodynamic favourability of assembly processes, enabling the preparation of highly tailored systems with a significant degree of structural and functional tunability.

1.2.3 Tuning Self-Assembly

The morphology and mechanical properties of the self-assembled structures, along with the degree of entanglement, type, and number of interactions between neighbouring structures, contribute to the bulk properties and functionality of the material. While self-assembly is driven by thermodynamically favourable interactions, the supramolecular structures are often kinetically trapped, in which the assembly process is dictated by the pathway rather than the thermodynamics of the final state.[92] Therefore, the properties of the resulting material are not only controlled by molecular structure but also by process-dependent factors, allowing a single material to be tuned to afford varying properties. While such tunability increases the complexity and utility of the materials that can be accessed, it also complicates the characterisation of these materials. Given the process dependence of the

resulting material, and therefore its characterisation, without strict standardisation of preparation and analysis, reproducibility and comparability of results are lost. While automated characterisation platforms are becoming more accessible, their inflexibility and prohibitive cost still necessitate custom-built solutions for some researchers, thereby undermining reproducibility in characterisation.[93] When such systems are inaccessible, one must rely on the standardisation of experimental processes, as it has been shown that without a well-designed and shared methodology, the preparation of supramolecular materials is difficult to reproduce.[94] Without interlaboratory comparability between experimental endpoints, the compilation of consistent datasets becomes problematic as significant experimental noise in the dataset can limit the possible insights they can afford. Subsequently, it is important to consider the preparation changes that can be made to tune the structures to ensure a consistent characterisation protocol is established.

Van der Tol *et al*. have noted that while supramolecular assemblies are becoming increasingly ubiquitous in the literature, the role of solvents in the self-assembly is often not addressed.[95] However, understanding the many interactions within such a system is crucial for the development of functional materials and understanding the pathway complexity. A preliminary assessment of the effects of varying solvents and the self-assembly of a triazine derivative (**S-T**) by means of atomic force microscopy, significant changes were observed (Figure 1.8). Low-polarity solvents afforded drop-cast films with significant aggregation, with network-like single supramolecular chains forming in toluene and higher-order aggregates of assembled fibres in ethyl acetate. However, as solvent polarity increased, the system afforded much smaller aggregates. In fact, additional *in situ* investigation of the solution identified the smaller aggregates in acetic acid to be a drying artefact of concentrating a molecularly dissolved solution.

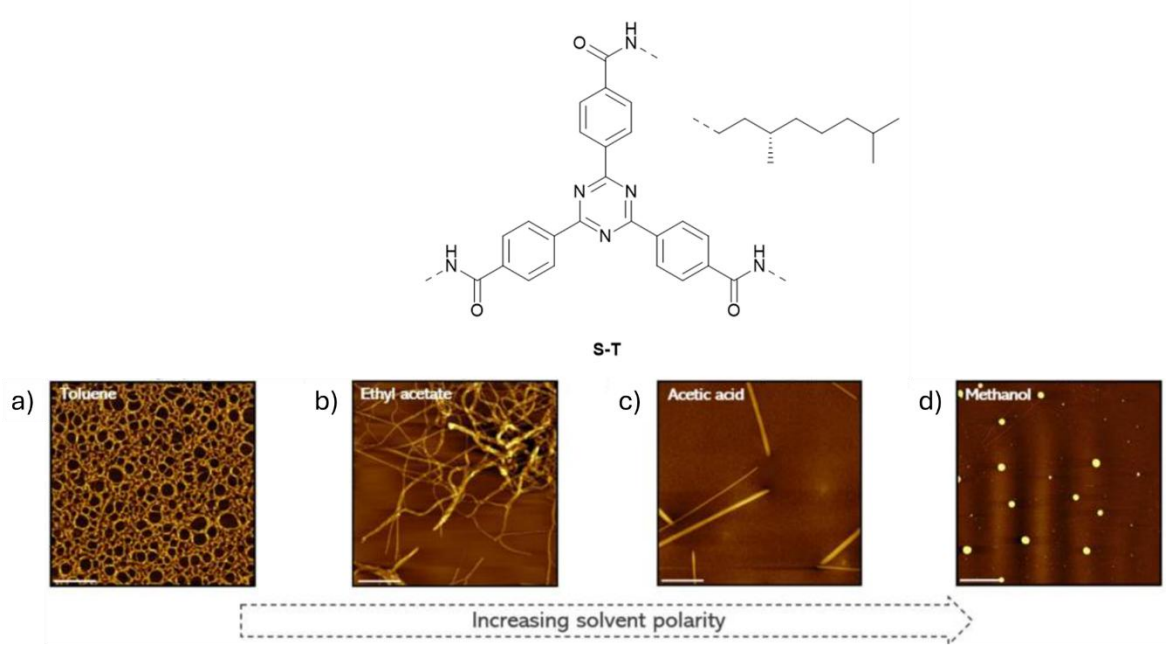


Figure 1.8. General structure of triazine **S-T** and AFM height images of dropcasted samples from a) toluene, b) ethyl acetate, c) acetic acid, and d) methanol demonstrating the solvent-dependent morphology of **S-T**. The inset scalebar represents 2 μm. Reproduced from van der Tol *et al.*[95]

The preliminary results demonstrated the crucial effect of solvents on the self-assembly of **S-T**. Subsequently, a Hansen solubility parameter (HSP) model was constructed in 58 solvents for **S-T** (Figure 1.9). HSP is described by three parameters δH, δD, and δP which represent the energy arising from hydrogen bonding, polar, and dispersion interactions, respectively, that hold the liquid together. It was shown that the HSP could describe the effects of the solvents well, classifying good solvents (providing dissolved monomers), bad solvents causing precipitation, bad solvents affording higher-order aggregates, and solvents inducing the formation of supramolecular polymers. The accuracy of the HSP model was sufficient to show rational trends that help understand solvent effects in supramolecular self-assembly. Such models also identified the influence of solute concentration on the self-assembly, with increasing concentration reducing the HSP space describing good solvents dramatically, while the space describing the solvents that afford supramolecular solvents remains mostly unaffected. The contraction of the good solvent space is indicative of the limited solubility, more likely to afford aggregates on the boundary of that space at higher concentrations, while the intermolecular solute-solute interactions dominate in the supramolecular polymerisation space. The work by van der Tol is a clear demonstration of the key role that solvents play in controlling supramolecular assemblies, as well as other process-dependent considerations, such as concentration.[95]

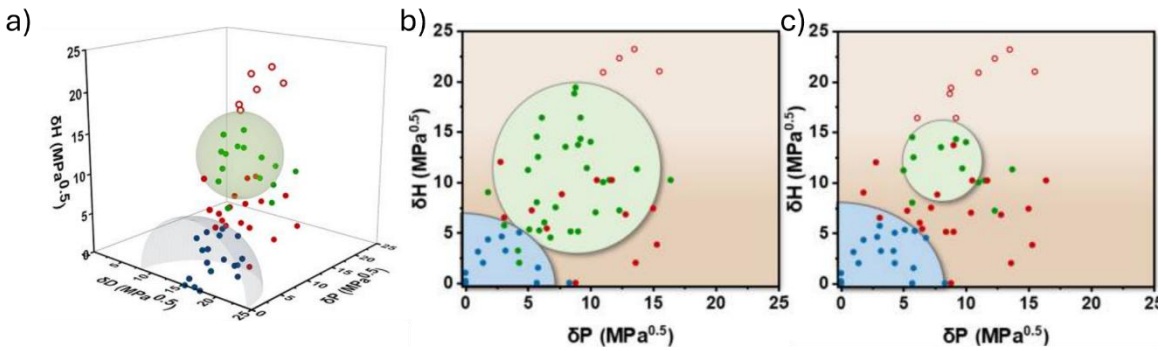


Figure 1.9. a) Experimentally determined 3D Hansen solubility space of ***S*-T** based on 58 different solvents. HSP is described by three parameters δH, δD, and δP which represent the energy arising from hydrogen bonding, polar, and dispersion interactions, respectively, that hold the liquid together. b-c) 2D representation of the Hansen solubility space of ***S*-T** at concentrations of 25 and 500 μM, respectively. The solvents are divided into four distinct categories: bad solvents causing precipitation (open red dots), bad solvents inducing higher-order aggregates (closed red dots), good solvents providing dissolved monomers (green dots), and solvents promoting the formation of supramolecular polymers (blue dots). Reproduced from van der Tol *et al*.[95]

Modifying the key intermolecular interactions between the solvent and aggregates enables control over the self-assembled materials. A straightforward way to induce material changes is by adjusting the preparation temperature. Debnath *et al*. demonstrated that the thermal history of solutions can bias the kinetic system towards different local minima. This control enables tuning of supramolecular assembly by balancing the contributions of hydrogen-bonding and π-π stacking interactions. At higher temperatures, less-ordered aromatic interactions are favoured, while at lower temperatures, assembly results in a more ordered hydrogen-bonded network.[96] Hisamatsu *et al*. have demonstrated how modulating temperature can be used to selectively prepare different topologies of self-assembled 4-aminoquinoline amphiphiles. When heated to 60°C, these molecules self-assemble into room-temperature-stable nanofibres. Heating the same starting solution or the nanofibre solution to 80°C enables the selective construction of perforated vesicles. While the metastable vesicles slowly transform into an intertwined nanofiber system at room temperature, brief heating restores the selective assembly of the perforated vesicle (Figure 1.10).[97]

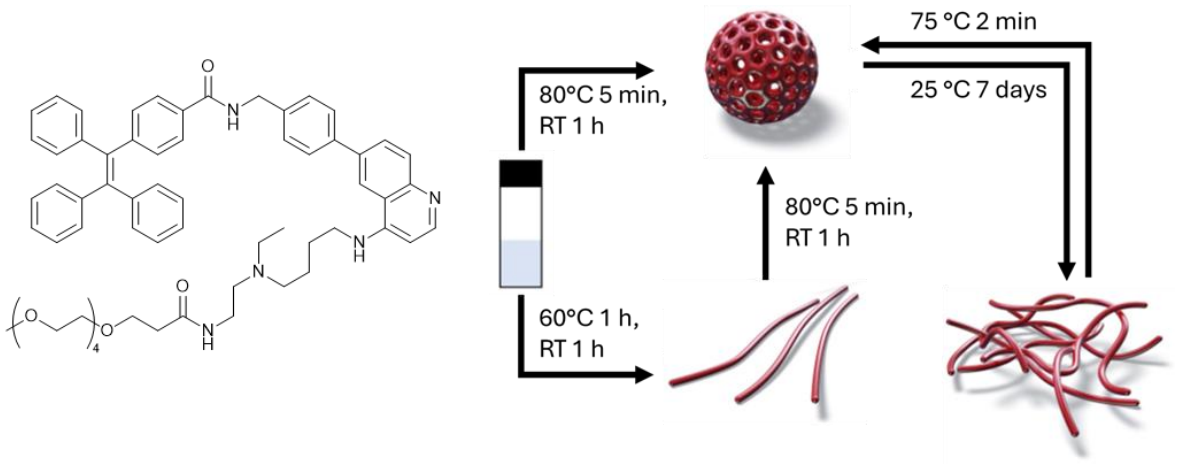


Figure 1.10. Chemical structure of the 4-aminoquinoline amphiphiles and a schematic representation of the selective construction of perforated vesicles and nanofibers by temperature control of the self-assembly.[97]

In aqueous systems, introducing ionisable moieties, such as carboxylic acids, provides a means of controlling assembly by modulating interactions between the aggregate and the solvent or other additives. For example, Nanda *et al.* demonstrated that preparing gels using a phenylalanine-functionalised pyrene to form hydrogels at different pHs (7.46–10.5) afforded significantly different materials from a single compound. At high pH (>12.0), self-assembly favoured the formation of achiral tape-like aggregates, while at low pH (7.46 – 11.0), left-handed helical structures were formed. In addition to controlling the nanoscale morphology, the macroscale property of thixotropy (upon mechanical shaking, the gel is solubilised and, after resting, re-forms into a gel) was observed only at lower pH. Despite achieving thixotropy in gels over the pH range 7.46 – 10.5, only those at pH 7.46 fully recovered their strength, highlighting the pH-dependent control of this unique bulk property.[98]

The presence of other ions in the solution can also affect supramolecular materials. Salts can significantly alter the solubility of a molecule and, therefore, the stability of their self-assembled aggregates based on their effect on the fibre-solution surface tension.[99] McAulay *et al*. showed how micellar aggregates of diphenylalanine-functionalised naphthalene formed at high pH varied in the presence of different alkali metal salts. Most notably, changing the salt can change the pK_a of the aggregate. Subsequently, gels form at different rates as pH decreases, thereby tuning the bulk properties of the resulting material.[100]

As discussed, for the utility of smart materials to be fully realised, they must be multifunctional. Therefore, combining the above tunability with responsive functionality will be key to designing the next generation of adaptive smart materials. Therefore, we need to identify and optimise a chemical system with such features.

1.2.4 Rylene Self-Assembly

Functionalised rylenes, including naphthalene diimide (NDI) and perylene bisimide (PBI), are small-molecule dyes that have become of significant interest due to their ability to self-assemble into a wide range of functional materials.[101, 102] The rylenes are polycyclic aromatic scaffolds substituted with two dicarboxylic acid imide groups (Figure 1.11). Rylenes possess high thermal, chemical, and photochemical stability, along with the propensity to self-assemble into tuneable, optoelectronic, responsive fibres. These properties make NDIs and PBIs ideal candidates for developing e-skin materials. Not only do e-skins need to be flexible, which supramolecular assemblies can provide, but they also need to withstand the heat, pressure, and salts the human body exudes, all while being lightweight and electronically responsive to unobtrusively monitor the surrounding environment.[103]

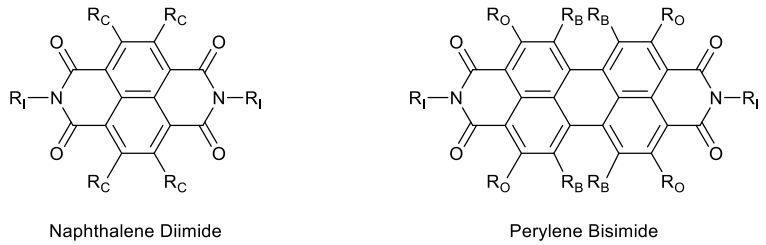


Figure 1.11. General chemical structure of a naphthalene diimide and perylene bisimide with hydrogens replaced with R_I, R_C, R_B, and R_O representing the imide, core, bay, and ortho position functionalisation, respectively.

Understanding how these supramolecular systems aggregate is important because the type of aggregation directly impacts the material's electronic properties. This is due to the spatial dependence of the intermolecular coulombic coupling (J_c) between the stacked molecules, which will depend upon the distance (R) and angle (θ) between two dipole moments (μ), as well as the dielectric constant of the medium (ε_r), as shown in (Equation 1.3 and Figure 1.12.[104]

$$J_C = \frac{\mu^2(1 - 3\cos^2\theta)}{4\pi\varepsilon_r R^3} \qquad \text{(Equation 1.3)}$$

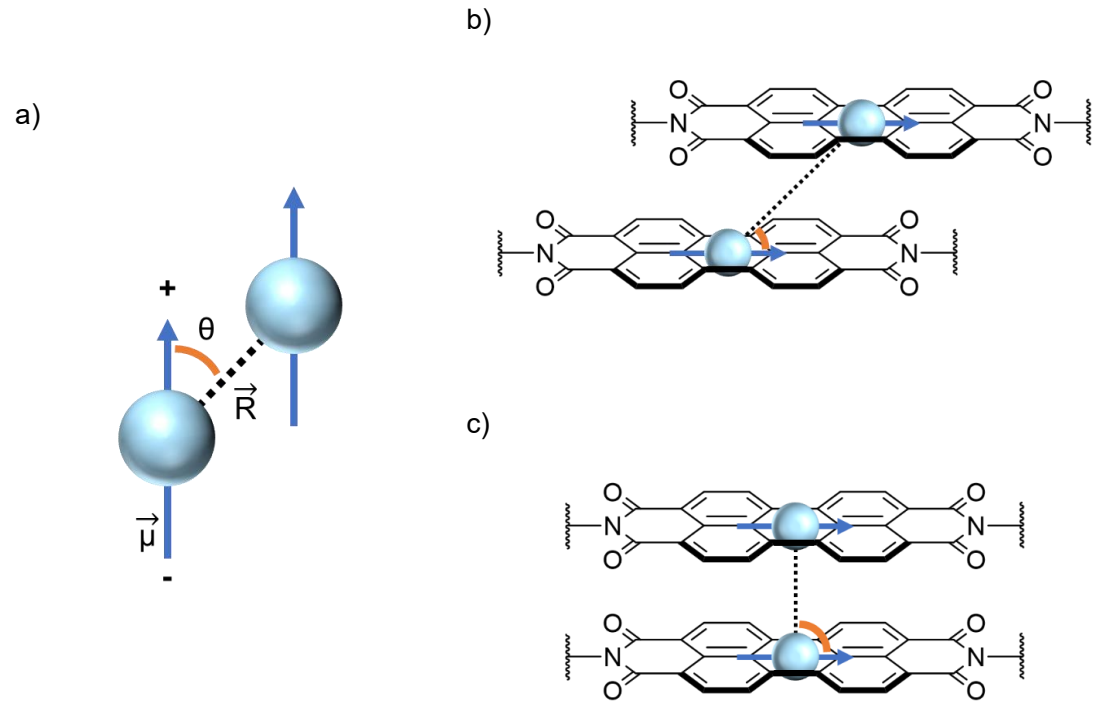


Figure 1.12. a) Relative orientation of transition dipole moments defining b) *J*-type aggregates ($\theta < \theta_M$) and c) *H*-type aggregates ($\theta_M < \theta < \pi/2$) in an PBI dimer.

J-type aggregates are defined by those with a dipole orientation less than the magic angle ($\theta_M = 54.7°$, for which $J_C = 0$), resulting in a "head-to-tail" orientation. In contrast, H-type aggregates have a dipole orientation between the magic angle and 90°, resulting in a "side-by-side" orientation. Subsequently, *J*-type aggregates result in a negative Coulombic coupling ($J_C < 0$) while *H*-type aggregates result in a positive Coulombic coupling ($J_C > 0$).[104]

The change in Coulombic coupling affects the electronic properties of the aggregate by a shift in the energies of the allowed ground state transitions of the molecule, resulting in the characteristic red-shifted and blue-shifted UV-vis absorption peaks of the *J* and *H* aggregates, respectively, compared to the monomer peak. This is because, in the molecular dimer, Coulombic coupling leads to linear combinations of two excited states. The out-of-phase transitions have a net-zero transition dipole moment, which is electronically forbidden, so the electronic properties are governed by the transition between the ground state and the in-phase excited state.[105] This gives rise to the red shift in *J*-type aggregates, as the negative Coulombic coupling leads to a decrease in the in-phase excited-state energy. The decline in transition energy can be rationalised by the reduction in electrostatic repulsions between in-phase transition dipole moments when arranged head-to-tail. Conversely, the positive Coulombic coupling leads to an increase in the in-phase excited state energy due to the rise in electrostatic repulsions, as shown in Figure 1.13.[106]

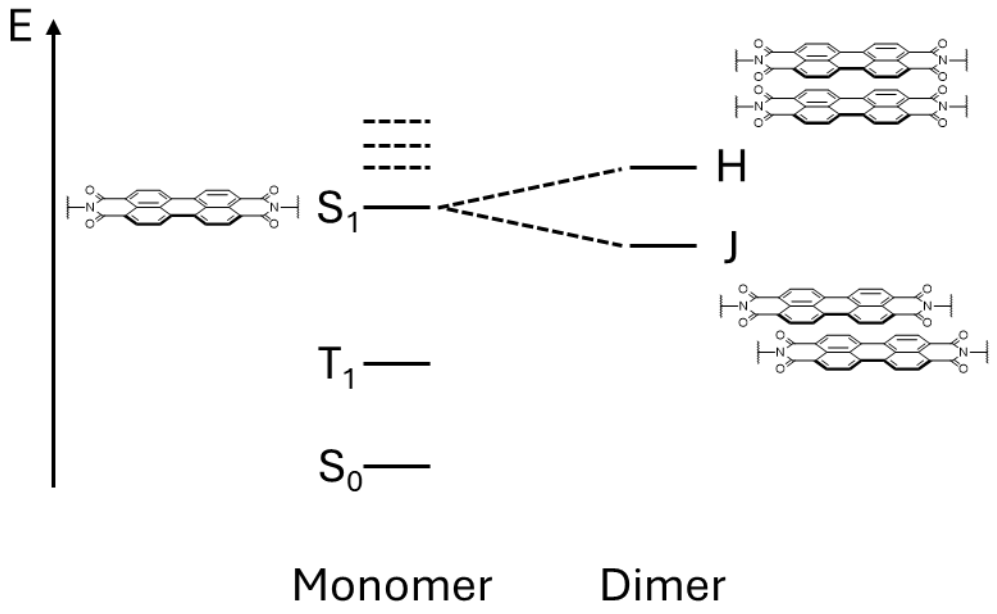


Figure 1.13. Simplified Jablonski diagram showing the energy levels of singlet and triplet states of parent PBI chromophore, including the *H*- and *J*-coupled states of a typical π-stacked dimer aggregate.[103]

The choice of solvent used to facilitate self-assembly impacts the molecular stacking.[107] While stacking, the PBI and NDI molecules can be selectively oriented to form *H*-type or J-type aggregates, depending on the solvent. Yang *et al*. demonstrated that NDIs self-assemble into *J* aggregated structures when precipitated from polar solvents, while *H* aggregated structures form when precipitated from non-polar solvents.[107] Selectively preparing the desired aggregate is important, as significant differences in the optoelectronic properties of each aggregate will lead to substantial changes in the material's properties. For example, the photoconductivity of self-assembled PBIs is dependent on the charge carrier mobility and exciton diffusion length, both of which are superior in *J*-type aggregated structures when compared to *H*-type aggregated structures.[108] This makes the ability to predict the predominant aggregation type critical to optimising the electronic capabilities of organic material a worthwhile challenge to overcome.

1.2.5 Tunability of Rylene Materials

Rylenes can be functionalised at different positions to target different properties of the resulting material. The rylene core remains electronically isolated from any substituents attached to the imide nitrogen. Therefore, the imide substituent has little effect on the position of the lowest-energy transition. This allows the rylene core to be highly functionalized without inhibiting its optoelectronic properties.[103] Subsequently, imide functionalisation is ideal for influencing the self-assembled nanostructures without

disrupting the optoelectronic benefits afforded by the aromatic core.[109] Limiting the investigation solely to imide functionality restricts the solubility, reactivity, and supramolecular assemblies that can be achieved with the target materials.[110] Functionalisation of the rylene core can tune properties such as the electronics and optical spectra.[109, 111] PBI core functionalisation at the bay position is well reported, as it is a more synthetically approachable target than at the ortho position.[109] However, steric interactions between a bay substitution and the adjacent hydrogen can force a distortion in the perylene core.[110] By twisting the core, the π-π interactions that afford self-assembly can be disrupted, potentially altering the resulting nanostructure.[112] Work by Egan *et al.*, however, showed that functionalising the bay position of PBI with a pyrrolidine group significantly reduced the band gap and extended the absorption range into the visible spectrum, which is critical for photovoltaic applications.[113] Unlike other core-substituted PBIs, which can only form dimers, the pyrrolidine functionalisation did not fully inhibit the formation of an extended supramolecular network.[112] However, compared to the unsubstituted bay analogues, significant changes in the nanostructure were observed.[113] While reports focus on the impact of imide functionalisation on self-assembly, it's clear that core functionalisation can have just as much impact. This provides an additional opportunity for greater control of these systems, which is yet to be fully realised. Therefore, to fully optimise and tune materials for their desired application, better design rules and prediction tools are required to understand the link between such functionalisation and both the self-assembled structures and bulk properties.

Not only do structural changes influence the self-assembly of rylenes, but so too does the solvent in which the materials are dissolved to afford aggregation. Chen *et al*. have demonstrated that the degree of formation of self-assembled structures of functionalised PBIs depends not only on the sided chain, but also on the solvent in which the material is dissolved.[114] Aggregation constants, a measure of binding strength between perylene monomers in a given solvent, were determined across 17 solvents of varying dipolarity, polarisability, and hydrogen bonding capacity by concentration-dependent UV-vis absorption analysis (Figure 1.14). It was shown that the amount of π-π interactions varies significantly with changing solvents. By modulating these interactions, the degree of aggregation could be tuned by using select solvents, with lower aggregation constants observed as polarity is increased for aliphatic and dipolar aprotic solvents, and increased aggregation constants observed as polarity is increased for dipolar protic solvents. It was

thereby shown that the aggregation of PBIs is dominated by electrostatic interactions in nonpolar solvents and by solvophobic interactions in protic solvents.

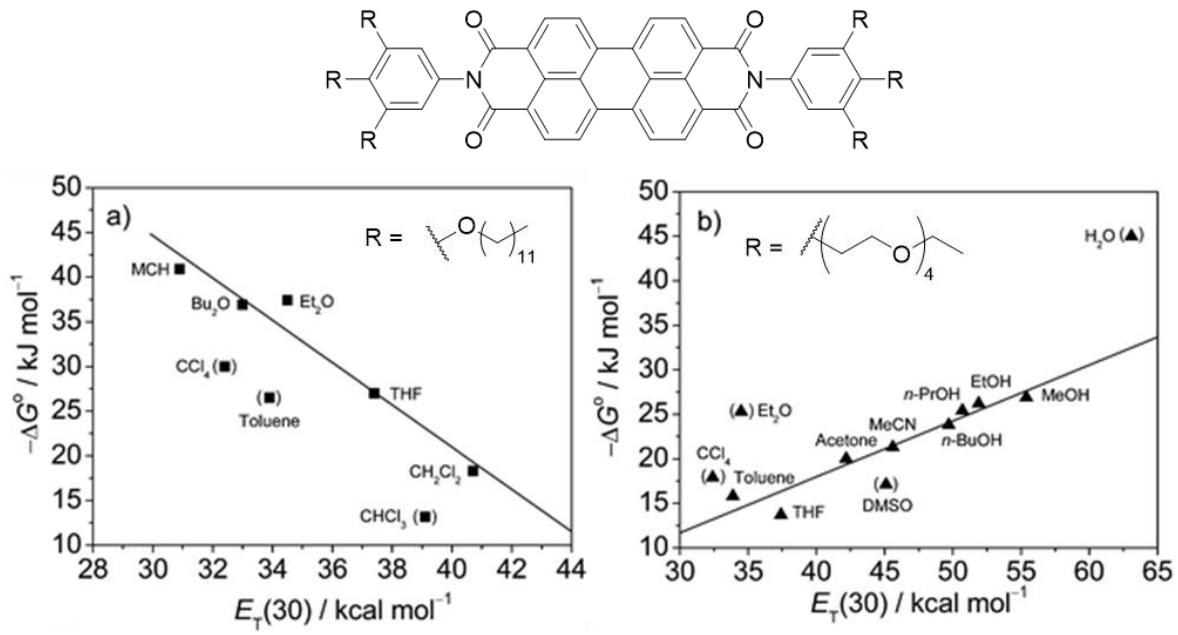


Figure 1.14. Plot of −ΔG° vs. the $E_T(30)$ solvent polarity scale. General chemical structure is provided where a) represents the data for molecule with R = $O(CH_2)_{11}CH_3$, and b) represents the data for molecule with R = $(CH_2CH_2O)_4CH_2CH_3$. Note that for methylcyclohexane (MCH), no $E_T(30)$ value is available; the cyclohexane value was used instead in a). Adapted from Chen *et al*.[114]

Solvents can affect both the degree of aggregation, and importantly, the topology of aggregation. Naranjo *et al.* have demonstrated how solvent choice impacts the behaviour of N-annulated perylene bisimides. In highly solvating solvents like chloroform, the PBI core tends to form dimers. When using the less polar methylcyclohexane, the process shifts to anti-cooperative supramolecular polymerisation, where dimers assemble into larger oligomers, illustrating the role of solvent–solute interactions, driven by the hydrophobic effect described in Section 1.2.2. These differences in assembly affect the resulting film characteristics, with those prepared from chloroform leading to small nanoparticles, whereas methylcyclohexane produces longer, rope-like fibres from a single molecule.[115] Such solvent-induced morphological changes were demonstrated by Chen *et al*., who showed how the choice of solvent could afford long 1D crystalline nanobelts in methanol, while the same compound would assemble nanoleaves in n-hexane. It was further demonstrated how the significant changes to the morphology directly impact the conductive ability of films prepared from each solution (Figure 1.15). By tuning the morphology to favour these nanobelt structures, the resulting highly ordered crystalline assembly enabled increased charge transport, affording conductivity 3 orders of magnitude greater than that of other

polymer systems.[116] These works illustrate the adaptability and tunability of self-assembling rylenes, with vastly different topologies and properties resulting from a single molecule.

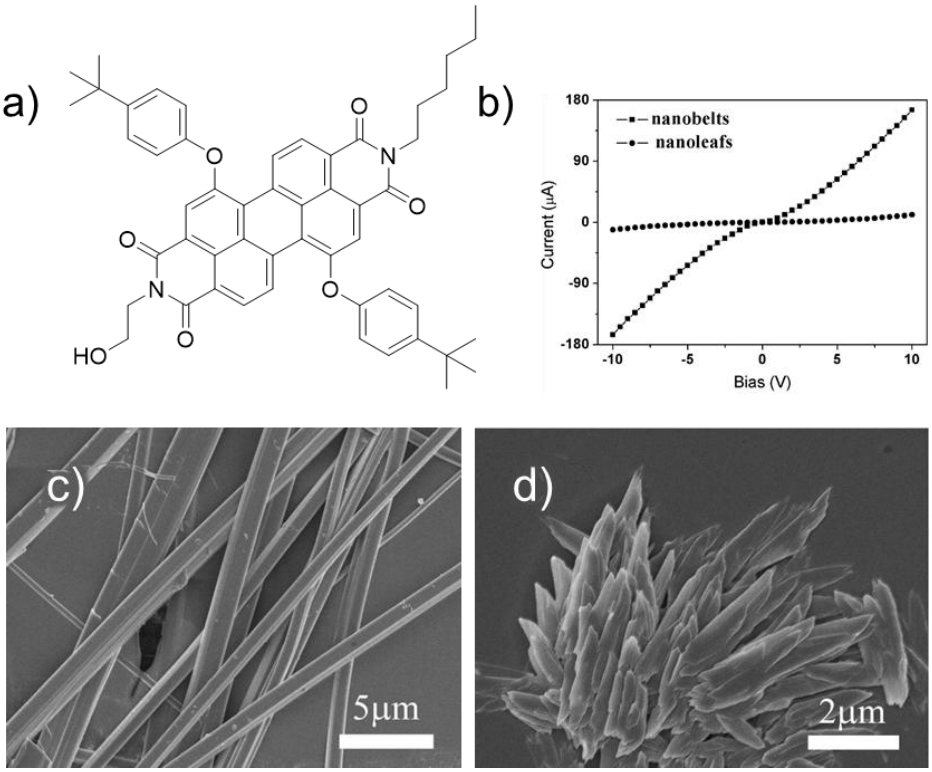


Figure 1.15. a) Chemical structure of functionalised PBI. b) I–V curves measured on nanobelts (solid squares) and nanoleaves (solid circles). c-d) Scanning electron microscopy images of the nanobelts from methanol and nanoleaves from n-hexane, respectively. Reproduced from Chen *et al*.[116]

The installation of amino acids at the imide position has been extensively investigated within the Draper group.[117-120] The addition of amino acids provides a synthetically accessible method of tuning the structure with many different side chains of varying hydrophobicity available, allowing the intrinsic solubility to be finely tuned with different aliphatic side chains or enhanced intramolecular interactions with aromatic moieties. By adding ionisable carboxylic acids, the highly aromatic, apolar rylenes can be more readily solubilised in polar solvents, enabling self-assembly in polar organic and aqueous systems. Furthermore, the pH-dependent ionisation of such groups enables external tuning to create a range of materials, such as gels and films. Recently, Adams *et al*. showed how the conductive properties of films made from alanine-functionalised PBI (PBI-A) exhibited mechanoresponsive changes in conductivity upon the application of a bending force.[121] It was shown that a simple change in amino acid functionalisation facilitated the mechanoresponsive effect. Films were cast using PBI-A, as well as phenylalanine- and histidine-functionalised PBI (PBI-F and PBI-H, respectively). PBI-F failed to cast films effectively, potentially due to an additional hydrophobic contribution from the phenylalanine functionalisation, which disrupted the

formation of large, self-assembled micellar structures. PBI-H was capable of forming conductive films, but upon bending, the film cracked, causing conductivity to cease by disrupting the continuous pathway. Conversely, PBI-A could withstand bending forces while decreasing current across the film in proportion to the degree of bending. The nanofibres formed in the PBI-A films were shown to be significantly longer than those formed by PBI-H. It was suggested that the longer fibres of PBI-A act like layers upon bending, creating longer conductive pathways.[121] The work by Adams *et al*. illustrates the connection between micelle topology resulting from amino acid functionalisation and the ability to exhibit mechanoresponsive conductivity. Further development of materials with desirable properties, such as mechanoresponsive conductivity, would be aided if these properties could be predicted rather than discovered serendipitously and later rationalised.

1.3 Predicting Self-Assembly

As discussed above, varying nanostructures afforded by molecular self-assembly using different chemical-scale and preparation modifications can significantly affect the performance of self-assembled materials.[118] Being able to link the chemical structure to the desired material and responsive properties is key to enabling the rational design of smart materials. Currently, there are very few reliable methods to accurately predict the effect of the multiscale interactions that give rise to a supramolecular material's unique properties.

1.3.1 Existing Techniques

Many methods have been used previously to model, understand, and predict the outcome of self-assembly in soft-matter systems with varying success. For example, critical packing parameters, density functional theory, and molecular dynamics have been used to tackle the limited understanding of molecular self-assembly. Each approach has its own advantages and disadvantages.

1.3.1.1 Critical Packing Parameter

The rationalisation of why surfactant molecules self-assemble into spherical micelles, cylindrical micelles, or spherical bilayer vesicles was a fundamental problem in early surfactant theories. Early approaches to predict packing considered the geometric and packing constraints afforded by the molecular structure. By quantifying the molecular volume of a surfactant (V), the head group area (a_0) and chain length (l_0), the geometry of aggregation can be related to the critical packing parameter (CPP) (Equation 1.4, Figure 1.16).

$$CPP = v/a_0 l_c \quad \text{(Equation 1.4)}$$

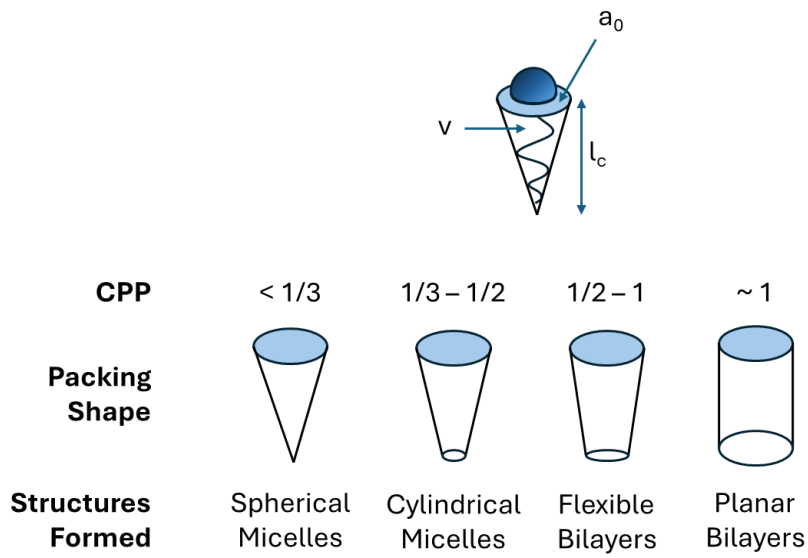


Figure 1.16. Packing shape and the larger structure topologies are formed for a given critical packing parameter based on the volume, head group area, and chain length of a surfactant.[122]

The dimensionless CPP quantity represents the geometric balance of hydrophilic and hydrophobic moieties at the interface of self-assembled structures. Such a simplified representation of assembly neglects many of the nuances which control self-assembly. It has been noted that the CPP parameter ignores the intermolecular interactions provided by the surfactant tail, leading to ineffective predictions when used alone.[123] Moreover, self-assembly is highly dependent on the chemical environment. While CPP is often used to explain self-assembly in aqueous environments, it becomes less relevant when other factors unrelated to the surfactant geometry dominate.[124] While it can be an important parameter for determining the expected type of self-assembled structures, the numerous thermodynamic factors that influence them can make the calculation of CPP from molecular structure alone unattainable. Experimental estimates of the CPP often deviate from the theoretical values, necessitating reliance on experimental observations to infer the structure.[125] Subsequently, calculating the critical packing parameter typically requires insight into molecular shape and self-assembly, making direct application to molecular design difficult.[125]

1.3.1.2 Quantum Mechanical Methods

Quantum mechanical (QM) calculations at different levels of accuracy can provide information about supramolecular assemblies at various length and time scales. QM techniques generally yield highly accurate properties; however, they also require significant computational power. Subsequently, QM calculations used to model supramolecular

materials are often limited to static or short dynamic simulations of single molecules or systems of up to ~500 heavy atoms.[126] It is well-known that DFT simulations are not well-suited to modelling long-range interactions, which are crucial for accurately modelling non-covalent interactions in supramolecular assemblies.[127] Despite its limitations, DFT methods have shown that they can provide valuable insights into supramolecular behaviour when modelling simplified assemblies of only a few molecules.

One of the earliest examples of DFT being used to probe supramolecular stacking was reported by Setnička *et al*. They used DFT to compute stable conformations of a simplified model of their self-assembling guanine derivative. This allowed them to calculate infrared and vibrational circular dichroism spectra for the most stable conformations of dimers to tetramers. While these calculations alone are insufficient to elucidate the aggregated structures, by comparing the calculated spectra with their corresponding experimental spectra, the most likely arrangement can be identified.[128, 129] DFT has since been used in several investigations of supramolecular assemblies. For example, Woods *et al*. sought to design self-assembling materials without the strong interaction sites often used to guide self-assembly. Without such features, molecules need to be carefully designed to ensure the desired high-order structures are entropically favourable.[130] It was suspected that designing a molecule with a saddle geometry would sufficiently restrict translational freedom in the stacked assembly to induce a greater degree of order. Microscopic imaging identified the expected 2D nanosheets, but to understand the geometry of molecular assembly, DFT was utilised. As with the work by Setnička et al., it was shown that DFT alone, while able to provide some insight, requires experimentation to interpret effectively. The theoretical calculations predicted that the carpyridine tetramers preferred an alternating stacking conformation. However, the X-ray single-crystal structure revealed that the cores instead adopt a slipped arrangement. DFT identified such slipped arrangements as a local minimum but found them to be slightly higher in energy than the alternating conformation ($\Delta E_{DFT} \sim 3$ kJ mol^{-1}) (Figure 1.17). Despite its limitations, DFT was key to demonstrating that engineering molecular shape is an important consideration in the design of supramolecular systems.[130]

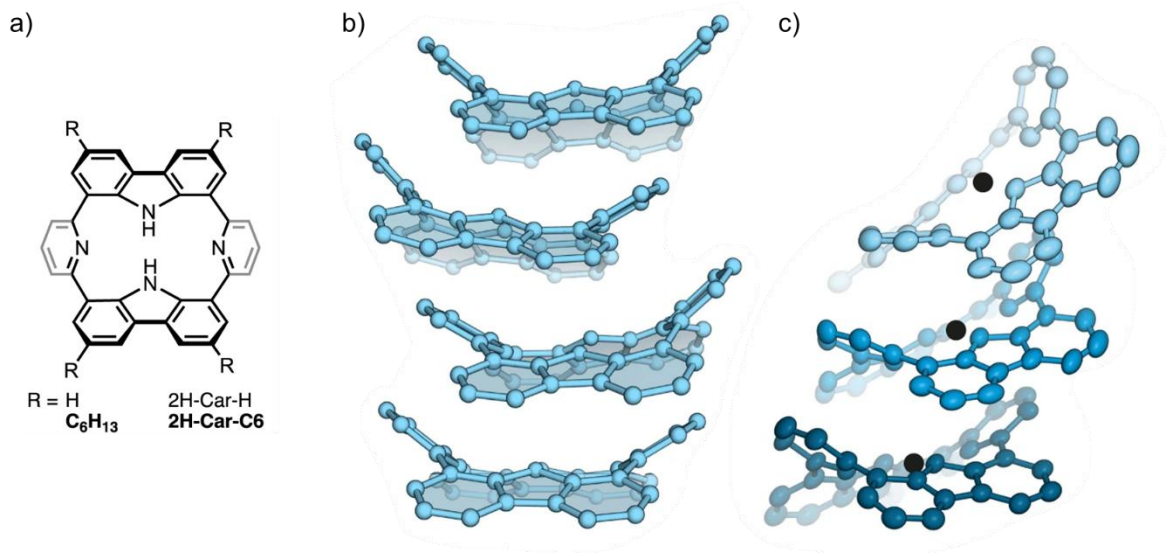


Figure 1.17. a) Generic chemical structure of carpyridine with R group modifications. b) DFT derived arrangement of a four-unit stack of the carpyridine core (2H-Car-H) showing the retention of the principal axis of curvature. c) Molecular arrangement derived from the X-ray single crystal structure of 2H-Car-C6, showing a slipped arrangement also predicted by DFT. Adapted from Woods *et al.*[130]

Draper *et al*. used DFT techniques to probe the self-assembly of alanine-functionalised perylene bisimide (PBI-A).[131] It was shown that by modelling small assemblies of up to four PBI-A molecules, one can explore how different packing motifs give rise to experimentally observed shifts in optical properties upon aggregation. It was suggested that reducing the pH did not change the mode of aggregation but did increase the size of the helically stacked aggregates. Pairing this insight with experimental results obtained from nuclear magnetic resonance (NMR) experiments supported previously proposed explanations for gelation processes. It was noted that changes in the computational setup predicted that neutral aggregates would be 4 – 40× larger than their doubly deprotonated counterparts. Such a high variance in predictions suggests that DFT is better at providing qualitative insights into molecular assemblies than at making quantitative predictions. It was also highlighted that as molecular complexity increases, manual structure determination used to inform DFT theory becomes increasingly unfeasible. Despite requiring many more DFT calculations, it was suggested that sampling molecular dynamics structures may be a better approach for finding low-energy structures.[131]

1.3.1.3 Molecular Dynamics Methods

As discussed earlier, understanding the solvation properties of small molecules is essential for understanding the morphology and properties of the resulting self-assembled structures. However, predicting solubility and the underlying thermodynamic properties that influence it are important not only in soft-matter prediction problems but also in a variety of other

chemical fields, especially pharmaceutical research.[132] Solvation free energy ΔG_{solv} is a key thermodynamic property defined as the change in Gibbs free energy when a molecule moves from the ideal gas phase to a solvent at a specified temperature and pressure. As the Gibbs free energy barrier decreases, the molecule's solubility increases. As computational hardware has improved rapidly, free energy calculations using highly optimised molecular dynamics software have become a standard tool in industry. While several molecular dynamics (MD) methods exist, these systems typically identify atoms that change between transition states and calculate the changes in carefully tailored forcefields. Ensuring that appropriate forcefields are used is critical, as they represent a pragmatic compromise between computational efficiency and physical accuracy. Classical force fields simplify the system into bonded and non-bonded contributions. Forcefields are derived from a combination of quantum-mechanical calculations on small molecular fragments and empirical fitting to reproduce experimental observables.[133] Early attempts to perform free energy calculations using MD methods suffered from an inability to model Coulombic interactions, which are key to self-assembled aggregation.[104, 132] A significant limitation of such free energy calculations is the accuracy of the necessary forcefields which often require refitting to bespoke DFT calculations for each new compound to achieve reasonable accuracy.[132] Such calculations are computationally costly, so the resulting predictions afforded by MD modelling using differing forcefields often necessitate a comparison with experimental data to identify the most appropriate choice for a given class of molecules.[134] While the commonly used forcefields (AMBER, CharmM, OPLS, and GROMOS) produce similar estimation of primary interactions afforded by bonding and non-bonding forces, each has significant differences when calculating the interactions affording secondary structures. For example, when applied to the prediction of properties in self-assembled peptides, it was found that CharmM-based force fields better estimated the conformation of the self-assembled aggregates of human islet amyloid polypeptide. In contrast, AMBER-based models more accurately estimated the chemical-scale property of carbon shift signals in nuclear magnetic resonance.[134] Therefore, given that the properties of supramolecular materials are afforded by their hierarchical nature, MD modelling across these different scales may require multiple force fields, which may be difficult to identify.

As techniques for characterising soft-matter systems improve, molecular dynamics may be better able to leverage *a priori* knowledge to better understand self-assembled structures. Diphenylalanine-functionalised naphthalene (2Nap-FF) is a well-studied self-assembling molecule, with the ability to self-assemble on the nano/micro scale in solution, ultimately

forming gels as pH is lowered. To better understand self-assembly, the nanoscale structure was investigated using small-angle scattering techniques to inform molecular dynamics simulations. In unbiased simulations, MD models were unable to recreate the hollow tubular assemblies previously observed with scattering techniques. Therefore, biased simulations using the *a priori* knowledge afforded by small angle scattering were necessary. Restricting the model to a cylindrical system yielded the experimentally observed hollow cylinder in the MD simulation, which, together with the scattering characterisation, confirmed the surfactant-like nature of the 2Nap-FF self-assembly.[135] However, the low resolution of the scattering data limited the ability to model the structure with sufficient resolution, and thus the actual packing remained undetermined.[136] The self-assembly in the solution state is shown to directly affect the structures achieved in the final material, with the ultimate structure arising from the collapse and subsequent stabilisation of the solution state morphology. Predicting and understanding the packing and self-assembly of solution-state aggregates are necessary to understand the morphology of the final material. Using cryogenic transmission electron microscopy (cryo-TEM), the highly ordered packing and helical symmetry of the 2Nap-FF self-assembly were directly observed, allowing high-resolution atomic models to be constructed. These showed that the micelle structure arises from eight protofilaments consisting of a trimer of 2Nap-FF molecules in various conformations. The cryo-TEM model corroborated previous scattering data and MD simulations; however, this technique was better able to show that packing was driven by burying the aromatic rings in the hydrophobic core.[136] While experimental interpretation of complex self-assembled structures by techniques such as high-resolution cryo-TEM may only be achieved with sufficient *a priori* knowledge, it is clear that MD and other theoretical techniques share the same requirements.[137] Subsequently, while MD can provide powerful insights into self-assembled structures, techniques that provide direct observations of properties of interest (if there is sufficient *a priori* knowledge to interpret them) may be a more reliable option than QM and thermodynamic modelling for a better understanding of self-assembled systems and their properties.

1.3.1.4 Statistical Thermodynamics Methods

Statistical thermodynamic models have been investigated as an alternative to particle-based calculations. The models treat each molecule as if it interacts with an average environment rather than tracking every individual interaction. By using statistical mechanics to calculate the most energetically favourable arrangement, the equilibrium structure of self-assembled aggregates can be determined. Molecular theory (MOLT) approaches pioneered by

Szleifer *et al*. provided a framework for modelling the average structure of self-assemblies in soft materials. [138, 139] In contrast to particle-based simulations, MOLT directly estimates the free energy of different possible structures. It can leverage the system's natural symmetries to minimise computational cost. The output of the molecular theory provides both structural and thermodynamic information, giving insight into the system's morphology and average properties. However, a significant limitation of molecular theory is its inability to provide information on the kinetics of self-assembly, which can be key to understanding the bulk properties of interest.[138] While progress has been made in the applications of MOLT to out-of-equilibrium systems, it remains an ongoing challenge in its development. While probing the equilibrium structure can provide important insights, it is important to note that many self-assembled systems can result from kinetic trapping effects. Such effects hinder the formation of thermodynamically stable structures determined by techniques such as MOLT. Therefore, to effectively characterise self-assembled aggregation, it may be necessary to use measurements that probe the material's dynamic state, rather than the idealised thermodynamic endpoint.

1.3.1.5 Limitations of Current Methods

Aside from the scientific limitations of QM and MD techniques, one significant drawback of QM models is their inaccessibility. High-performance computing and QM environments are crucial for computational research but are challenging for researchers with limited computational knowledge to utilise. This makes *in silico* prediction inaccessible to many wet-lab scientists. Newer software packages, such as the Amsterdam Modelling Suite[140] (AMS) has been developed with intuitive user accessibility in mind. Such systems are necessary to avoid making QM modelling prohibitively complex, giving the typical wet-lab chemist access to important predictive tools such as the COnductor-like Screening MOdel for Real Solvents (COSMO-RS). COSMO-RS was developed to predict thermophysical data for liquid systems, enabling the efficient calculation of many properties that other methods cannot handle. It has become an effective alternative to force field-based molecular simulation methods.[141] This allows thermodynamic properties key to molecular self-assembly, such as solubility and pK_a, to be predicted. However, to predict solubility using COSMO-RS with AMS, some physical characteristics of the molecule must already be known, including the melting point. For high-melting-point solids, accurate melting points may be difficult to obtain. Being unable to accurately characterise melting points reduces the reliability of predictions provided by COSMO-RS. As discussed in section 1.2.4, we are interested in the properties of compounds such as amino acid functionalised rylenes. While

the installation of the hydrophilic and ionisable carboxylic acid moiety imparts many useful characteristics, increasing hydrophilicity can make the molecule hygroscopic. Additionally, compounds with high melting points may decompose chemically before reaching their melting point. These together can make it difficult to quantify the melting point of compounds such as amino acid-functionalised rylenes. Such limitations can be observed in the differential scanning calorimetry (DSC) of leucine-functionalised naphthalene diimide. DSC shows the occurrence of multiple exothermic events, which may or may not relate to the evaporation of absorbed water, melting, or chemical breakdown (Figure 1.18).

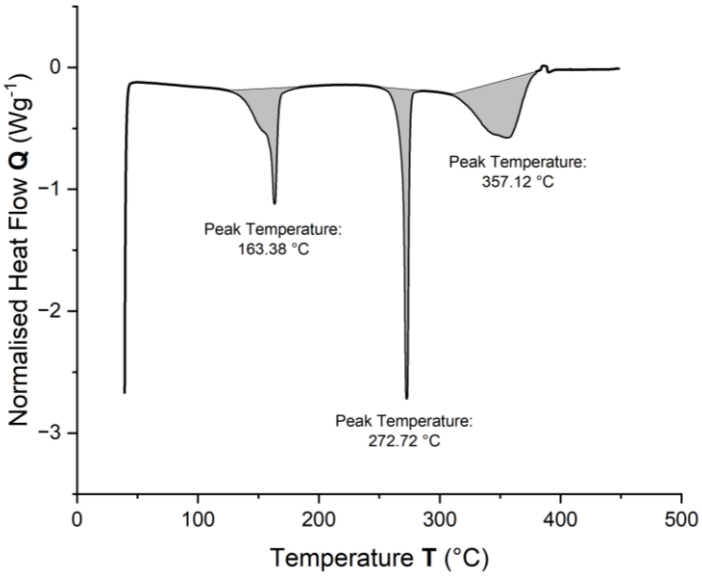


Figure 1.18. Differential scanning calorimetry (DSC) of leucine functionalised naphthalene diimide. DSC shows the occurrence of multiple exothermic events across a wide temperature range. Data was captured for this project using a TA Instruments DSC 25 with RCS90 refrigerated cooling system.

Given the inadequate quantification of necessary physical characteristics and the limitations of theoretical models, in many cases, the direct measurement of the property of interest may be required. However, if one wishes to predict such properties before synthesis, then predictive systems derived from direct characterisations of the property, rather than theorised predictions afforded by QM methods, may be a better alternative.

1.3.2 Machine Learning

Empirically elucidating an unknown relationship between inputs and outputs of a system can be difficult to achieve, especially in fields as complex as chemistry. Machine learning (ML) uses algorithms to analyse complex data and predict the state of new data. The development of these powerful tools has gained significant interest in the field of chemistry, especially in pharmaceutical discovery, where the quantity of available information is unfathomably vast.[142] While machine learning models can be classified by the types of models produced

or the problems they are used to solve, the traditional categorisations are reinforcement, unsupervised, and supervised learning (Figure 1.19).[143] As ML methods continue to be developed, they do not all fall into these categories; however, they all share the goal of generating models that formalise data and solve problems for which identifying distinct and critical features of the input data can be difficult.

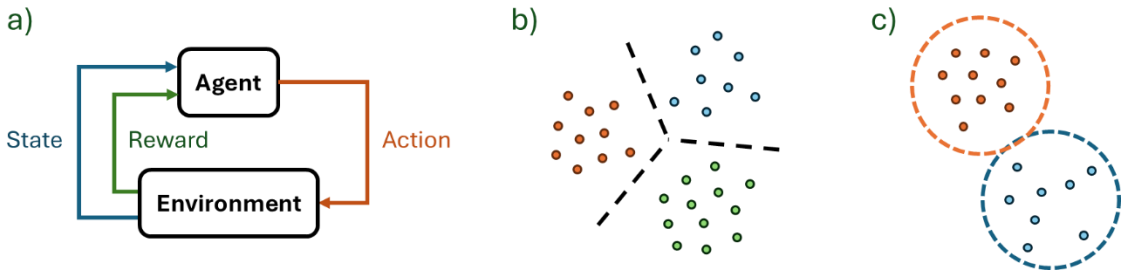


Figure 1.19. Cartoons illustrating a) reinforcement learning, b) unsupervised learning, and c) supervised learning.

Choosing a model depends on the data available for training the algorithm. When datasets are very large but lack output data, reinforcement learning mimics human learning by learning from actions and outcomes, improving through feedback. If only input data exists, unsupervised learning analyses clusters and identifies trends within the dataset. If both input and output data are accessible, supervised learning is used to create an algorithm that learns a correlation between them.[144]

Reinforcement learning (RL) uses models that learn how changes to a system or environment affect an objective, enabling it to maximise a reward. Basic reinforcement learning is modelled as a Markov decision process (MDP) where the model seeks useful experience about states, actions, transitions, and rewards to learn how to behave optimally.[145, 146] In the context of building models capable of predicting which chemical structures exhibit the properties of interest, these features can be thought of as:

- States: the possible chemical structures
- Action: modifications to the chemical structure
- Transition: the probability of obtaining a new structure after an action
- Reward: property-based evaluation of the resulting structure

The reward function is an important part of the MDP that implicitly specifies the learning goal. The function guides how the system should be controlled.

Despite the successful application of reinforcement learning models across a variety of prediction problems, there are significant fundamental limitations that make it a poor choice

for tackling chemical prediction. Firstly, designing an appropriate reward function can be a non-trivial task that requires considerable expertise, and incorrect reward signals can lead to suboptimal predictions.[147] To generate effective models, RL requires a substantial number of interactions to identify effective actions and, subsequently, an unachievable large number of states (or, in the case of chemical prediction, compounds) to make adequate predictions.[147] Lastly, RL performance is significantly reduced as the number of available actions increases.[147] Given that even for small molecules, there are hundreds of potential chemical modifications, it can be challenging for RL models to explore and learn chemical space efficiently. Even if an effective model for known data is found, RL struggle to generalise their experience to new environments, which may make predicting unknown compounds difficult and ineffective.[148]

Unsupervised learning (UL) uses models which identify patterns in datasets for which no classification or labels are provided.[143, 149] These models can identify noisy and irrelevant features which are detrimental to classification. The remaining labels are used to cluster the data by identifying high-density regions separated by relatively low-density volumes, which may correspond to meaningful subgroups within the dataset.[149] However, interpreting these clusters requires substantial domain expertise, as the algorithm itself provides no semantic meaning for the groups it discovers. Assigning chemical relevance to a cluster can therefore be challenging, particularly when the underlying physicochemical factors are complex or poorly understood. This reliance on expert interpretation, combined with the difficulty of mapping unlabelled clusters to specific chemical properties, limits the usefulness of UL for predicting chemical properties. As a result, UL methods are generally not well-suited for tasks that require explicit, quantitative prediction of molecular properties.

In contrast to unsupervised learning, supervised learning (SL) builds models to map the input features to known outputs. The output labels can be discrete values for classification models or continuous values for regression models.[143] However, the no-free-lunch theorems of supervised learning indicate that there is no one-size-fits-all learning algorithm that works best on all problems.[150] There is a wide range of SL models, each with its own strengths and weaknesses. SL models were used extensively in this project, serving as the basis for our ML system. Subsequently, Chapter 2 provides an in-depth discussion of SL models and their application to this work.

1.3.2.1 Machine Learning in Material Chemistry

With the advent of tools such as ChatGPT,[151] Copilot,[152] and Gemini,[153] the terms machine learning (ML) and large language models (LLMs) have become somewhat synonymous with artificial intelligence (AI). While closely connected, these are independent concepts which should be disambiguated. Artificial intelligence encompasses a broad field of technologies that build machines capable of mimicking human cognitive reasoning. Machine learning is the application of AI that enables the extraction of knowledge from data, which can be used to learn new information about a given problem. Large language models are a specific type of ML system that are trained on immense datasets of text to generate human language, predicting the next most probable token (a fundamental unit of text) in a sequence to create coherent text that can answer questions, among other useful tasks. Given the success of LLMs in predicting human language, which provide convincing answers to a wide range of questions, there is a naïve assumption that LLM systems can be applied to problems in chemistry. As the popularity of LLM tools such as ChatGPT continues to grow, there has been an increasing number of investigations into the viability of applying such systems to chemical prediction problems, with varying success. Recently, Runcie *et al.* investigated the ability of different LLM models with varying levels of reasoning and computational time to intuit and understand chemical structures.[154] While it has been shown that LLMs can now directly solve advanced chemistry problems without external tools, current reasoning models are still far from perfect. It was observed that allowing the model to reason for longer allowed for improved performance across eight distinct chemical reasoning tasks (Figure 1.20).[154]

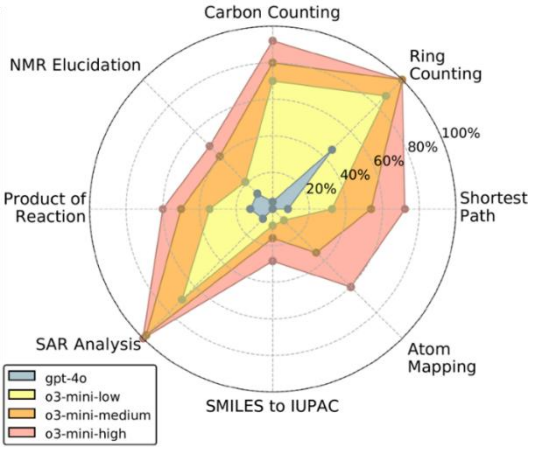


Figure 1.20. Radar plot summarising the performance of LLM across eight chemical problems, assessing each model's capacity for molecular comprehension and chemical reasoning. Results show a trend that increased reasoning correlates with higher success rates. [154]

The best-performing model, o3-mini-high, only achieved 100% accuracy on ring counting, a relatively simple task, and on basic structure-activity relationship analysis, which primarily involves mathematical reasoning rather than chemical insight. It is not clear whether, if current models were allowed to reason for longer, performance on the remaining problems would continue to improve.[154] While the results of this investigation indicate that the advancements in LLMs could give rise to general-purpose chemical reasoning tools, it is important to consider the demands of such models and the economic and environmental implications of relying on LLMs. Given the well-established negative impacts of greenhouse gas (GHG) emissions on the environment, many companies and governments have committed to significantly reducing GHG emissions.[155] However, the intense demands of LLMs necessitate the construction of resource-intensive data centres, which are incompatible with such goals. For example, in 2021, Google set a target to reduce its GHG emissions by 50% from the 2019 baseline by 2030. In its 2024 environmental report, Google revealed that its GHG emissions have increased by 48% since 2019, primarily due to higher energy consumption in data centres.[156] Not only do these facilities require immense power, but they also consume an ever-growing quantity of water to operate. Lower estimates indicate that global AI's water use is projected to reach 4.2–6.6 billion cubic meters in 2027, more than half of the United Kingdom's total annual water withdrawal.[157] In addition to this, the lack of transparency in reporting the emission impacts of LLM model training makes it difficult to assess the environmental cost of such developments. The current trend toward reduced transparency regarding AI's environmental impact contributes to misinformation and hinders informed decision-making at all levels, from individual researchers and developers to organisations and policymakers.[158] As AI becomes cheaper and faster, total usage expands, intensifying environmental strain despite the increased efficiency. Without structural shifts in how LLMs are designed, deployed, and used, these invisible costs will continue to rise, threatening to offset the societal benefits that made these systems valuable in the first place.[159] Given that current worldwide policies already fail to meet the environmental target of avoiding warming of greater than 1.5 ºC above pre-industrial levels set out by the Paris agreement (Figure 1.21), it is clear that these environmentally demanding technologies need to be developed and used more appropriately. With such limitations in mind, we must consider more efficient, economical, and targeted ML systems when tackling chemical prediction problems. While systems such as GPT-3 are estimated to use 1.3 GWh during the training phase (and over 0.5 GWh per day to fulfil the millions of requests made to ChatGPT[151]), we estimate using Codecarbon[160] that the models prepared for this project required only approximately 50 kWh.[161]

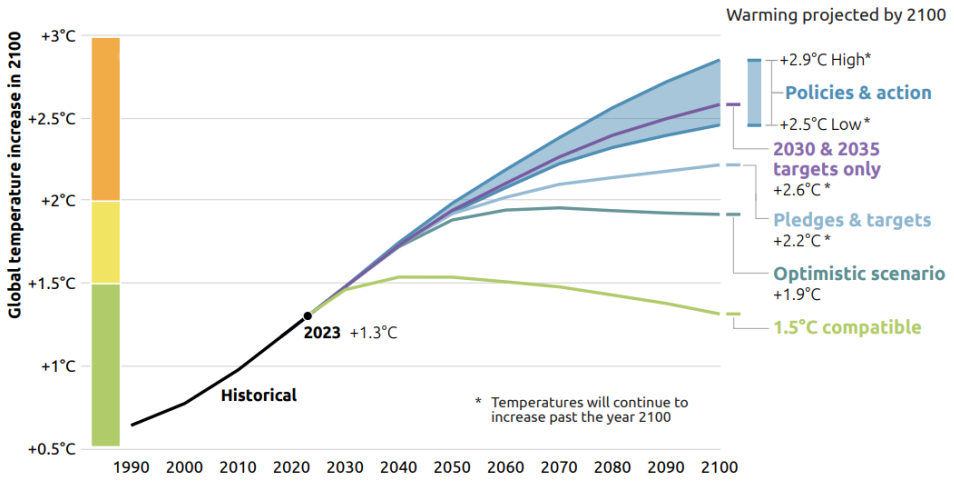


Figure 1.21. Projected emissions and expected warming based on pledges and current policies. Adapted from "Warming Projections Global Update", Clim. Anal., © 2025, Climate Analytics and NewClimate Institute.[162]

Limitations in material discovery and prediction are not confined to the field of organic materials chemistry. The discovery of energetically favourable inorganic crystals is important to the development of several applications. Due to the high costs associated with synthetic complications, the experimental determination of crystal structures remains a time-consuming endeavour, with only 20,000 of the over 200,000 entries in the Inorganic Crystal Structure Database of computationally stable structures being catalogued. Such limitations make it very difficult to discover new candidate crystal structures using traditional methods. The space of possible materials is far too large to sample meaningfully without relying on prior design rules, which biases the types of materials investigated. However, using graph neural network (GNN) machine learning techniques trained on known structures, Merchant *et al*. have developed the first models capable of accurately predicting crystal structure stability.[163] Structural candidates generated by modifying the available crystal structures to produce over 10^9 possible crystal compositions, far more than could be screened conventionally. Using the GNN, these candidates were screened for those with adequate stability (decomposition energy), expanding the possible stable structures by an order of magnitude. One hundred compounds were randomly selected, and the structure stability was computationally predicted using DFT calculations. This model predicted computationally stable crystals with a hit rate of 33%, which is much higher than the typical 1% in previous work. Furthermore, at the time of publication, 736 of the computed structures had already been independently synthesised. Of note was the model's ability to make out-of-distribution predictions, providing insight into viable candidate classes omitted from the training data. This highlights the capacity for GNN and other ML systems to drive material discovery by overcoming the limitations of relying solely on traditional chemical intuition.[163]

Much of the investigation and application of smaller-scale machine learning systems in soft-matter material chemistry has focused on enabling faster, cheaper approaches to determining quantum-mechanical parameters of self-assembly. As performing QM calculations to investigate molecular self-assembly can be expensive and time-consuming, it has become an attractive target for efficient machine learning models. Nonetheless, these calculations provide valuable insights, enabling rapid prediction of key parameters typically obtained from QM methods, thereby greatly accelerating research on supramolecular materials. Van Teijlingen *et al.* examined the use of QSPR in predicting the aggregation propensity (AP) scores of short peptides. Many peptides self-assemble in water, forming a variety of structures. Although short (di-to-octa-) peptides are synthetically accessible, it is impossible to experimentally determine each sequence's propensity to self-assemble due to the exponential growth in the number of compounds as peptide length increases (4×10^2 dipeptides to 2.56×10^{10} octapeptides). The AP score was previously recognised as a useful predictor of self-assembly, so it was selected as the target parameter for modelling.[164] The AP score is a parameter derived from MD simulations, calculated from the solvent accessible surface area (SASA) of the peptide in both the initial and final configurations of the simulation (Equation 1.5).

$$AP = \frac{SASA_{initial}}{SASA_{final}} \qquad \text{(Equation 1.5)}$$

Using an active supervised learning approach, support vector machine models iteratively learned how to identify peptides with high AP scores. By validating the model performance with the already modelled tripeptides, it was shown that after 15 iterations of the learning process, 35 of the 39 top-performing tripeptides were identified. The 15 model iterations produced outputs for 150 compounds, and MD modelling was used to determine the 35 highest-scoring candidates. Compared with the 8,000 simulations required to map chemical space fully, this represents a >50× acceleration in the discovery process.[165] It was highlighted that, given the size of the compounds being represented, descriptors can become just as prohibitively expensive to calculate. Therefore, low-resolution and low-cost descriptors were used to pre-screen the immense number of candidate molecules, before predicting the property of interest using more computationally expensive descriptors.[165] Although the method was shown to be effective at determining the AP score, such predictive models may have limited applicability in predicting the actual endpoint of interest. Using the calculated endpoint means the model is effective at predicting the QM simulation output but does not necessarily indicate whether the high-AP-scoring peptides do indeed self-assemble, as any

method-based biases will influence the results. Subsequently, basing predictive models on experimental endpoints may be more effective at affording true property prediction.

While machine learning studies on predicting inorganic material properties have been effective, the hierarchical nature of supramolecular materials makes predicting bulk properties from molecular structure a nontrivial task. Despite that, Inokuchi *et al*. were able to utilise ML models to correlate dissipative particle dynamics (DPD) simulations to modelled bulk properties of surfactant systems.[166] Using coarse-grained molecular simulations, surfactants can be reduced to models of molecular beads of varying hydrophobicity connected by harmonic springs.[166] From these models, the bulk dispersion and viscosity properties have been reliably estimated.[167] Machine learning models were trained to model the relationship between the structure of the simplified bead model and the calculated bulk properties with high accuracy. These models identified that the design of high-performance surfactants (high dispersion and low viscosity) requires hydrophilic beads, suggesting that surfactants containing multiple amine or phosphate groups will exhibit high-performance properties. While this crucial study shows that machine learning has the potential to address the lack of predictive rules between molecular-scale inputs and bulk properties, it also has major limitations. By oversimplifying the input structure, significant molecular information is lost. Subsequently, while such systems can hint at molecular design rules, direct correlations between molecular structures and properties cannot be achieved. Therefore, important nuances and insights may not be well captured by such models. Furthermore, to build this model, the authors had to assume that "the molecular structure data of the surfactant, and the output dispersion and viscosity, which were obtained from simulation, [are] correct".[166] This assumption highlights the tenuous nature of using calculated endpoints to predict properties of interest. The resulting models may or may not provide truly predictive insights and design rules, but without experimental validation, one must simply assume they are correct.

1.4 Quantitative Structure Property Relationships

1.4.1 Fundamental Principles of QSPRs

Quantitative structure-activity relationships (QSARs) utilise molecular descriptors, data closely related to molecular structure, to make predictions by quantitatively describing biological activity. The development of QSARs is closely related to the history of molecular structure theories, in which toxicologists sought to quantify the relationship between chemical structure and acute toxic effects. Early in the development of QSAR theory, it was observed that a relationship existed between the toxicity of primary aliphatic alcohols and their water solubility. It was known that solubility is determined by chemical structure, thereby demonstrating a link between structures and biological activity.[168]

Crum-Brown and Fraser later developed one of the earliest formalisations of a QSAR derived from the proposed correlation between biological activity and the molecular constitution of alkaloids, as shown in (Equation 1.6. Physiological activity in a biological system (ϕ) is defined by a function (f) of its chemical constitution (C). From this relationship, it was proposed that a change in the chemical constitution, ΔC, will affect a change in the physiological activity, $\Delta\phi$.[169]

$$\phi = f(C) \qquad \text{(Equation 1.6)}$$

It was suggested that with sufficient variations of the chemical constitution, an approximate solution for the expected activity in every case could be obtained.[169] This concept has seen significant development in the following decades, with applications and models covering a wide range of chemistries, most notably in the development of quantitative structure-property relationships (QSPRs). QSPRs utilise molecular descriptors, features representative of the molecular structure, to draw inferences about the properties of a compound, rather than biological activity in QSARs. These models can not only reproduce experimental data but also extrapolate them to provide reliable estimates of the same properties for an unknown compound.[168]

As the interest in these methods grew, particularly in the pharmaceutical field where the lead candidates are too numerous to characterise directly, regulatory agencies sought to develop guidelines and acceptability criteria to increase confidence in the application of QSARs to drug development, resulting in the following recommendations for acceptable QSAR/QSPRs by the Organisation for Economic Co-operation and Development (OECD):[170, 171]

- **A defined endpoint**
  - Clear scientific and regulatory purposes
  - Transparency of the underlying experimental data
  - Quality of the underlying experimental data
- **Unambiguous algorithm**
  - Description of the algorithm and/or software
  - Inputs and other options
  - Model accessibility
- **A defined domain of applicability**
  - Clear definition of the applicability domain and limitations of the model
- **Appropriate measures of goodness-of-fit, robustness and predictivity**
  - Goodness-of-fit, robustness and predictivity
- **Mechanistic interpretation**
  - Plausibility of the mechanistic interpretation

The development of pharmaceutical therapies is an expensive and time-consuming process. To advance a single compound to market, hundreds of thousands of compounds must be synthesised and screened, collectively costing US pharmaceutical companies $39.4 billion in 2005 and increasing.[172] Several computational techniques have been utilised across the pharmaceutical industry to minimise the cost and lead time of drug development. QSARs have already proven effective in this regard, with their use assisting in the discovery of both Merck's antibacterial Norfloxacin and Eisai's anti-Alzheimer's treatment Donepezil.[172] Just as QSARs have been instrumental in focusing pharmaceutical investigations on more promising candidates, so too could QSPRs be utilised to design optimised supramolecular systems.

QSPRs utilise molecular descriptors, data closely related to molecular structure, to make predictions by quantitatively describing properties of interest. Molecular descriptors for QSPRs may include data as simple as counts of hydrogen-bond donors and acceptors, as well as molecular mass. Such descriptors are crucial in chemical prediction problems, such as drug design, as evidenced by their inclusion in Lipinski's rule of five, a method for predicting oral bioavailability that uses easy-to-compute descriptors of candidate molecules.[173] There are many steps involved in designing a robust and effective QSPR system. An important consideration in building a QSPR model is ensuring that the data is accurate and of high-quality. One of the major issues plaguing QSPR datasets is the prevalence of incorrect chemical structure representations.[174] Therefore, before modelling

can even begin, high-quality data must be obtained from well-documented and reliable sources. This data is then processed to obtain property labels and descriptors for each input, which are used in the model's training. Once a model is built, various methods can be used to assess its effectiveness on both the training data used to develop the model and on independent data, ensuring that truly predictive models have been identified. These effective models can then be probed to identify the key molecular descriptors that visualise how the chemical structure affects the modelled property. Ultimately, such models can be deployed to predict the properties of candidate molecules. This would narrow the scope of experimental investigations to those molecules more likely to exhibit the desired properties, enabling them to be rationally designed (Figure 1.22). [168, 175]



Figure 1.22. General steps central to building effective QSPRs

The development of a QSPR framework capable of tackling chemical classification processes was central to this project. Subsequently, Chapter 2 presents a focused discussion of the methods and techniques used to build the models in this project.

1.4.2 QSPRs in Supramolecular Applications

QSPRs are trained to address specific chemical prediction problems, but there is no one-size-fits-all approach to finding the most efficient model. However, previously developed QSPRs offer insight into what QSPRs can achieve in material chemistry. To date, there has been limited investigation into the development of QSPRs to predict the properties of supramolecular materials, with no known examples for rylene property prediction.

While many applications of machine learning in material chemistry aim to predict the outcome of QM modelling, using the QM characteristics of the molecules as descriptors for QSPR rather than as the target endpoint may be more insightful. In the investigation of self-assembling polymers, QM descriptors were an effective alternative to the typical structural descriptors used for small molecules. For large molecules and polymers, such descriptors are too complex to extract meaningful correlations. To overcome this, Wu *et al.* used QM methods to describe subunits of self-assembling polymers, and autocorrelated their values with those of other subunits to yield topological descriptors that encode both structural and numerical property information.[176, 177] Using these descriptors, two QSPRs were developed to predict interesting bulk properties: the drug loading capacity and critical micelle

concentration. Using multiple linear regression supervised learning models, highly interpretable, statistically significant, and well-performing models were identified. It was shown that model interpretability was essential to gaining insight into the structural parameters that govern the modelled properties. For the CMC model, it was demonstrated that descriptors related to the interaction energies between the polymer and the aqueous solvent, the number of charges, and the number of branched molecular structures were significant. However, despite identifying these key parameters, no additional independent validation of the relationship was undertaken. Therefore, Wu *et al.* showed that supramolecular properties can be modelled by QSPR, exemplifying the importance of developing interpretable systems, but failed to apply these predictions to validate the insights obtained.

To ensure QSPRs are robust and reliable, it is essential to independently validate their predictions by experimental investigation of the predicted outputs. This confirms that a genuinely predictive relationship has been identified and modelled. Gupta *et al.* used experimental validation to investigate the effectiveness of their QSPR for predicting supramolecular gelation.[178] Reliably predicting whether a molecule can yield a self-supporting gel from supramolecular assemblies enables the fabrication of uniquely tailored functional materials. As discussed previously, self-assembly is driven by thermodynamically favourable interactions, resulting in supramolecular structures that are often kinetically trapped.[92] This preparation pathway dependence makes the prediction of supramolecular gelation particularly challenging to resolve. A significant advantage of their QSPR system was the use of multiple models to make an ensemble prediction of the gelation property. Given that no single model provides a one-size-fits-all solution to these classification problems, building predictions on a consensus of models, each with its own advantages and disadvantages, allows more reliable identification of complex properties, such as gelation. Seven ML models were optimised and benchmarked on a test set of compounds with sufficient chemical similarity. Ineffective models were disregarded, leaving three models with robust predictive ability for molecules used both in and independent of the training procedure. The consensus prediction of these models was experimentally validated on nine molecules with the highest prediction confidence. The experimental validation showed that each prediction was correct, indicating that truly predictive models had been developed. Another significant advantage of the model is the use of a well-defined applicability domain to identify candidate molecules with sufficient chemical similarity to the training set, thereby increasing the likelihood that they will be well modelled by the QSPR, further providing

confidence in the predicted property.[178] Similar work by Li *et al*. involved a QSPR system built on a larger, more diverse set of dipeptides. It was highlighted that the class imbalance, that is, the proportion of molecules which do not gel compared to those that do, in the training set led to a significant decrease in predictive performance. While the application of data augmentation techniques such as the synthetic minority over-sampling technique (SMOTE) and adaptive synthetic sampling (ADASYN) was necessary to resolve the class imbalance, the precision of the predictions was still significantly poorer than the model reported by Gupta *et al*.[178, 179] This indicates that while expanding the datasets is necessary to extend the applicable chemical space of a QSPR, a robust QSPR may be inaccessible without sufficient data in each property class.

To truly design materials, it is necessary to predict whether materials can be made and, importantly, whether the resulting material exhibits the required properties. In stem cell culturing, gel stiffness affects stem cell differentiation, so different storage moduli are needed based on the specific application.[180] While the design of several independent systems could allow the necessary elasticity to be achieved, it has been shown that the well-defined, tuneable elasticity of a collagen gel system can enable the matrix to specify the lineage towards the neurons, myoblasts, or osteoblasts from identical serum conditions.[180] This work highlights how enabling the prediction of supramolecular material properties can enable the complex processes to be guided to enable the desired effects with sufficient knowledge of tuneable organic materials. Expanding on the work of Gupta *et al*., work from Simpson *et al*. recently demonstrated the first QSPR for predicting the rheological properties of peptide-based gelators.[181] Using an ensemble of ML methods, it was shown that the rheological properties of G′ and G″ (storage and loss moduli: measures of the solid-like and liquid-like character of a material, respectively) could be effectively predicted, despite QSPRs previously failing to model kinetically trapped materials, such as hydrogels. Given the complexity of such properties, it was highlighted that the use of simple, computationally cheap methods enables further iterations to be developed as new data is obtained. Using methodologies such as Shapley Additive exPlanations (SHAP), the impact of each descriptor on the modelled properties could be calculated, allowing important structural features to be identified, and a mechanistic explanation for the properties to be proposed. It was also shown that the presence of valine and isoleucine simultaneously promotes G′ while diminishing G″ in the resulting gel. This insight, currently unexplored in supramolecular literature, may enable the design of more rigid gels for new applications.[181] Relying on an intuitive understanding based on the available literature would be unable to elucidate these design

principles, which are crucial for designing and tailoring these materials. It is clear that robust machine learning tools could pave the way for the design of a new generation of functional materials.

1.5 Aims of this Thesis

Supramolecular materials are of increasing interest due to their tuneable and functional nature. These significant advantages make predicting their material properties challenging, limiting the capacity for rational design. If self-assembling materials are to be tuned and designed rather than found serendipitously, then more reliable tools are required to enable robust predictions of material properties. Such tools would narrow the necessary scope of investigation, allowing viable candidates to be targeted and better understood. To that end, this project has focused on building quantitative structure-property relationships that link small-molecule structural features to properties of interest and lay the foundation for new tools to better design functional materials for smart devices.

Chapter 2 will focus on developing my QSPR system in KNIME. Breaking down each module and how each part contributes to the system, focusing on their theoretical basis and their implementation into these workflows. We will then show how earlier versions of these systems were benchmarked against previously modelled QSAR classification problems to classify the biological activity of the PLA2-IIA inhibitor dataset.

In Chapter 3, we apply the classification framework to predict organic solvent solubility. We explore the limitations of QSPR predictions on small datasets and address them by augmenting our dataset. We show that QSPRs can make accurate predictions when trained on sufficient data and address the limitations of current models, making future iterations more robust.

Chapter 4 investigates the ability of QSPRs to predict larger-scale self-assembly interactions. Using SAXS, we collected high-throughput and high-quality scattering profiles to develop a robust training set capable of probing a small molecule's ability to self-assemble across various solvents. We explore additional techniques to identify significant structural moieties governing the predictions, showing that such QSPR models may reveal design principles.

By developing systems to address the predictive gaps in supramolecular material design, the complete control and design of these functional materials could be fully realised. Given the extensive experimentation required to identify useful candidates, which are often assigned to applications retrospectively, enabling tailored design could help address the unique requirements of each application. Achieving this would enable the replacement of traditional materials with higher-performance, sustainable, and adaptable alternatives.

1.6 References

1. S. Bahl, H. Nagar, I. Singh and S. Sehgal, *Materials Today: Proceedings*, 2020, **28**, 1302–1306.
2. X. Fu, W. Cheng, G. Wan, Z. Yang and B. C. K. Tee, *Chemical Reviews*, 2024, **124**, 9899–9948.
3. Z. Khayat, *Supramolecular Chemistry*, 2023, **34**, 271–294.
4. J. Liu, C. Du, W. Huang and Y. Lei, *Biomaterials Science*, 2024, **12**, 8–56.
5. Y. Liu, X. Zhang, C. Xiao and B. Liu, *Materials Today Bio*, 2023, **20**, 100668.
6. N. K. Rege, N. F. B. Phillips and M. A. Weiss, *Current Opinion in Endocrinology, Diabetes and Obesity*, 2017, **24**, 267–278.
7. C. M. Wells, M. Harris, L. Choi, V. P. Murali, F. D. Guerra and J. A. Jennings, *J Funct Biomater*, 2019, **10**.
8. I. Ali, M. R. Islam, J. Yin, S. J. Eichhorn, J. Chen, N. Karim and S. Afroj, *ACS Nano*, 2024, **18**, 3871–3915.
9. Y. Zhou, X. Dong, Y. Mi, F. Fan, Q. Xu, H. Zhao, S. Wang and Y. Long, *Journal of Materials Chemistry A*, 2020, **8**, 10007–10025.
10. M. N. Mustafa, M. A. A. Mohd Abdah, A. Numan, A. Moreno-Rangel, A. Radwan and M. Khalid, *Renewable and Sustainable Energy Reviews*, 2023, **181**, 113355.
11. E. R. Draper, J. R. Lee, M. Wallace, F. Jäckel, A. J. Cowan and D. J. Adams, *Chemical Science*, 2016, **7**, 6499–6505.
12. T. Dutta, P. Chaturvedi, I. Llamas-Garro, J. S. Velázquez-González, R. Dubey and S. K. Mishra, *RSC Advances*, 2024, **14**, 12984–13004.
13. R. Pallotta, S. Cavalli, M. Degani and G. Grancini, *Small Structures*, 2024, **5**, 2300448.
14. M. Kök, İ. N. Qader, F. Dagdelen and Y. Aydoğdu, *El-Cezeri*, 2019, **6**, 755–788.
15. S. Kamila, *American Journal of Applied Sciences*, 2013, **10**.
16. M. Elsherif, A. E. Salih, M. G. Muñoz, F. Alam, B. AlQattan, D. S. Antonysamy, M. F. Zaki, A. K. Yetisen, S. Park, T. D. Wilkinson and H. Butt, *Advanced Photonics Research*, 2022, **3**, 2100371.
17. K. Uchino, in *Advanced Piezoelectric Materials (Second Edition)*, ed. K. Uchino, Woodhead Publishing, 2017, DOI: https://doi.org/10.1016/B978-0-08-102135-4.00001-1, pp. 1–92.
18. B. Wang, T. L. Lee and Y. Qin, *Materials*, 2023, **16**, 7206.
19. L. Zhu, W. Zhang, S. Luan, J. Wei, Y. Yang and J. Miao, *iScience*, 2025, **28**, 113126.
20. A. D. B. L. Ferreira, P. R. O. Nóvoa and A. T. Marques, *Composite Structures*, 2016, **151**, 3–35.
21. Y. S. Oh, J.-H. Kim, Z. Xie, S. Cho, H. Han, S. W. Jeon, M. Park, M. Namkoong, R. Avila, Z. Song, S.-U. Lee, K. Ko, J. Lee, J.-S. Lee, W. G. Min, B.-J. Lee, M. Choi, H. U. Chung, J. Kim, M. Han, J. Koo, Y. S. Choi, S. S. Kwak, S. B. Kim, J. Kim, J. Choi, C.-M. Kang, J. U. Kim, K. Kwon, S. M. Won, J. M. Baek, Y. Lee, S. Y. Kim, W. Lu, A. Vazquez-Guardado, H. Jeong, H. Ryu, G. Lee, K. Kim, S. Kim, M. S. Kim, J. Choi, D. Y. Choi, Q. Yang, H. Zhao, W. Bai, H. Jang, Y. Yu, J. Lim, X. Guo, B. H. Kim, S. Jeon, C. Davies, A. Banks, H. J. Sung, Y. Huang, I. Park and J. A. Rogers, *Nature Communications*, 2021, **12**, 5008.
22. Y. Yu, J. Li, S. A. Solomon, J. Min, J. Tu, W. Guo, C. Xu, Y. Song and W. Gao, *Science Robotics*, 2022, **7**, eabn0495.
23. H. Wu, G. Yang, K. Zhu, S. Liu, W. Guo, Z. Jiang and Z. Li, *Advanced Science*, 2021, **8**, 2001938.
24. W. Heng, S. Solomon and W. Gao, *Advanced Materials*, 2022, **34**, 2107902.

25. C. Zhao, J. Park, S. E. Root and Z. Bao, *Nature Reviews Bioengineering*, 2024, **2**, 671–690.
26. W. Chen, C. Wang, W. Liu, B. Zhao, Z. Zeng, F. Long, C. Wang, S. Li, N. Lin and J. Zhou, *Advanced Materials*, 2023, **35**, 2209041.
27. R. Tällberg, B. P. Jelle, R. Loonen, T. Gao and M. Hamdy, *Solar Energy Materials and Solar Cells*, 2019, **200**, 109828.
28. Z. Zheng, X. Li, G. Jiang, W. Cheng, D. Zhao and H. Yu, *Macromolecular Rapid Communications*, **n/a**, e00557.
29. T. I. Awan, S. Afsheen and A. Mushtaq, in *Influence of Noble Metal Nanoparticles in Sustainable Energy Technologies*, Springer Nature Switzerland, Cham, 2025, DOI: 10.1007/978-3-031-80983-5_3, pp. 49–78.
30. M. Yildirim and Z. Candan, *Materials Today: Proceedings*, 2023, DOI: https://doi.org/10.1016/j.matpr.2023.10.116.
31. S. S. Anandakrishnan, M. Tabeshfar, M. Nelo, J. Peräntie, H. Jantunen, J. Juuti and Y. Bai, *RSC Sustainability*, 2024, **2**, 961–974.
32. O. Guillon, *Advanced ceramics for energy conversion and storage*, Elsevier, Amsterdam, 2020.
33. J. Li, H. Huang and T. Morita, *Sensors and Actuators A: Physical*, 2019, **292**, 39–51.
34. M. Ling, L. Yuan, Z. Luo, T. Huang and X. Zhang, *Actuators*, 2022, **11**, 134.
35. K. Raj and A. P. Das, *Environmental Chemistry and Ecotoxicology*, 2023, **5**, 79–85.
36. *On the restriction of the use of certain hazardous substances in electrical and electronic equipment*, 2011, **2011/65/EU**, OJ L295/291.
37. M. Zacchini, *Chemosphere*, 2024, **360**, 142414.
38. R. Sharif, A. Khalid, S. W. Ahmad, A. Rehman, H. G. Qutab, H. H. Akhtar, K. Mahmood, S. Afzal and F. Saleem, *Nanoscale Advances*, 2023, **5**, 3803–3833.
39. R. Lin, J. Xu, M. Wei, Y. Wang, Z. Qin, Z. Liu, J. Wu, K. Xiao, B. Chen, S. M. Park, G. Chen, H. R. Atapattu, K. R. Graham, J. Xu, J. Zhu, L. Li, C. Zhang, E. H. Sargent and H. Tan, *Nature*, 2022, **603**, 73–78.
40. Fraunhofer Institute for Solar Energy Systems, Fraunhofer ISE Develops the World's Most Efficient Solar Cell with 47.6 Percent Efficiency, https://www.ise.fraunhofer.de/en/press-media/press-releases/2022/fraunhofer-ise-develops-the-worlds-most-efficient-solar-cell-with-47-comma-6-percent-efficiency.html, (accessed 28 March 2025).
41. National Renewable Energy Laboratory, Best Research-Cell Efficiency Chart, https://www.nrel.gov/pv/cell-efficiency.html, (accessed 28 March 2025).
42. J. Guo, Y. Liang, S. Zhang, D. Ma, T. Yang, W. Zhang, H. Li, S. Cao and B. Zou, *Green Energy and Resources*, 2023, **1**, 100007.
43. D. T. Gillaspie, R. C. Tenent and A. C. Dillon, *Journal of Materials Chemistry*, 2010, **20**, 9585–9592.
44. S. Zhang, Y. Ding, B. Liu and C.-c. Chang, *Waste Management*, 2017, **65**, 113–127.
45. Z. Wang and R. Liu, *Materials Today Electronics*, 2023, **4**, 100036.
46. Z. Guo, T. Sato, Y. Han, N. Takamura, R. Ikeda, T. Miyamoto, N. Kida, M. Ogino, Y. Takahashi, N. Kasuya, S. Watanabe, J. Takeya, Q. Wei, M. Mukaida and H. Okamoto, *Communications Materials*, 2024, **5**, 26.
47. M. Seiti, A. Giuri, C. E. Corcione and E. Ferraris, *Biomaterials Advances*, 2023, **154**, 213655.
48. S. Doshi, M. O. A. Forner, P. Wang, S. E. Hadwe, A. T. Jin, G. Dijk, K. Brinson, J. Lim, A. Dominguez-Alfaro, C. Y. J. Lim, A. Salleo, D. G. Barone, G. Hong, M. L. Brongersma, N. A. Melosh, G. G. Malliaras and S. T. Keene, *Adv Mater*, 2025, **37**, e2415827.
49. K. Fu, R. Lv, B. Na, S. Zou, R. Zeng, B. Wang and H. Liu, *Polymer*, 2019, **184**, 121900.

50. E. Chamanehpour, S. Thouti, H.-G. Rubahn, A. Dolatshahi-Pirouz and Y. K. Mishra, *Advanced Materials Technologies*, 2024, **9**, 2301392.
51. S. T. M, A. B. Arshad, P. T. Lin, J. Widakdo, M. H. K, H. F. M. Austria, C.-C. Hu, J.-Y. Lai and W.-S. Hung, *RSC Advances*, 2021, **11**, 9638–9663.
52. Z. Song, X. Liu, X. Sun, Y. Li, X. Nie, W. Tang, R. Yu and J. Shui, *Carbon*, 2019, **151**, 36–45.
53. M. M. Abdul Hameed, S. A. P. Mohamed Khan, B. M. Thamer, N. Rajkumar, H. El-Hamshary and M. El-Newehy, *Polymers for Advanced Technologies*, 2023, **34**, 6–23.
54. A. Balcerak-Woźniak, M. Dzwonkowska-Zarzycka and J. Kabatc-Borcz, *Materials (Basel)*, 2024, **17**.
55. K. Imato and Y. Ooyama, *Polymer Journal*, 2024, **56**, 1093–1109.
56. A. Fattah-alhosseini, R. Chaharmahali, S. Alizad, M. Kaseem and B. Dikici, *Hybrid Advances*, 2024, **5**, 100178.
57. C. C. Graham and D. Perker, 2024.
58. A. Hernandez-Martinez, *Archives of Pulmonology and Respiratory Care*, 2022, **8**, 010–011.
59. S. Panja and D. J. Adams, *Chemical Society Reviews*, 2021, **50**, 5165–5200.
60. R. K. Mishra, S. Das, B. Vedhanarayanan, G. Das, V. K. Praveen and A. Ajayaghosh, in *Molecular Gels: Structure and Dynamics*, ed. R. G. Weiss, The Royal Society of Chemistry, 2018, DOI: 10.1039/9781788013147-00190, p. 0.
61. Y. Hou, L. Zou, Q. Li, M. Chen, H. Ruan, Z. Sun, X. Xu, J. Yang and G. Ma, *Materials Today Bio*, 2022, **15**, 100327.
62. X. Li, J. Yang and Y.-W. Yang, *Materials Chemistry Frontiers*, 2023, **7**, 1463–1481.
63. B. J. Kim, D. Yang and B. Xu, *Trends in Chemistry*, 2020, **2**, 71–83.
64. T. A. Welsh and E. R. Draper, *RSC Advances*, 2021, **11**, 5245–5264.
65. S. V. Bhosale, M. Al Kobaisi, R. W. Jadhav, P. P. Morajkar, L. A. Jones and S. George, *Chemical Society Reviews*, 2021, **50**, 9845–9998.
66. B. O. Okesola, V. M. P. Vieira, D. J. Cornwell, N. K. Whitelaw and D. K. Smith, *Soft Matter*, 2015, **11**, 4768–4787.
67. F. Khan and S. Das, *ChemistrySelect*, 2022, **7**, e202200205.
68. D. K. Smith, in *Molecular Gels: Structure and Dynamics*, ed. R. G. Weiss, The Royal Society of Chemistry, 2018, DOI: 10.1039/9781788013147-00300, p. 0.
69. A. D. O'Donnell, S. Salimi, L. R. Hart, T. S. Babra, B. W. Greenland and W. Hayes, *Reactive and Functional Polymers*, 2022, **172**, 105209.
70. E. R. Draper and D. J. Adams, *Chem*, 2017, **3**, 390–410.
71. C. A. Dreiss, in *Wormlike Micelles: Advances in Systems, Characterisation and Applications*, eds. C. A. Dreiss and Y. Feng, The Royal Society of Chemistry, 2017, DOI: 10.1039/9781782629788-00001, p. 0.
72. E. R. Draper and D. J. Adams, *Langmuir*, 2019, **35**, 6506–6521.
73. C. D. Jones and J. W. Steed, *Chemical Society Reviews*, 2016, **45**, 6546–6596.
74. R. J. Staples, *Acta Crystallogr E Crystallogr Commun*, 2025, **81**, 895–905.
75. C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallographica Section B*, 2016, **72**, 171–179.
76. K. A. Houton, K. L. Morris, L. Chen, M. Schmidtmann, J. T. A. Jones, L. C. Serpell, G. O. Lloyd and D. J. Adams, *Langmuir*, 2012, **28**, 9797–9806.
77. I. Ramos Sasselli, P. J. Halling, R. V. Ulijn and T. Tuttle, *ACS Nano*, 2016, **10**, 2661–2668.
78. P. Flowers, K. Theopold, R. Langley and W. R. Robinson, OpenStax, Houston, Texas, 2019.
79. J. Allday and S. Hands, *Introduction to entropy : the way of the world*, CRC Press, Taylor & Francis Group, Boca Raton, FL, 2024.

80. J. W. Steed, D. R. Turner and K. J. Wallace, *Core concepts in supramolecular chemistry and nanochemistry*, John Wiley, Chichester, England, 2007.
81. J. N. Israelachvili, *Intermolecular and surface forces*, Academic Press, 3rd edn., 2015.
82. E. E. Meyer, K. J. Rosenberg and J. Israelachvili, *Proc Natl Acad Sci U S A*, 2006, **103**, 15739–15746.
83. D. T. Infield, A. Rasouli, G. D. Galles, C. Chipot, E. Tajkhorshid and C. A. Ahern, *J Mol Biol*, 2021, **433**, 167035.
84. P. J. Cragg, in *Supramolecular Chemistry: From Biological Inspiration to Biomedical Applications*, Springer Netherlands, Dordrecht, 2010, DOI: 10.1007/978-90-481-2582-1_1, pp. 1–48.
85. Z. Mu, Q. Shao, J. Ye, Z. Zeng, Y. Zhao, H. H. Hng, F. Y. C. Boey, J. Wu and X. Chen, *Langmuir*, 2011, **27**, 1314–1318.
86. Z. Nazarian, C. M. Forsyth and S. E. Wheeler, *Chemistry of Materials*, 2026, **38**, 3879–3903.
87. W.-R. Zhuang, Y. Wang, P.-F. Cui, L. Xing, J. Lee, D. Kim, H.-L. Jiang and Y.-K. Oh, *Journal of Controlled Release*, 2019, **294**, 311–326.
88. A. S. Mahadevi and G. N. Sastry, *Chemical Reviews*, 2013, **113**, 2100–2138.
89. A. M. Reilly and A. Tkatchenko, *Chemical Science*, 2015, **6**, 3289–3301.
90. F. Biedermann, W. M. Nau and H.-J. r. Schneider, *Angewandte Chemie International Edition*, 2014, **53**, 11158–11171.
91. F. Biedermann, V. D. Uzunova, O. A. Scherman, W. M. Nau and A. De Simone, *Journal of the American Chemical Society*, 2012, **134**, 15318–15323.
92. E. Mattia and S. Otto, *Nature Nanotechnology*, 2015, **10**, 111–119.
93. C. E. Shields, A. M. Scholes and A. G. Slater, *Chemical Science*, 2025, **16**, 20124–20138.
94. E. R. Draper and D. J. Adams, *Nature Materials*, 2024, **23**, 13–15.
95. J. J. B. van der Tol, G. Vantomme and E. W. Meijer, *J Am Chem Soc*, 2023, **145**, 17987–17994.
96. S. Debnath, S. Roy, Y. M. Abul-Haija, P. W. J. M. Frederix, S. M. Ramalhete, A. R. Hirst, N. Javid, N. T. Hunt, S. M. Kelly, J. Angulo, Y. Z. Khimyak and R. V. Ulijn, *Chemistry – A European Journal*, 2019, **25**, 7881–7887.
97. Y. Hisamatsu, G. Toriyama, K. Yamamoto, H. Takase, T. Higuchi and N. Umezawa, *Chemistry – A European Journal*, 2024, **30**, e202400134.
98. J. Nanda, A. Biswas and A. Banerjee, *Soft Matter*, 2013, **9**, 4198–4208.
99. R. E. Ginesi and E. R. Draper, *Soft Matter*, 2024, **20**, 3887–3896.
100. K. McAulay, P. A. Ucha, H. Wang, A. M. Fuentes-Caparrós, L. Thomson, O. Maklad, N. Khunti, N. Cowieson, M. Wallace, H. Cui, R. J. Poole, A. Seddon and D. J. Adams, *Chemical Communications*, 2020, **56**, 4094–4097.
101. N. Hazra, S. Hazra, K. Gayen, P. Ghosh, A. Palui and A. Banerjee, *ACS Applied Polymer Materials*, 2024, **6**, 859–869.
102. J. J. Walsh, J. R. Lee, E. R. Draper, S. M. King, F. Jäckel, M. A. Zwijnenburg, D. J. Adams and A. J. Cowan, *The Journal of Physical Chemistry C*, 2016, **120**, 18479–18486.
103. F. Würthner, C. R. Saha-Möller, B. Fimmel, S. Ogi, P. Leowanawat and D. Schmidt, *Chemical Reviews*, 2016, **116**, 962–1052.
104. N. J. Hestand and F. C. Spano, *Chemical Reviews*, 2018, **118**, 7069–7163.
105. P. J. Walla, in *Modern Biophysical Chemistry: Detection and Analysis of Biomolecules, 2nd Edition*, Wiley-VCH, Weinheim, 2014.
106. M. Kasha, H. R. Rawls and M. Ashraf El-Bayoumi, *Pure and applied chemistry*, 1965, **11**, 371–392.

107. X. Yang, X. Xu and H.-F. Ji, *The Journal of Physical Chemistry B*, 2008, **112**, 7196–7202.
108. E. R. Draper, B. J. Greeves, M. Barrow, R. Schweins, M. A. Zwijnenburg and D. J. Adams, *Chem*, 2017, **2**, 716–731.
109. R. K. Dubey, S. J. Eustace, J. S. van Mullem, E. J. R. Sudhölter, F. C. Grozema and W. F. Jager, *The Journal of Organic Chemistry*, 2019, **84**, 9532–9547.
110. A. G. Slater, E. S. Davies, S. P. Argent, W. Lewis, A. J. Blake, J. McMaster and N. R. Champness, *The Journal of Organic Chemistry*, 2013, **78**, 2853–2862.
111. M. Al Kobaisi, S. V. Bhosale, K. Latham, A. M. Raynor and S. V. Bhosale, *Chemical Reviews*, 2016, **116**, 11685–11796.
112. Z. Chen, U. Baumeister, C. Tschierske and F. Würthner, *Chemistry – A European Journal*, 2007, **13**, 450–465.
113. J. G. Egan, University of Glasgow, 2023.
114. Z. Chen, B. Fimmel and F. Würthner, *Organic & Biomolecular Chemistry*, 2012, **10**, 5845–5855.
115. C. Naranjo, A. Doncel-Giménez, R. Gómez, J. Aragó, E. Ortí and L. Sánchez, *Chemical Science*, 2023, **14**, 9900–9909.
116. Y. Chen, Y. Feng, J. Gao and M. Bouvet, *Journal of Colloid and Interface Science*, 2012, **368**, 387–394.
117. E. R. Draper, J. J. Walsh, T. O. McDonald, M. A. Zwijnenburg, P. J. Cameron, A. J. Cowan and D. J. Adams, *Journal of Materials Chemistry C*, 2014, **2**, 5570–5575.
118. E. R. Draper, L. J. Archibald, M. C. Nolan, R. Schweins, M. A. Zwijnenburg, S. Sproules and D. J. Adams, *Chemistry – A European Journal*, 2018, **24**, 4006–4010.
119. T. A. Welsh, O. Matsarskaia, R. Schweins and E. R. Draper, *New Journal of Chemistry*, 2021, **45**, 14005–14013.
120. T. A. Welsh, J. G. Egan, B. Dietrich, N. Rafferty, R. E. Ginesi, J. Doutch, R. Schweins and E. R. Draper, *Journal of Physics: Materials*, 2023, **7**, 015004.
121. V. Adams, J. Cameron, M. Wallace and E. R. Draper, *Chemistry – A European Journal*, 2020, **26**, 9879–9882.
122. S. Dutt, P. F. Siril and S. Remita, *RSC Advances*, 2017, **7**, 5733–5750.
123. R. Nagarajan, *Langmuir*, 2002, **18**, 31–38.
124. S. Abbott, *Update*, 2016, **1**, 2–26.
125. Y. Ishiwatari, T. Yokoyama, T. Kojima, T. Banno and N. Arai, *Molecular Systems Design & Engineering*, 2024, **9**, 20–28.
126. R. Van Lommel, W. M. De Borggraeve, F. De Proft and M. Alonso, *Gels*, 2021, **7**, 87.
127. G. A. DiLabio and A. Otero-de-la-Roza, *Reviews in computational chemistry*, 2016, **29**, 1–97.
128. V. Setnička, J. Nový, S. Böhm, N. Sreenivasachary, M. Urbanová and K. Volka, *Langmuir*, 2008, **24**, 7520–7527.
129. R. Van Lommel, W. M. De Borggraeve, F. De Proft and M. Alonso, *Gels*, 2021, **7**.
130. J. F. Woods, L. Gallego, P. Pfister, M. Maaloum, A. Vargas Jentzsch and M. Rickhaus, *Nature Communications*, 2022, **13**, 3681.
131. E. R. Draper, L. Wilbraham, D. J. Adams, M. Wallace, R. Schweins and M. A. Zwijnenburg, *Nanoscale*, 2019, **11**, 15917–15928.
132. J. H. Moore, D. J. Cole and G. Csanyi, *arXiv preprint arXiv:2405.18171*, 2024.
133. L. Joubert, G. Picard and J.-J. Legendre, *Journal of Alloys and Compounds*, 1998, **275-277**, 934–939.
134. P. Zhou, C. Yuan and X. Yan, *Current Opinion in Colloid & Interface Science*, 2022, **62**, 101645.

135. E. R. Draper, B. Dietrich, K. McAulay, C. Brasnett, H. Abdizadeh, I. Patmanidis, S. J. Marrink, H. Su, H. Cui, R. Schweins, A. Seddon and D. J. Adams, *Matter*, 2020, **2**, 764–778.
136. R. R. Sonani, S. Bianco, B. Dietrich, J. Doutch, E. R. Draper, D. J. Adams and E. H. Egelman, *Cell Reports Physical Science*, 2024, **5**, 101812.
137. C. R. M. MacDonald and E. R. Draper, *Beilstein J Org Chem*, 2024, **20**, 2608–2634.
138. G. Zaldivar, Y. A. Perez Sirkin, G. Debais, M. Fiora, L. L. Missoni, E. Gonzalez Solveyra and M. Tagliazucchi, *ACS Omega*, 2022, **7**, 38109–38121.
139. I. Szleifer and M. A. Carignano, in *Advances in Chemical Physics*, 1996, DOI: https://doi.org/10.1002/9780470141533.ch3, pp. 165–260.
140. *Journal*, 2025.
141. A. Klamt, F. Eckert and W. Arlt, *Annu Rev Chem Biomol Eng*, 2010, **1**, 101–122.
142. J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, *Nature Reviews Drug Discovery*, 2019, **18**, 463–477.
143. E. F. Morales and H. J. Escalante, in *Biosignal Processing and Classification Using Computational Learning and Intelligence*, eds. A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda and O. Mendoza-Montoya, Academic Press, 2022, DOI: https://doi.org/10.1016/B978-0-12-820125-1.00017-8, pp. 111–129.
144. H. B, V. G, Vaibhav, K. S. Anantharaju and R. K. Pai, *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 2024, **7**, 100458.
145. S. L. Brunton and J. N. Kutz, *Data-driven science and engineering: machine learning, dynamical systems, and control*, Cambridge University Press, Cambridge, United Kingdom, New York, NY, Second edn., 2022.
146. M. van Otterlo and M. Wiering, in *Reinforcement Learning: State-of-the-Art*, eds. M. Wiering and M. van Otterlo, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, DOI: 10.1007/978-3-642-27645-3_1, pp. 3–42.
147. A. Srinivasan, *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 2023, **07**.
148. K. Cobbe, O. Klimov, C. Hesse, T. Kim and J. Schulman, 2019.
149. Z. Ghahramani, in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, eds. O. Bousquet, U. von Luxburg and G. Rätsch, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, DOI: 10.1007/978-3-540-28650-9_5, pp. 72–112.
150. T. F. Sterkenburg and P. D. Grünwald, *Synthese*, 2021, **199**, 9979–10015.
151. OpenAI, ChatGPT, https://chatgpt.com/, (accessed 21st Aprtil 2026).
152. Microsoft, Copilot, https://copilot.cloud.microsoft/, (accessed 21st April 2026).
153. Google, Gemini, https://gemini.google.com/, (accessed 21st April 2026).
154. N. T. Runcie, C. M. Deane and F. Imrie, *arXiv preprint arXiv:2505.07735*, 2025.
155. J. Delbeke, A. Runge-Metzger, Y. Slingenberg and J. Werksman, in *Towards a climate-neutral Europe*, Routledge, 2019, pp. 24–45.
156. Google 2024 Environmental Report, https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf, (accessed 5th January 2026).
157. P. Li, J. Yang, M. A. Islam and S. Ren, *Communications of the ACM*, 2025, **68**, 54–61.
158. S. Luccioni, B. Gamazaychikov, T. A. da Costa and E. Strubell, *arXiv preprint arXiv:2506.15572*, 2025.
159. N. Jegham, M. Abdelatti, C. Y. Koh, L. Elmoubarki and A. Hendawi, *arXiv preprint arXiv:2505.09598*, 2025.

160. B. Courty, V. Schmidt, Goyal-Kamal, M. Coutarel, B. Feld, J. Lecourt, L. Connell, SabAmine, inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, A. Saboni, H. d. Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, alencon, M. Stęchły, C. Bauer and Lucas-Otavio, *Journal*, 2024, DOI: 10.5281/zenodo.11171501.
161. A. de Vries, *Joule*, 2023, **7**, 2191–2194.
162. Climate Action Tracker, *Climate Analytics and New Climate Institute: Berlin, Germany*, 2025.
163. A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, *Nature*, 2023, **624**, 80–85.
164. P. W. J. M. Frederix, R. V. Ulijn, N. T. Hunt and T. Tuttle, *The Journal of Physical Chemistry Letters*, 2011, **2**, 2380–2384.
165. A. van Teijlingen and T. Tuttle, *Journal of Chemical Theory and Computation*, 2021, **17**, 3221–3232.
166. T. Inokuchi, N. Li, K. Morohoshi and N. Arai, *Nanoscale*, 2018, **10**, 16013–16021.
167. E. S. Boek, P. V. Coveney, H. N. W. Lekkerkerker and P. van der Schoot, *Physical Review E*, 1997, **55**, 3124–3133.
168. R. Todeschini, V. Consonni and P. Gramatica, in *Comprehensive Chemometrics*, eds. S. D. Brown, R. Tauler and B. Walczak, Elsevier, Oxford, 2009, DOI: https://doi.org/10.1016/B978-044452701-1.00007-7, pp. 129–172.
169. A. C. Brown and T. R. Fraser, *J Anat Physiol*, 1868, **2**, 224–242.
170. OECD, *OECD Series on Testing and Assessment*, 2023, DOI: 10.1787/d96118f6-en.
171. A. Gissi, O. Tcheremenskaia, C. Bossa, C. L. Battistelli and P. Browne, *Computational Toxicology*, 2024, **31**, 100326.
172. K. B. Lipkowitz, T. R. Cundari and ProQuest, *Reviews in computational chemistry*, Wiley, Hoboken, NJ, 2007.
173. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Advanced Drug Delivery Reviews*, 2001, **46**, 3–26.
174. L. Zhao, W. Wang, A. Sedykh and H. Zhu, *ACS Omega*, 2017, **2**, 2805–2812.
175. J. K. Gupta, D. J. Adams and N. G. Berry, *Chem Sci*, 2016, **7**, 4713–4719.
176. W. Wu, R. Zhang, S. Peng, X. Li and L. Zhang, *Chemometrics and Intelligent Laboratory Systems*, 2016, **157**, 7–15.
177. B. Hollas, *Journal of Mathematical Chemistry*, 2003, **33**, 91–101.
178. J. K. Gupta, D. J. Adams and N. G. Berry, *Chemical Science*, 2016, **7**, 4713–4719.
179. F. Li, J. Han, T. Cao, W. Lam, B. Fan, W. Tang, S. Chen, K. L. Fok and L. Li, *Proceedings of the National Academy of Sciences*, 2019, **116**, 11259–11264.
180. A. J. Engler, S. Sen, H. L. Sweeney and D. E. Discher, *Cell*, 2006, **126**, 677–689.
181. J. D. Simpson, L. Thomson, C. M. Woodley, C. M. Wallace, B. Dietrich, A. S. Loch, D. J. Adams and N. G. Berry, *Advanced Materials*, 2025, **37**, 2415031.

Chapter 2: Developing a quantitative structure-property relationship framework in KNIME

This chapter is adapted, in part, from the following publication:

"An accessible property classification framework to predict the solubility of functionalised naphthalenes and rylenes in organic solvents"

C. R. M. MacDonald, Alex S. Loch, Neil G. Berry and E. R. Draper

CRMM was responsible for researching, creating figures, and writing the publication. CRMM wrote the initial draft of the manuscript, to which all authors contributed for the final publication.

2.1 Introduction

As discussed in Chapter 1, quantitative structure-property relationships (QSPRs) have gained increasing interest across various fields of chemistry as computing power has become more abundant and machine learning capabilities have improved.[1-3] However, as machine learning tools become more complex, they can often become uninterpretable and inaccessible to a typical wet-lab chemist.[4] Subsequently, open-source, accessible tools are required if such systems are to be effectively utilised across various fields of chemistry.

The limited research applying QSPRs to material chemistry problems provides a promising opportunity to develop useful design principles using machine learning, which typical chemical intuition has failed to realise.[5-7] The quality of QSPR models directly depends on the quality of the training data; therefore, the limited availability of material chemistry data may significantly impact the predictive ability of the resulting models. Therefore, to assess the effectiveness of the QSPR framework, it is necessary to separate the limitations of the modelling procedure from those of the input data.

The capacity of our framework to build competent models was first investigated using high-quality, large-scale pharmacological datasets known to be amenable to quantitative structure-activity relationship (QSAR) modelling, to ensure that the framework can build models with data of sufficient quality. By proving that the framework can build effective models when data quality is sufficient, we can better understand whether the limitations of future models arise from deficiencies in the data or in the model framework.

The following chapter will discuss the development of a quantitative structure-property relationship framework in KNIME (Section 2.3.2). We break down each module and how each part contributes to the system, focusing on its theoretical basis and implementation into these workflows. We show how earlier versions of these systems were benchmarked against previously modelled QSAR classification problems to classify the biological activity of the PLA2-IIA inhibitor dataset. Figure 2.1 presents a simplified schema of the QSPR workflow built in KNIME for this project. The workflow identifies how each phase of the QSPR process consists of multiple subprocesses. Each of these are processes are described and discussed herein to provide a complete understanding of our QSPR system.

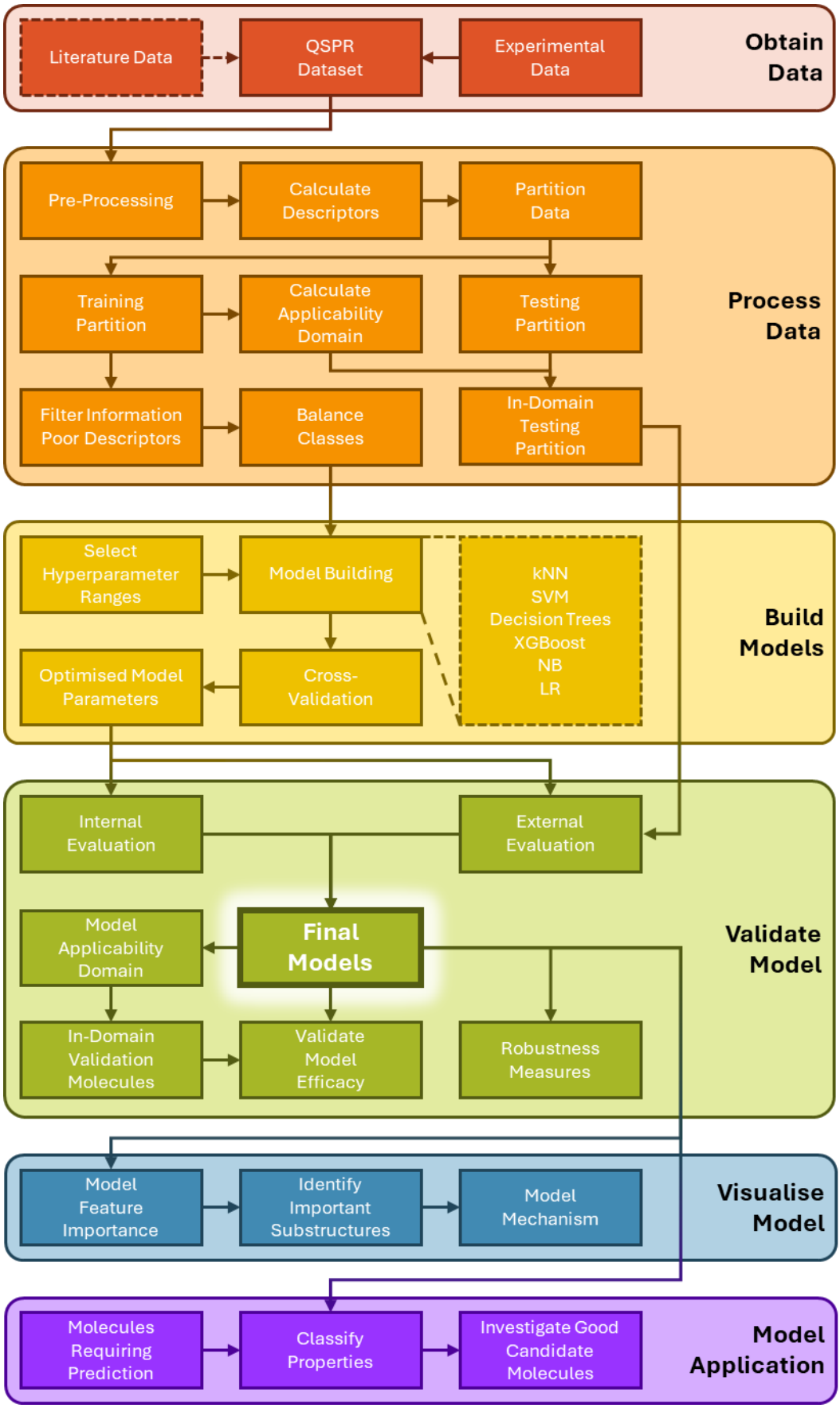


Figure 2.1. Overall QSPR classification modelling, testing, and validation workflow. The simplified workflow highlights the numerous processes required to obtain and process data, build and select models using internal test sets, and externally validate with independent predictions.

2.2 Dataset Acquisition

Before a QSPR model can be built, training data accurately describing the endpoint of interest is required. This endpoint must have a clear scientific purpose and be built on transparent experimental data.[8, 9] The quality of training data is dependent entirely on the reliability of the data sources.[10] As a result, data curation is among the most important steps in QSPR development.[11] A fundamental assumption in cheminformatics studies is that the training data accurately represent the underlying chemical structures.[10] It has been demonstrated that a small number of erroneous structures represented by erroneous descriptors will have a significant detrimental effect on a QSPR model's predictivity.[12, 13] Subsequently, to develop effective QSPR models, data must be collected and curated to afford high-quality datasets from reproducible sources.

2.2.1 Experimental Data

To ensure complete control over input data quality, the endpoint is best probed through experimentation to collect training data for models to learn from. By making QSPR tools accessible to wet-lab chemists with intimate knowledge of the chemical systems of interest, they can better develop processes and capture data to ensure a high-quality description of the desired endpoint. A significant error in QSPR development is the failure to account for data heterogeneity.[14] There are often several ways to investigate a given endpoint, each yielding different results. Furthermore, what seems to be a simple, well-defined endpoint may be interpreted differently by different researchers. Therefore, for the highest-quality dataset, the data should be collected using a single protocol.

QSPRs are of interest precisely because experimental investigations are costly and time intensive. Thus, capturing quality data often comes at the cost of dataset size. An investigation into the effects of dataset size by Rácz *et al*. highlights the importance of ensuring a sufficient number of datapoints to train QSPR models (Figure 2.2).[15] Across three case studies, it was clear that while some models are less dependent on dataset size, the predictive performance of machine learning models decreases significantly as the training dataset is reduced. Subsequently, it is not only important to ensure that data collection methods are robust but also afford sufficient throughput to yield adequate dataset sizes. While a single experimentalist may be able to produce a dataset amenable to modelling, the limitations in data collection may hinder the efficacy and scope of the resulting predictions. Subsequently, methods are often sought to augment QSPR datasets.

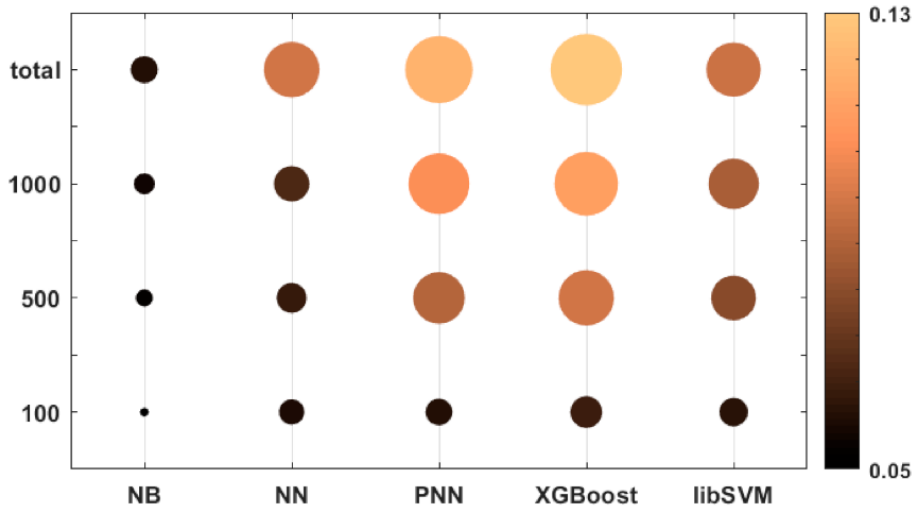


Figure 2.2. Bubble plot of the performances of different machine learning algorithms at different dataset sizes when modelling cytochrome P450 inhibitors from the Pubchem Bioassay database. A larger, brighter bubble indicates better performance corresponding to the average of the 25 normalised performance parameter values. Reproduced from Rácz *et al.*[15]

2.2.2 Literature Mining

While the continuous publication of new results provides access to an ever-expanding pool of data to build QSPRs, only a small portion of all available research is available in a well-structured format (Figure 2.3).[16] In the absence of large databases, mining the literature for appropriate data can be an arduous process. Most chemical knowledge is reported in unstructured natural language, making rules-based information extraction difficult, especially because data reporting varies across applications.[16] Subsequently, tools are being developed to extract crucial yet disconnected data from across the breadth of the literature. Machine learning has been employed to aid the extraction of useful datasets from papers and patents; however, these systems often perform less effectively when retrieving information for complex chemical structures and can struggle to handle the heterogeneity of scientific documents.[17] Large language models have been able to better tackle document heterogeneity by interpreting figures using vision-language models. However, even LLM data mining still faces several limitations, including an inability to interpret routine methods when they are not explicitly provided, a bias toward scientific literature, and a lack of benchmarks to improve these tools further.[16] Subsequently, manual curation is often necessary to ensure the preparation of high-quality datasets.

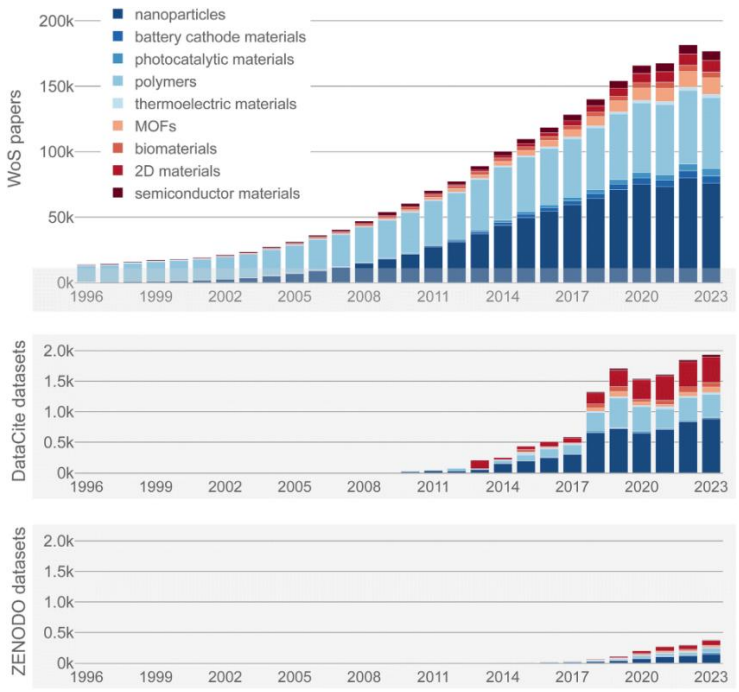


Figure 2.3. Number of research papers *vs* datasets deposited in data repositories in materials science and chemistry per year. The top graph shows the number of publications from 1996 to 2023. The two graphs below show the number of datasets in chemistry and materials science deposited in the Zenodo and DataCite repositories from 1996 to 2023. Reproduced from Schilling-Wilhhelmi *et al.*[16]

QSARs are at a distinct advantage due to the standardisation of assays and protocols employed across biochemical research. Extensively research endpoints, such as measuring inhibitory concentration (IC_{50}) for biological functions critical to developing new medicines, mean that there is a significant backbone of harmonised literature that make them ideal targets for modelling. The importance of such endpoints, together with the wealth of measurements, has led to extensive, manually curated databases of biochemical data, such as PubChem and ChEMBL, the two largest open-access databases of molecular activity data.[18-20] Such databases have been used to develop effective QSARs.[21, 22] However, while highly standardised in practice, real-world implementations vary widely across laboratories and assays. A lack of machine-interpretable language to describe biological assays, both in the primary literature and in aggregated databases, makes integrating multiple databases problematic.[23] It has also been found that, for a given database, separating assays by materials and methods is necessary to avoid building models with no predictive ability, further highlighting the impact of protocol standardisation on the generation of datasets amenable to QSPR modelling.[23]

Given the limitations of literature mining and manually curated databases, it is naïve to assume that any dataset will yield performative models. To adequately benchmark our QSPR framework, it is necessary to use training data which is known to be amenable to modelling.

Should the modelability of the input data be undetermined, the failure to yield effective models could either be from deficiencies in the dataset or an ineffective modelling procedure. By utilising datasets previously used to produce effective QSPR models, any deficiency in the predictive performance is indicative of limitations in the modelling procedure. Subsequently, we used a ChEMBL dataset which provides IC_{50} values for inhibition of phospholipase A2 group IIA (PLA2-IIA) activity (ChEMBL ID: CHEMBL3474).[19] The PLA2-IIA dataset was previously shown to be modelled well by Woodley, and was therefore an ideal dataset to benchmark a new QSPR system.[24] An ideal dataset would be free experimental noise which could hinder the modelling procedure, however it has previously been reported that only 650 assays have data for more than 100 distinct compounds, and there are only 54 assays with data for more than 500 distinct compounds.[25] This has made the practice of combining assays a common practice for the development of ML systems. In the case of the PLA2-IIA target, while no single assay provides enough IC_{50} data to support effective modelling (with the most frequent assay in the dataset describing only 84 molecules), aggregating the data across sufficiently similar assays measuring the same target affords a dataset of sufficient quality for modelling.[24] Given the previous success of modelling the CHEMBL3474 dataset shows that limiting training data to IC50 data to assays that standardised activities on a single human protein target. With over 650 data points, the PLA2-IIA IC_{50} data is also sufficient to minimise the performance impact of small training sets in the modelling process. It is important to note, however, that CHEMBL3474 aggregates data from more than 50 different assays. This is largely unavoidable given that approximately three-quarters of ChEMBL assays contain data for fewer than ten compounds.[25] Although all measurements relate to a single biological target, substantial heterogeneity may still exist between assays, including differences in buffer composition, pH, temperature, assay duration, analyte concentration, and assay technology. These sources of variability may be compounded by database-related issues, such as the assignment of distinct protein variants or cell types to the same ChEMBL identifier, as well as transcription errors introduced during the transfer of data from the primary literature into the database.[25] Consequently, while the previous work of Woodly established CHEMBL3474 as a useful benchmark for evaluating the predictive performance of the QSPR framework, the experimental noise inherent to aggregated ChEMBL datasets may limit the mechanistic insight that can be extracted from the resulting models. Future studies should therefore employ rigorous data curation procedures to maximise dataset homogeneity and minimise experimental noise. However, achieving this level of curation may prove challenging, as

identifying and accounting for assay-specific sources of variation often requires substantial domain expertise.

2.3 Machine Learning Platforms

Having acquired a sufficiently well-defined dataset, the data must be processed and modelled to build a QSPR. As computational systems become more functional, they often become more complex. As we discussed in Chapter 1, software that enables quantum mechanics and molecular dynamics modelling is often prohibitively complex, requiring significant expertise to utilise efficiently and limiting its use by typical wet-lab chemists. If a QSPR framework is to be accessible and utilised appropriately, it must not only be open source but also interpretable by chemists with limited ML knowledge.

2.3.1 Python

Python is an incredibly versatile programming language which has exploded in popularity across scientific projects.[26] Owing to its beginner-friendly syntax, a continuously growing library of tools, and an active development community, Python has emerged as an effective system for various applications, particularly in machine learning.[27] With the release of open-source scientific toolbox libraries in Python, such as scikit-learn,[28] the ability to develop and train machine learning models is becoming increasingly accessible. While scikit-learn alone may not provide every required functionality, Python libraries can be easily integrated in tandem to build the required system. While compared to other languages, Python is much more concise and easier to interpret, the code required to tackle complex tasks is still non-trivial to parse and understand.[29] To ensure that our framework remains accessible, we require the versatility of Python, but an interface much more accessible to the typical chemist.

2.3.2 Visual Programming

Visual programming encompasses tools that enable the development of programs visually, typically by connecting boxes with pre-defined programming functionality, allowing non-programmers to develop programs.[30] Chemists have already used such tools to build highly effective QSPRs. Gupta *et al*. previously developed the first predictive machine-learning models for functionalised mono- and dipeptides using the visual programming platform Pipeline Pilot.[31, 32] While Pipeline Pilot is a well-established cheminformatics tool and was previously regarded as a necessary specialist product, as free and open-source alternatives emerge, it may now be considered a commodity product.[33] By developing QSPR tools

behind prohibitive paywalls, they become less accessible, reducing their adoption among typical chemists. While the advantages of visual programming support the OECD's requirement for an unambiguous algorithm set for the development of effective QSPRs, the inaccessibility of Pipeline Pilot somewhat undermines the clear advantages of the platform. With that in mind, the KoNstanz Information MinEr[34] (KNIME) is a promising alternative for developing accessible QSPR tools. While KNIME originated as a "business intelligence" product, its adoption by computational chemists and pharmaceutical companies has led to the development of numerous cheminformatics functionalities in KNIME. It has previously been shown to be an effective tool for chemical data mining, pre-processing for QSAR development, and machine learning model development.[11, 35] In addition to the advantages in accessibility, the focus on backwards compatibility ensures that the tools developed within the platform can still be executed reliably, even as the platform is adjusted and upgraded.[35] Given the numerous advantages of the KNIME platform and its proven ability to tackle QSPR problems, it was chosen as the optimal system for building our QSPR framework.

Workflows in KNIME are created by connecting nodes with predefined yet tuneable functionality to manipulate data for a range of purposes. While KNIME provides a range of nodes, many highly specific tasks may not be natively supported. However, given the open-source nature of KNIME, there are several freely available extensions which provide nodes with additional functionality. Where such nodes do not exist, highly customisable nodes allow Python code to manipulate the data as required. Each of these nodes can be connected in series and parallel to shuttle data through a series of manipulations and functions to build a QSPR. To enhance the system's readability, nodes can be grouped into subunits known as metanodes and components. This allows the system to be broken down into manageable units, making it easier for a chemist unfamiliar with QSPR to understand, thereby enhancing its accessibility. As well as moving data between nodes, KNIME utilises "flow variables" which can be manually defined or output by the system. These can be used to control the tuning of each node, for example, by defining what data are relevant to a given process or how the node should be calibrated, given features of the input data. Using these functions, we developed a complete QSPR system capable of classifying molecules according to the desired criteria. The QSPR framework takes the data and transforms it into a machine-readable format. The pre-processed data is subsequently partitioned to build the models and then validate their performance (Figure 2.4). While the earliest iterations of our system offered much more limited functionality than the final framework, herein we will discuss the concepts used across the prediction problems modelled across Chapters 2, 3, and 4.

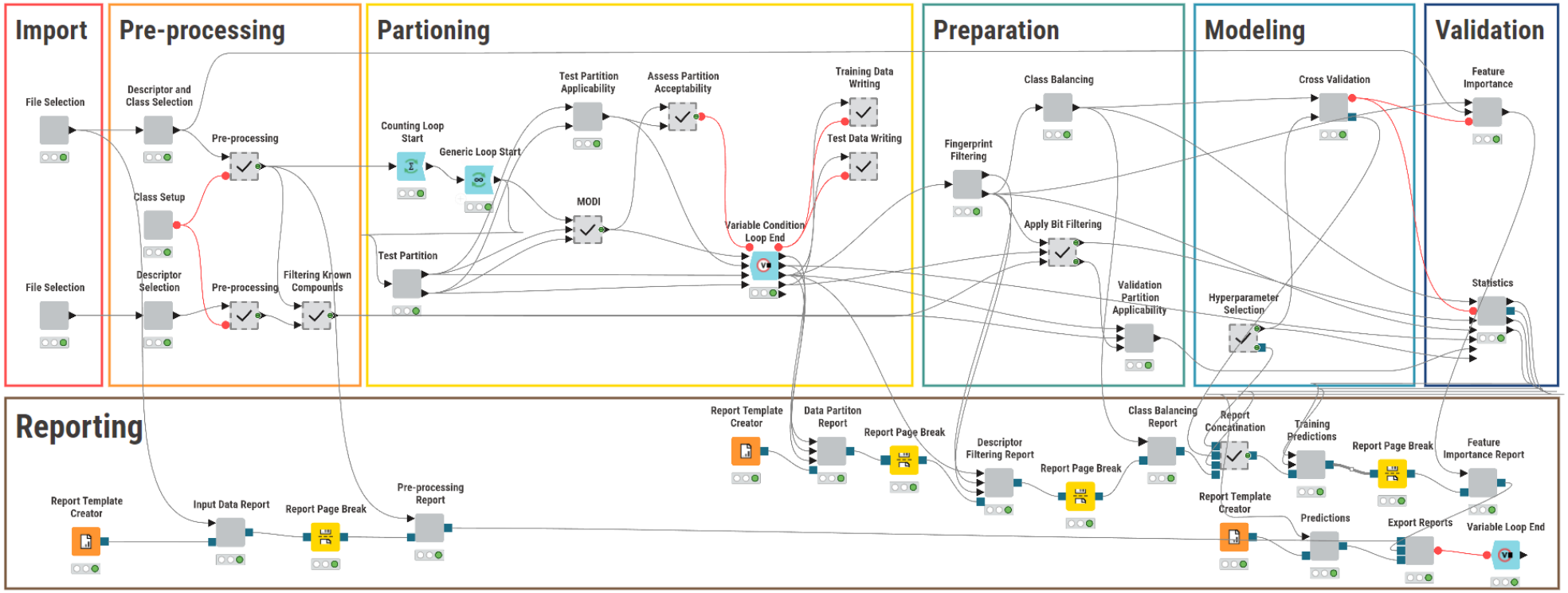


Figure 2.4. Top-level view of the final iteration of the QSPR framework. The framework comprises a series of interconnected nodes, metanodes, and components capable of receiving the required data, processing it, constructing predictive models, and performing validation analyses on the resulting outputs. Although the workflow contains hundreds of individual nodes, the grouping of related processes into visual modules allows it to remain interpretable, analogous to the simplified schematic presented in Figure 2.1. Such transparency would be more difficult to achieve using traditional programming approaches.

2.4 Dataset Pre-processing

To begin working with our data in KNIME, we must import it appropriately. In this project, we utilise both simple comma-separated value (CSV) files; Excel files; and more complex structure data files (SDFs), a widely used text-based cheminformatic file format that collates the MOLFile representing a compound's identity and structure with additional property metadata for a library of molecules.[36] The earliest iterations of our QSPR system required substantial manual intervention to adapt between datasets. Most nodes require some input to define the data of interest for their given function. Typically, nodes in KNIME process data in columns and therefore require the column header to be named to identify the appropriate data. Therefore, if the name of the column of interest changes across different input data files, the system will fail to identify the correct data. In the case of PLA2-IIA, the data is available in SDF format, which cannot be edited as easily as a CSV file. Subsequently, the columns corresponding to chemical structure and target property cannot be easily unified across problems. While the nodes can be updated manually, failing to correctly update any nodes dependent on these columns with the appropriate header name causes the workflow to fail. Such problems are where flow variables are essential in adapting the framework. By developing a component that allows the user to identify the column defining the structural data and the target property, and to express them as flow variables, the appropriate information can be passed to the necessary nodes with minimal manual intervention (Figure 2.5). Such functions are essential to minimise human error, which may result in the misapplication of the QSPR, and provide a clearer interface for the user.

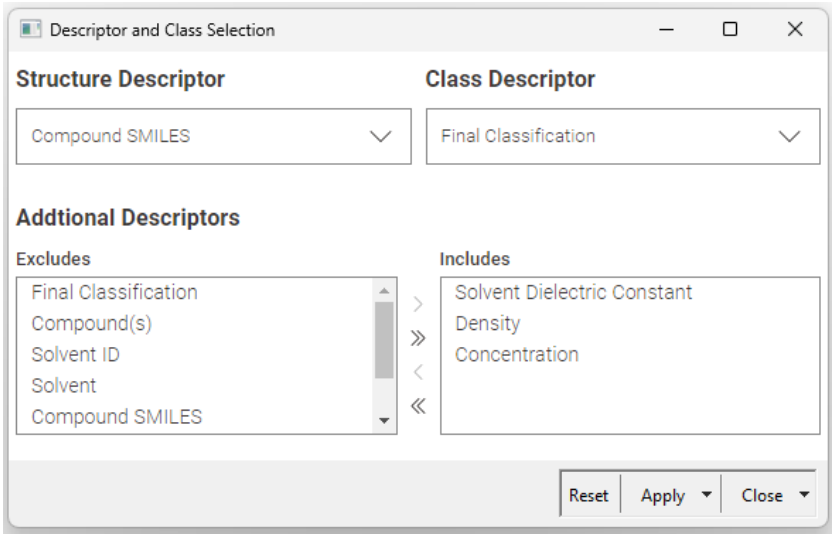


Figure 2.5. Simplified user interface for the selection of structural and descriptor data in KNIME, minimising the manual reconfiguration required throughout the workflow.

2.4.1 Data Curation

While manually curating input data is the ideal approach to ensuring a high-quality dataset, it is important to include quality assurance checks in the QSPR workflow, especially when handling larger datasets. Such processes ensure that not only are human errors identified, but so too are any unexpected outputs of the data workflow. To begin processing our data, we first convert our structural data into a machine-interactable format. In this project, we used RDKit-based nodes to construct the molecular structure of each input molecule from the provided SDF or the simplified molecular input line entry system (SMILES) representation.[37-39] SDFs provide a single or multiple concatenated Molfile structural descriptions which contain information that may include (but is not limited to) atomic identity, coordinates, and connectivity.[36] SMILES uses rules describing the atoms, bonds, rings, aromaticity, *etc*., to reduce molecular structure into a simple, machine-interpretable text string representation.[39] Information from the structure can be calculated to help identify molecules that may have been misinterpreted. For example, data can be screened for nonsensical molecular descriptor values (such as a molar mass of 0) to exclude data with missing structural information, thereby further cleaning the input data. Furthermore, the input structures are desalted, removing common salts of varying complexity from simple inorganic and organic salts such as alkali metal ions and ammonium to more complex organics such as trifluoroacetic acid from the molecular representation. This ensures that the QSPR is built solely considering organic molecular structures. Making such structural standardisations is necessary to harmonise input structures and enable more effective predictive systems.[35] Another critical analysis is screening for duplicates to ensure the dataset contains unique entries for each molecule. While large databases may use unique identification numbers, the human errors plaguing such datasets render them inadequate for identifying equivalent structures.[40] Instead, identifying duplicates based on the structural representations themselves is necessary. By removing duplicate entries based on their SDF representation recalculated from the RDKit representation, predictive performance is expected to improve by minimising model overfitting to overrepresented chemical spaces. We note that the use of the SDF representations recalculated identified all but one duplicate molecule since SDF generation is not guaranteed to be identical for the same molecule. Instead, canonised SMILES, that is, SMILES representations which have been reordered to afford a consistently unique representation for each molecule, should be used to identify duplicate entries in future models.

To train supervised machine learning models, it is also necessary to include property labels representing the value one wishes to predict. When modelling parameters such as IC_{50} values, there are two approaches to apply these as labels. The first is to use continuous values to build a regression model that predicts an absolute value for the property. The second is to segregate the data into discrete classes based on some characteristic and build a model that predicts to which class the molecule belongs, known as a classification model. Given the narrow range of endpoint values investigated in Chapter 3, which makes for poorer regression models,[41] and the discrete classes investigated in Chapter 4, we focused on building an effective classification framework. The input data can be manually prepared with classification labels; however, to minimise the influence of human error, the workflow was designed such that the classes can be defined using a component to assign the appropriate labels as part of the workflow (Figure 2.6). For example, in the IC_{50} dataset for PLA2-IIA, molecules were classified as active if IC_{50} was less than 100 nM. This classification was selected to afford a model with a well-defined endpoint with clear scientific application. While molecules may be defined as "active" at different thresholds depending on the target, 100 nM is a common threshold concentration indicating a reasonably active inhibitor.[42] Subsequently, the resulting QSAR aims to identify and label molecules as "Active" or "Inactive" depending on their likelihood of falling on either side of this threshold.

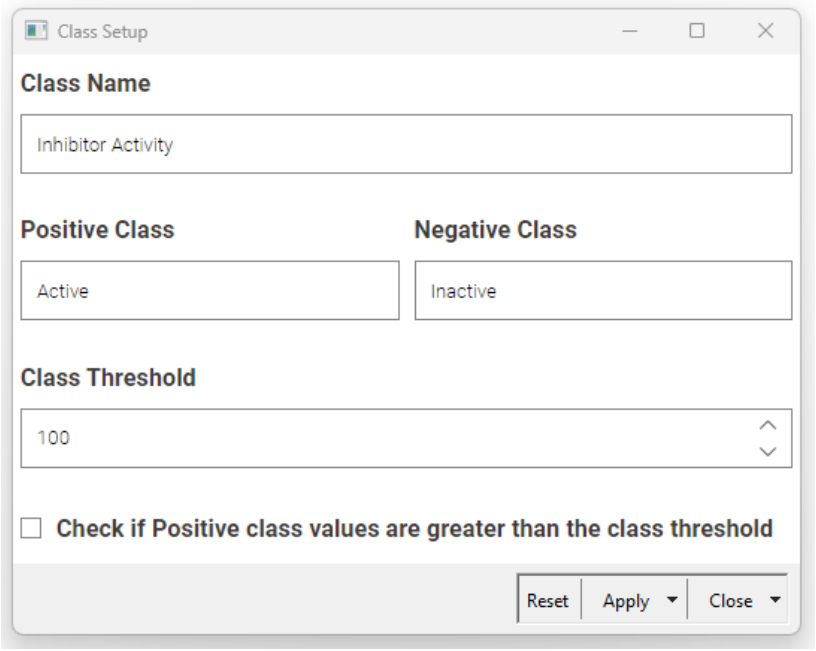


Figure 2.6. Class setup interface in KNIME, allowing the simple assignment and adjustment of data classification thresholds. Data is propagated throughout the framework, minimising the manual reconfiguration required throughout the workflow.

With the molecular structure generated, erroneous rows screened out, and property labels appropriately assigned to each molecule, we can build structural representations more amenable to use in QSAR modelling.

2.4.2 Molecular Descriptors

To build a QSAR model capable of predicting a property from chemical structures, the model must first interpret the structural features of the molecule. Molecular descriptors are the basis of QSAR modelling. Each descriptor captures some representative information about each molecule. By identifying relationships between these molecular descriptors and the target endpoint, one can not only build a highly predictive system but also elucidate the structural features that govern the target property. There exist innumerable molecular descriptors; however, they can often be categorised by their informational content. Zero-dimensional (0D) descriptors are the simplest representations, independent of molecular structure and connectivity, defined by simple numerical properties (e.g., physicochemical properties, atom counts, molecular weight). 1D descriptors describe substructures such as molecular fragments and functional groups, affording partial descriptions of a molecular structure. 2D descriptors fully describe the complete connectivity of the molecule based on some ending of a molecular graph. 3D descriptors capture the geometric arrangement of the structure in addition to its connectivity. Additional information can be incorporated into a single descriptor, yielding four-, five-, and six-dimensional descriptors. While increasing the descriptor dimensionality increases information content, it comes at the cost of increased computational complexity (Figure 2.7).[32, 43, 44]

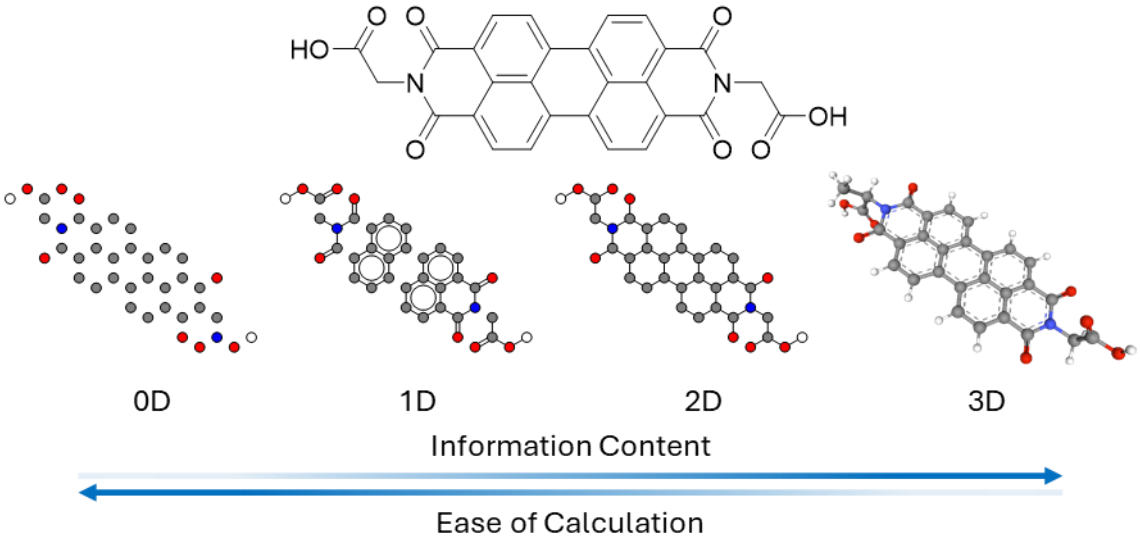


Figure 2.7. Graphical representation of molecular descriptors of varying dimensionality for PBI-G, and the relationship between information content and ease of calculation.[43]

In addition to limitations in calculating high-dimensional descriptors, probing and deconstructing such representations for QSAR applications can be difficult. To ensure that the construction of effective datasets remains accessible, the QSAR framework was built using interpretable, accessible descriptors. While one might expect lower-dimensional molecular descriptors to limit the development of QSARs, they have been instrumental in building many effective predictive systems. One such well-established use of low-dimensional descriptors is the prediction of oral bioavailability, a critical property in drug design, using Lipinski's rule of five.[45] This rule establishes that a molecule is predicted to be orally bioavailable on the basis of molecular weight and LogP (0D descriptors) and the number of hydrogen-bond donors and acceptors (1D descriptors). More recently, Gupta *et al*. developed highly predictive QSPR models using only 0D physicochemical descriptors and 2D fingerprint descriptors.[32] Subsequently, the final workflow built predictive models using these simpler low-dimensional descriptors.

2.4.2.1 Fingerprint Descriptors

Molecular fingerprints are 2D descriptors that can be utilised to encode the structural characteristics of a molecule as a vector. Molecular fingerprints have previously been utilised for chemical similarity comparisons and QSPRs.[46] Molecular fingerprints, such as the Morgan fingerprint, quantify extended connectivity by iteratively describing larger substructures around each atom to capture information about direct connections between atoms as well as the larger structures encoded as an integer.[47] The integer is then converted into a bit-vector fingerprint, which can be processed computationally. The resulting fingerprint encodes the presence or absence of each unique substructure within the molecule, up to a specified radius.[48] It is important to consider that, by reducing the molecular structure to a fingerprint, there is an inherent loss of information due to bit collisions (two different substructures being represented by a single bit). Bit collision will add noise to the analysis and can make the interpretation of important features more difficult, but it can be minimised by utilising a larger fingerprint (Figure 2.8).

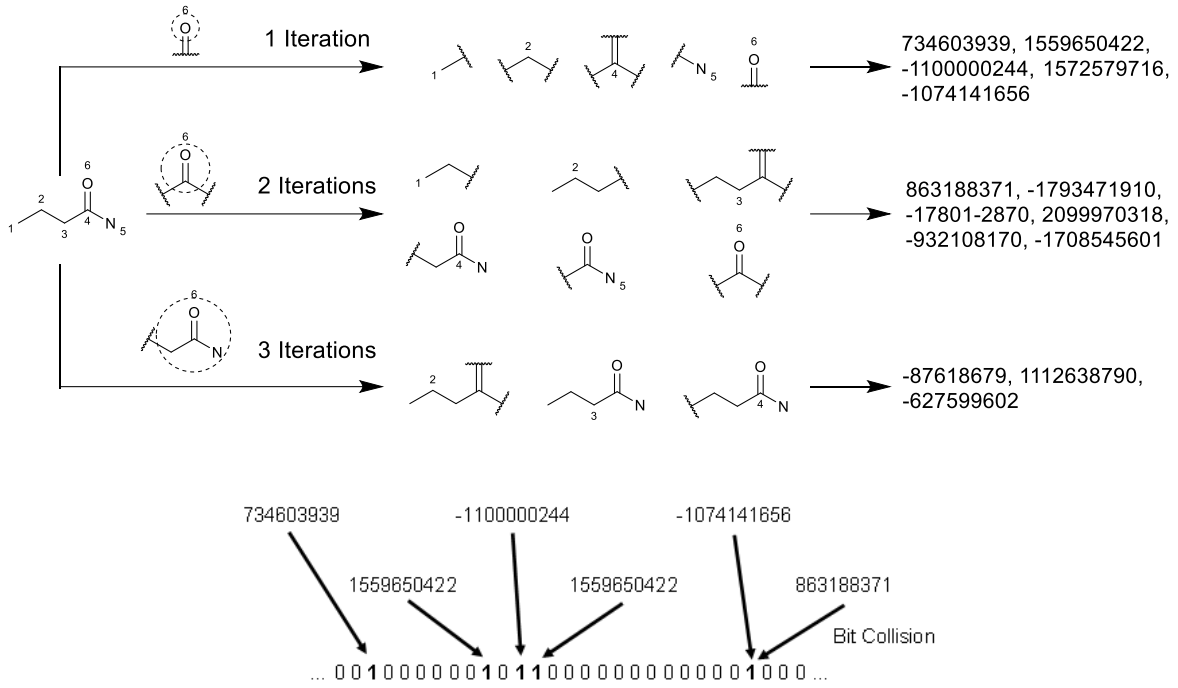


Figure 2.8. Morgan fingerprint generation. Iterative fragments around each atom are assigned integer values. A hash function reduces integers to a bit-vector representation. Adapted from reference [49].

The Morgan fingerprint was chosen for this project because it is the most widely used fingerprint and is among the best performers in small-molecule screening and target-prediction problems.[46] Despite its popularity, the Morgan fingerprint has limitations in capturing global and important structural features in larger molecules.[46] Given that molecular arrangement strongly influences supramolecular assembly, future models may benefit from fingerprints that address such limitations. For example, atom-pair fingerprints are defined by the atomic environments and the shortest path between each pair of atoms in a molecule.[50] More recently, the development of the minhashed atom-pair fingerprint unifies the circular connective fingerprints, which are effective at describing small and topological molecules, with those for large molecules.[46] Therefore, future QSAR frameworks used to model supramolecular systems, which are governed both by local chemistry and global topology, may greatly benefit from such unified descriptors.

2.4.3 Modelability Index

While it is obvious that changes to the QSAR workflow will result in varying predictivity of the resulting model, so too will the constitution of the dataset influence the predictive power of the QSAR. Given that cheminformatic systems employ many descriptors, machine learning techniques, and validation investigations in combination, QSAR systems can require a reasonable amount of time and resources to train adequately.[51] A significant difficulty in QSAR modelling is the presence of “activity cliffs”. These activity cliffs are

described by pairs of structurally similar molecules that differ significantly in the modelled property of interest. These cliffs represent discontinuity in the property landscape, which can make model optimisation and property prediction much more difficult.[52] To probe the quality of the dataset and its amenability to modelling, an *a priori* estimation of QSAR modelling feasibility can narrow the scope of investigation to datasets more likely to generate performative models. The modelability index (MODI) is one such metric that allows a dataset to be quantitatively assessed at minimal computational cost, assessing its ability to produce predictive QSARs.

The Morgan fingerprints of the bit integers were used to calculate molecular similarity for the modelability index (MODI).[51, 53] The Manhattan distance was used as a chemical similarity metric, as it has been previously determined to be the better metric for calculating chemical similarity in high-dimensional space, such as the Morgan fingerprints described in Section 2.4.2.1 which can be considered as a vector described by 1024 dimensions.[54] The Manhattan distance (D) of two fingerprints (**p** and **q**) is calculated by summing the absolute difference between each pair of corresponding coordinates (i) (Equation 2.1).

$$(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i| \qquad \text{(Equation 2.1)}$$

By calculating the Manhattan distance between each pair of Morgan fingerprints, and determining the lowest distance combination for each molecule, the most chemically similar molecule could be identified for each training set molecule. The number of times the classification (c = 2 for binary classification) of the i^{th} activity class is the same as nearest neighbour (N_i^{same}) is compared to the number of molecules in that same class (N_i^{total}). The average ratio of these across each class describes the MODI (Equation 2.2).

$$\text{MODI} = \frac{1}{c} \sum_{i=1}^{c} \frac{N_i^{same}}{N_i^{total}} \qquad \text{(Equation 2.2)}$$

For the PLA2-IIA dataset, classifying activity at the 100 nM threshold yields a MODI of 0.80. The typical threshold for expecting an effective dataset is 0.65.[51] This provides further confidence in the use of the PLA2-IIA IC_{50} molecules as the benchmark model for the framework. This MODI result indicates that there are relatively few activity cliffs across the property landscape, suggesting it should be much easier to build predictive models.

2.4.4 Applicability Domain

A significant consideration in QSAR development is defining appropriate applicability domains (ADs), as highlighted by the OECD's recommendations (Chapter 1, Section 1.4.1).[55] Based on the QSAR model's desired endpoint, a set of molecules with known properties can be collected to form the dataset. These molecules must cover a range of structures so that the model is suited to making predictions within the same chemical space, known as the applicability domain. Interpolated predictions are more reliable than extrapolated predictions, meaning predictions within the applicability domain are more reliable. The reductionist nature of QSPRs means that there will always be limitations to the types of structures that can be reliably predicted from the molecules used to build the model. Given that a molecular fingerprint is directly related to the chemical structure, measuring the similarity between chemical fingerprints is an effective way to assess molecule similarity and establish the applicability domain. Early iterations of our model utilised Tanimoto coefficients (T_C) because they are among the most effective measures of chemical similarity. The Tanimoto coefficient for dichotomous variables is described in (Equation 2.3 (T_c of molecule A and molecule B ($T_{C\ (A,B)}$) is defined by the number of *on* bits (those which indicate the presence of a structure represented by the hashed bit) unique to molecule A (a), the number of *on* bits unique to molecule B (b), and the number of *on* bits shared by molecule A and molecule B).[56]

$$T_{C\ (A,B)} = \frac{c}{a + b - c} \qquad \text{(Equation 2.3)}$$

The greatest T_C for each molecule is compared to a predefined threshold. Molecules exceeding the threshold are considered chemically similar to the training set, and it is expected that the QSPR will yield reliable predictions. While more recent discussions have shown that T_C depends on the fingerprint used, a threshold of 0.85 is still widely used as a measure of molecular similarity.[57] However, only 40% of the test molecules were deemed to lie within the calculated applicability domain despite the chemical similarity across the PLA2-IIA IC_{50} test set, a large proportion of which containing indole or quinoline derived structures. This indicates that simply choosing a threshold of 0.85, as reported in the literature, is insufficient to define the domain. This should be expected, as the range of similarity measurements depends not only on the distance measure but also on the type of

molecular fingerprint, so a one-size-fits-all threshold for the Tanimoto coefficient will not represent the data well.[57]

To overcome the limitations of the Tanimoto coefficient in defining the applicability domain, other well-defined domains were considered for future models. Sahigara *et al.* characterised several ADs, including better-defined distance-based approaches. These methods involve defining the average point across each vector of the training set fingerprints, known as the centroid. The maximum distance, average distance (d) and standard deviation (s) between the centroid and training data are then calculated. From these, a series of ADs can be defined, including: simply using the maximum distance as the boundary (maxdist); using multiples of the average distance (2d and 3d); using d+(s×z) with an arbitrary parameter z, typically 0.5 (dsz); and choosing a boundary which captures the 95th percentile of the test data (p95) (Figure 2.9).[58]

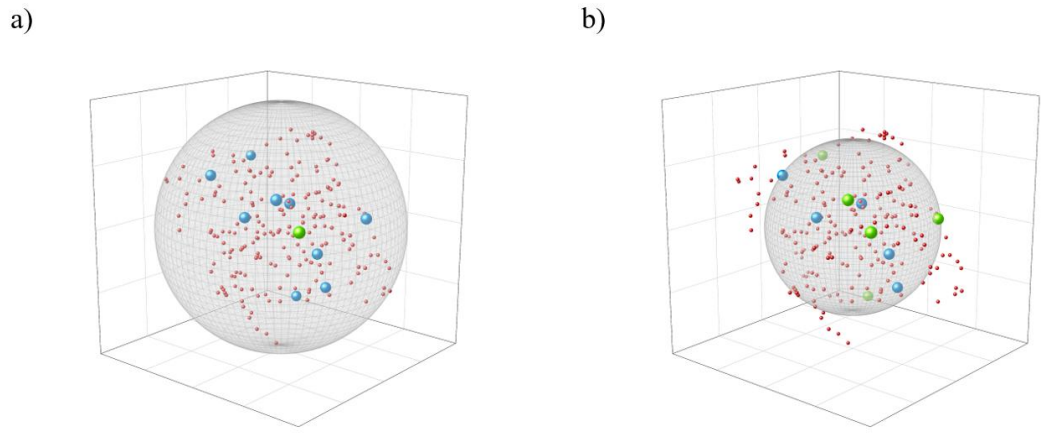


Figure 2.9. A plot of hypothetical chemical data described by three descriptors, each represented by an axis, where the red, blue, and green data represent training, in-domain test, and out-of-domain test data, respectively. Using the training data, well-defined domains of applicability can be calculated that can be used to identify the proximity, and therefore chemical similarity, of the test compounds to the training set. a) illustrates the maxdist boundary, while b) illustrates the more pessimistic dsz boundary.

Again, given that there is no single AD that fits all datasets, providing a set of well-defined thresholds enables assessment of their impact on model predictions. One would expect those within an appropriate boundary to be better predicted by the model; consequently, the most appropriate boundary can be selected based on which best excludes the poorly described molecules, thereby improving the model's predictive statistics. By selecting well-defined metrics based on the data for each predictive problem, rather than standard literature thresholds independent of the model of interest, those molecules best described by the model can be identified, and the confidence of the prediction can be greatly improved.

2.4.5 Dataset Preparation

The final step before modelling can begin is to prepare the datasets. An important step is partitioning the dataset to enable effective assessment of the model's performance. A machine learning model's prediction will be biased by the data used to train it, leading to overfitting to the training data and reducing the reliability of its predictions as a metric of performance on independent data. Subsequently, the metrics used to assess the model's performance on the training data will overestimate its predictive ability. Instead, a dataset independent of the training procedure but similar to the training data is required to estimate the model's predictive performance.[59] By selecting models based on their performance on unseen data, a predictive system can better interpolate the training data to make *a priori* predictions of the property of interest. A larger training set can help the model learn more effectively, while a larger testing set may provide more diversity from which a fairer assessment of model performance can be made. To achieve a good balance, typically 80 percent of the input dataset is used to train the model, while the remaining 20 percent are used as the independent test set.

It is important to consider the number and quality of descriptors used to build the QSAR relationship. It is necessary to ensure that each descriptor provides useful insight to the model. The amount of information provided by a given descriptor can be easily measured by calculating its population variance (σ^2), defined by the value of each descriptor (x), the average value of the descriptor across each molecule ($\bar{x}$), and the number of molecules (n) (Equation 2.4).[60] In the extreme case where a threshold variance of 0 is chosen, only descriptors that are homogeneous across each molecule are filtered. As this threshold is increased, descriptors with increasing variability are removed. Therefore, it is important to strike a balance between filtering descriptors that provide only noise and those that contain key information for finding a relationship between the descriptor and the property.

$$\sigma^2 = \frac{\sum(x - \overline{x})^2}{n} \qquad \text{(Equation 2.4)}$$

A significant error often made across QSPR development is the use of highly collinear descriptors.[14] Without addressing such descriptors, significant bias can be introduced into model training (depending on the machine learning method), as the same information is used twice, detracting from the model's ability to effectively interpret a robust relationship between the descriptors and the endpoint.[14] To mitigate collinearity, Spearman's rank correlation (r_s) was used to identify redundant descriptors for filtering (Equation 2.5).[61] For

each descriptor, the numerical values across the molecules are replaced with their ranks, i.e., their positions in the sorted list of that variable's values across all data points. The strength of the monotonic association between each of the ranked variables is then quantified using the difference between the ranks in each observation (d_i) and the number of observations (n).

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \qquad \text{(Equation 2.5)}$$

Calculating Spearman's rank correlation is trivial and well-defined. However, there is no clear consensus on the choice of an appropriate intercorrelation threshold across QSPR studies.[62] Filtering descriptors with rank correlation near unity ensures only highly correlated descriptors are removed. For each group of descriptors deemed to be highly correlated, the descriptor with the most correlated columns is carried forward, and all other correlated columns are filtered out. This procedure is repeated until no more correlated columns can be identified (Figure 2.10).

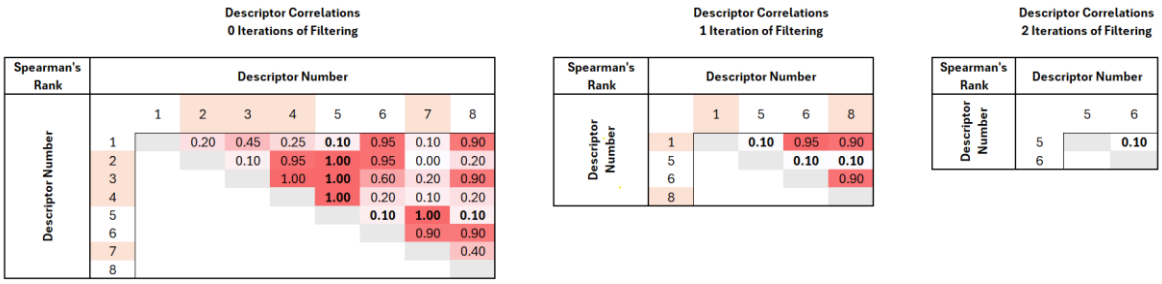


Figure 2.10. Correlation filtering on illustrative data using a correlation threshold of 0.9. With the greatest decree of correlation between descriptors (as determined by Spearman's rank), descriptor 5 is carried forward and the highly correlated descriptors 2, 3, 4, and 7 are removed. Following the first iteration, correlations remain, with 6 being the most highly correlated descriptor, therefore it is retained, while descriptors 1 and 8 are removed. Descriptors 5 and 6 have a correlation below the target threshold and so filtering is complete.

As the threshold is lowered, descriptors with a lower correlation are selectively filtered, which may further remove redundant descriptors. However, as the threshold is lowered further, the performance of the resulting models may be diminished because crucial information is lost. It has been shown that QSPR studies use thresholds ranging from 0.70 to 1.00, which can significantly affect the resulting models. Such thresholds are often omitted from reporting, making it difficult to reproduce these models.[62] Therefore, we report here that we utilised a variance threshold of 0.1 and a correlation threshold of 0.9 throughout the project. By applying these thresholds to the PLA2-IIA training set, the 1024-bit Morgan fingerprint was significantly reduced: 100 descriptors after the low-variance filter, then 87 after the subsequent correlation filter. While the number of descriptors can be further reduced by techniques such as principal component analysis (PCA), its use in our models was

avoided as it can reduce the interpretability of the resulting model. Principal components are linear combinations of the original descriptors that best explain the variance across the variables, provides an approximation of the original data using fewer major components.[63] While the reduction of data dimensionality may reduce overfitting and improve predictive performance, the principal components may represent contributions from several of individual structural features which is difficult to deconvolute. Subsequently, to ensure that descriptors remain as direct interpretations of chemical structure PCA was avoided, however future models may be able to incorporate PCA with the subsequent reconstruction of the original parameters after modelling to achieve the improved modelling performance with complete interpretability.[64]

Not only is the quality of the descriptors important to consider, but so too is the quantity. As the number of descriptors involved in a calculation increases, the modelling process becomes increasingly time-consuming.[65] Furthermore, including too many descriptors can result in models which are susceptible to overfitting and chance correlations.[66] The concept of "Occam's Razor" is a useful consideration in QSAR development. The ideal model uses a small number of simple molecular descriptors.[41] The filtering methods above increase the robustness of these models by improving both descriptor quality and quantity, thereby affording a simpler model. To ensure a model has sufficient data for the number of descriptors used, a guideline known as the "Topliss and Costello rule" recommends 5 molecules per descriptor for simple linear regression methods.[41] While the abundance of biochemical databases makes such a guideline easy to achieve in QSAR models, property prediction problems with very limited data may fall far below the ideal ratio. Subsequently, not only are predictive performance metrics necessary, but robustness metrics are also required to identify overfit models. Such metrics are discussed later in this chapter.

2.4.6 Class Balancing

Once the molecular descriptors are filtered to remove highly correlated and redundant descriptors, the molecules and their classes must be optimised. When building a classification model, the relative frequency of each class can affect the model's effectiveness. In pharmaceutical and chemical research, it is often the case that only a small number of molecules exhibit the property of interest, resulting in a minority class with few data points from which to extrapolate.[10] This limitation means that even with *a priori* knowledge of the class imbalance, adequate sampling of both classes is not always possible, in which case *post hoc* approaches must be applied. The most common *post hoc* methods involve down-

sampling the majority class or up-sampling the minority class (Figure 2.11). There are various ways to achieve this, though there is no universal approach due to significant dataset dependence.[67] While it is important to ensure class imbalance is addressed in the training partition of the data, the test set should retain a class imbalance representative of the original data, to better assess predictive performance under the expected dataset distribution.[67]

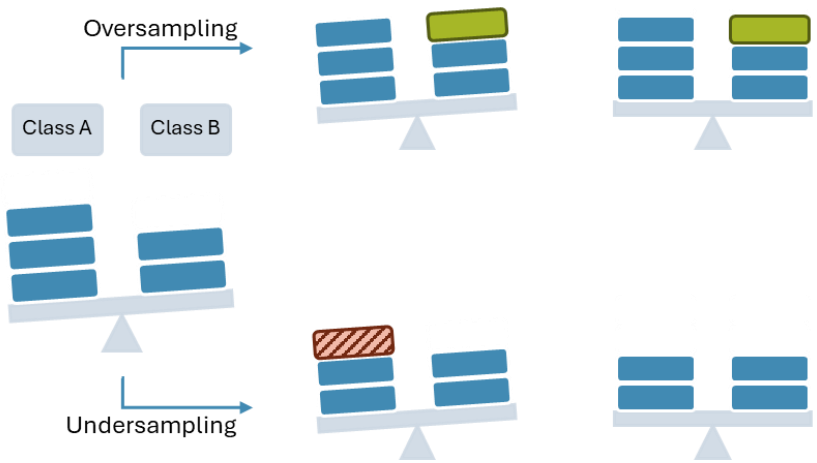


Figure 2.11. Cartoon illustrating the difference between undersampling and oversampling techniques to resolve class imbalance.

2.4.6.1 Undersampling

Undersampling is a technique that reduces the number of majority-class samples to address class imbalance. By reducing the number of samples in the majority class, the ratio of samples in the minority class is artificially increased, thereby improving the model's ability to predict the minority class accurately. In addition, undersampling techniques can reduce model training time by decreasing the dataset size while improving the model's robustness to noisy and outlying data points. These benefits can come at the cost of losing important information necessary to build effective models, particularly in small datasets, where any loss of information can significantly impact performance.[68-70] Furthermore, as datasets grow in size, undersampling becomes less effective, potentially limiting its applicability across datasets of all sizes.[68] As the simplest and easiest to implement, we utilised random undersampling, which randomly removes a portion of the majority-class samples until the target majority:minority class ratio is achieved, resulting in a more balanced dataset. Random undersampling has previously been utilised across chemical prediction problems in drug discovery.[68] However, due to the impacts of data loss and model intricacy, future models may benefit from undersampling methods, which can better consider the samples removed to preserve the most valuable modelling information, such as the NearMiss[71] and Tomek Links approaches.[72]

2.4.6.2 Oversampling

In contrast to undersampling, oversampling addresses class imbalance by duplicating or generating synthetic minority-class data while preserving the information available from the majority class. This allows the original feature distribution to be maintained while minimising the model's overfitting to the majority class. However, such methods can be computationally expensive, introduce excess noise to the data, and hinder the modelling of complex decision boundaries.[68] An early and popular approach to class imbalance is the synthetic minority oversampling technique (SMOTE).[73] Using the average of a randomly selected data point and its nearest neighbours from the minority class, a new synthetic data point upsamples the minority class. The new synthetic samples are very similar to other samples in the dataset, allowing the minority class to be augmented without further expanding the dataset. This is achieved by selecting a data point from the minority class and creating a new sample of the same class between one of its nearest neighbours (Figure 2.12). SMOTE has been effective in material design, having previously been used to build predictive models of mechanical properties in polymer materials.[74] The limitations of SMOTE have led to the development of more refined methods to handle class overlap and decision boundary complexity better. Introducing such techniques in future iterations of the QSAR framework may lead to more effective dataset class balancing, yielding better predictive performance of the output models.

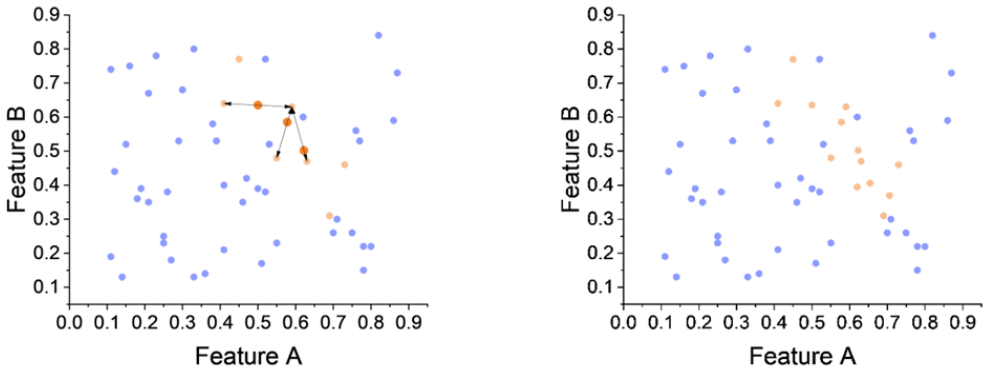


Figure 2.12. Illustration of the SMOTE oversampling technique. The orange minority class is augmented by generating new data points between neighbouring samples of the minority class.

One such refined oversampling method is the adaptive synthetic (ADASYN) sampling algorithm. Like SMOTE, ADASYN augments the minority class to balance the dataset; however, ADASYN focuses on augmenting regions of chemical space that are harder to model. While ADASYN may improve future models, it is not currently implemented in KNIME, either natively or through its extensions. While the open-source nature of KNIME

means a range of community-developed nodes are available, these nodes are not well-maintained or stable like those curated by KNIME and its affiliates. For example, while we implemented a community node used to enable ADASYN class balancing using Python[26] and the imblearn[75] library, changes to the imblearn library since the component's development rendered the code inoperative. While our knowledge of the Python programming language enabled the error to be identified (the renaming of the "fit_sample" attribute to "fit_resample" in the current versions of the imblearn library),[76] the typical chemist may not be able to perform such troubleshooting, which could make the framework unusable. Subsequently, we focused on implementing stable nodes provided by KNIME and its affiliates, where possible, to ensure the framework's long-term stability.

2.5 Machine Learning Model Training

With an optimised training set, the machine learning models can be trained. As discussed in Chapter 1, an ever-increasing number of machine learning models are being made available, many of which are easily implemented in KNIME. Given that the focus of this project is property prediction, the most appropriate machine learning model type is supervised learning. Therefore, the goal of the machine learning models is to establish a relationship between the optimised descriptors and the property class provided for each molecule. Each machine learning method has its own advantages and limitations. If one model struggles to capture the subtleties in a particular region of chemical space, combining multiple models can yield more reliable classification predictions. Using various models to confirm predictions can greatly enhance the accuracy of results on unseen candidates.[3] Identifying multiple effective predictive models also boosts our ability to design new materials with specific properties, as averaging the outputs of several good models typically yields results closer to real measurements than relying on a single model.

At the core of machine learning training is the optimisation of hyperparameters. Hyperparameters are variables that control and configure the behaviour of each machine learning model. The impact of hyperparameters on the predictive performance of the resulting models can be significant, yet despite this, their optimisation remains a challenging task across the field of machine learning.[77] Herein, we describe the models used throughout this project, the advantages and limitations of each (Table 2.1), and the processes by which hyperparameters were optimised.

Table 2.1. Comparison of popular machine learning classification models.[78, 79]

Model	Advantages	Disadvantages
k Nearest Neighbours	Simple No training phase Adaptable	Slow with large datasets Sensitive to noise
Decision Trees	Easy to interpret Handles non-linear data	Prone to overfitting Unstable
Random Forest	Reduces overfitting Robust Handles large datasets	Slower to train Less interpretable than a single tree
XGBoost	High accuracy Effective for complex patterns Handles non-linear data	Computationally expensive Prone to overfitting
SVM	Handles high-dimensionality Flexible kernels	Slow with large datasets Sensitive to the choice of kernel Sensitive to parameters
Naïve Bayes	Simplicity Handles high-dimensionality Interpretable	Incorrect assumptions Susceptible to data imbalance
Logistic Regression	Fast Interpretable Effective for linear problems	Assumes linearity Ineffective with non-linear data

2.5.1 K-Nearest Neighbours

Nearest neighbour classification is perhaps the simplest of the models used. This model will take the unclassified data and, using the distance between descriptors, assign the test molecule to the class of its nearest training point.[80] K-nearest neighbours (kNN) expands upon this by considering a majority vote among the nearest neighbours (Figure 2.13). This makes kNN a simple model to implement, as it requires only optimising the number of neighbours to consider and is easy to interpret. A balance must be found: a large k will provide more information to determine the class, but may also lead to the inclusion of data that is not close enough to yield an accurate estimate.[80] Since kNN must assess the distance between each pair of molecules to identify the nearest neighbours, kNN models become increasingly more expensive as the dataset size increases.

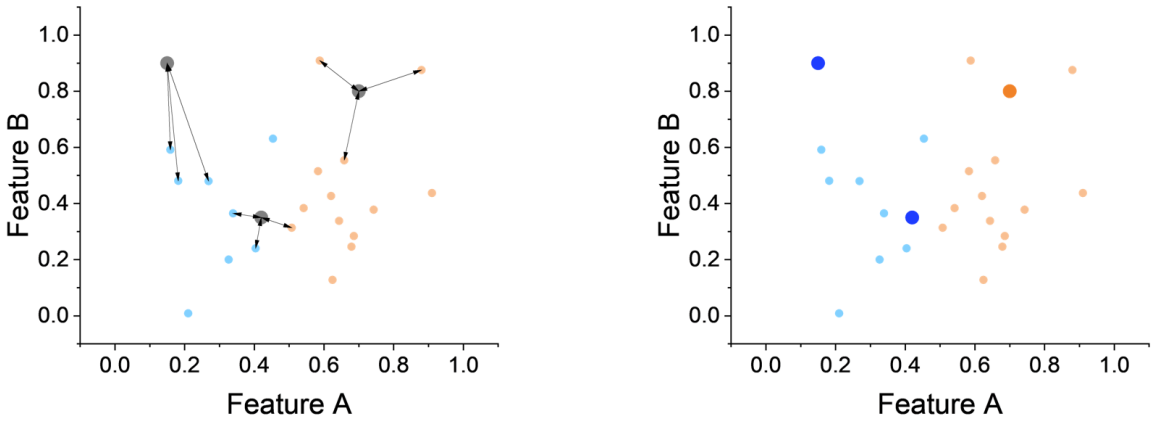


Figure 2.13. Visualisation of k nearest neighbour classification, where k = 3. Classification of an unknown molecules property is made by considering the consensus of class of the three nearest neighbours.

2.5.2 Decision Trees

Decision tree learning is a supervised model, the foundations of which were laid out by Breiman *et al*. and have since been used successfully throughout industry.[81, 82] These models have proven popular, in part due to their interpretability, even for those unfamiliar with machine learning; they can process a variety of data and handle large-scale datasets. Their fundamental interpretability makes them a good starting point for classification problems. Decision tree algorithms begin by searching for the best possible split from among the available variables. The variables will be compared, and the split that yields the optimal labelling of the data will be retained. In the case where all data points in a cluster are of the same class, the cluster is known as a leaf. If a branch creates a cluster which is not a leaf, the process is repeated. Another way to interpret how a decision tree classifies data is to geometrically partition the descriptor space, making the populations of each rectangle more homogeneous.[82] Figure 2.14 illustrates a basic decision tree classification and its geometric interpretation.

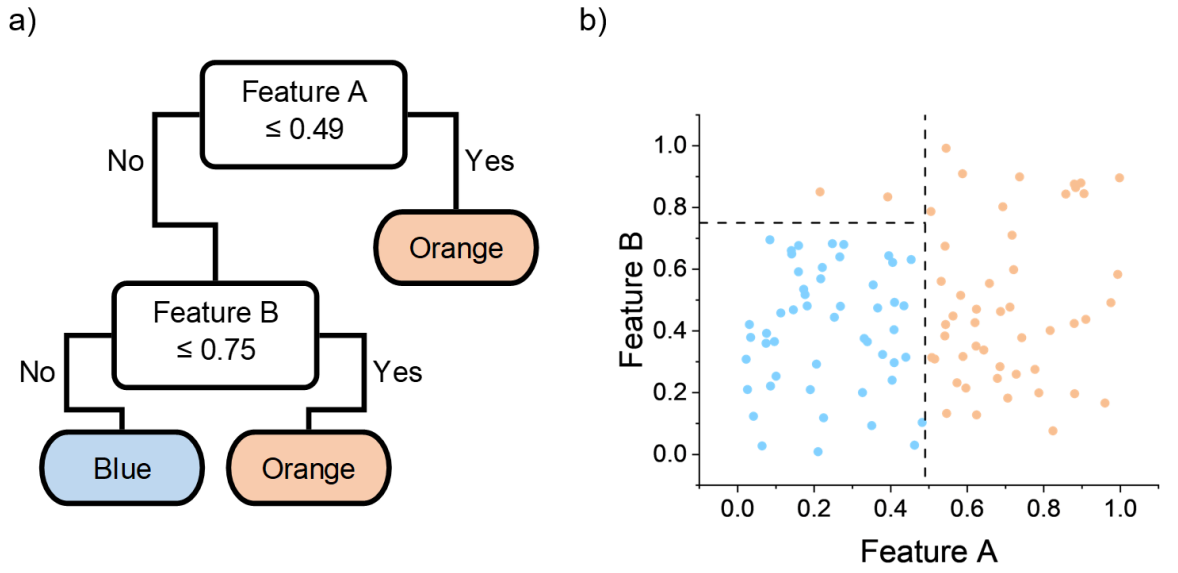


Figure 2.14. Visualisation of the decision tree splitting procedure. (a) shows a single decision tree with describing the rules of classification. (b) is the analogous geometric interpretation, illustrating the classification rules as decision boundaries segregating classes.

Individual decision trees are generally not robust when trained on different subsamples of the data; thus, selecting different subsets of descriptors to train a tree can yield notably different classifications.[81]

2.5.2.1 Ensemble Learning

Decision tree models can be improved by combining multiple decision trees and allowing them to vote, classifying data based on the consensus of the models, a process known as ensemble learning.[83] As the number of descriptors and complexity of the data increase, the data becomes increasingly difficult to partition and more prone to overfitting. Random forest (RF) models overcome this by generating a set of unique trees, each selecting a different subset of descriptors and samples. The algorithm then makes a prediction with each decision tree in the forest and outputs the class that is most frequently predicted across all trees. This ensemble method is less susceptible to noise from individual outliers, providing a more robust generalised model (Figure 2.15). However, this improved performance comes at the cost of increased complexity. The large number of trees generated by ensemble models is intrinsically less interpretable than a single tree. Therefore, the introduction of more complex models necessitates the development of feature importance metrics to ensure such models can be used to develop useful design principles.

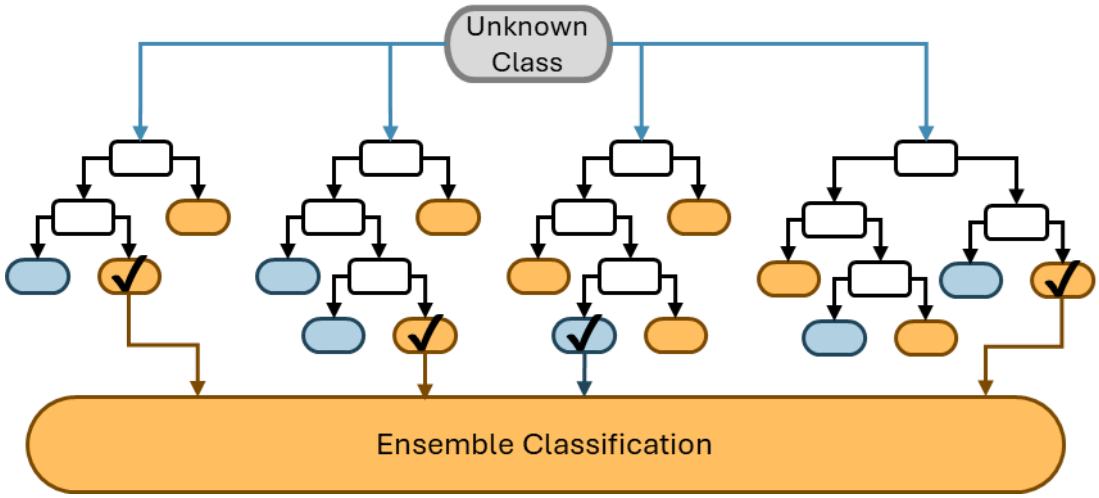


Figure 2.15. Visualisation of ensemble decision tree classification. Multiple trees are developed, each making its own prediction of the property classification. The majority consensus provides the final classification output.

2.5.2.2 Gradient Boosted Learning

Gradient boosted decision trees were first described by Friedman *et al.*[84] These models, which include the popular XGBoost algorithm, are also ensemble methods that, instead of using a consensus of independent trees, sequentially develop trees by considering the errors of previous iterations and summing them to approach a high-performance model (Figure 2.16). XGBoost models have already proven effective for predicting chemical systems.[85]

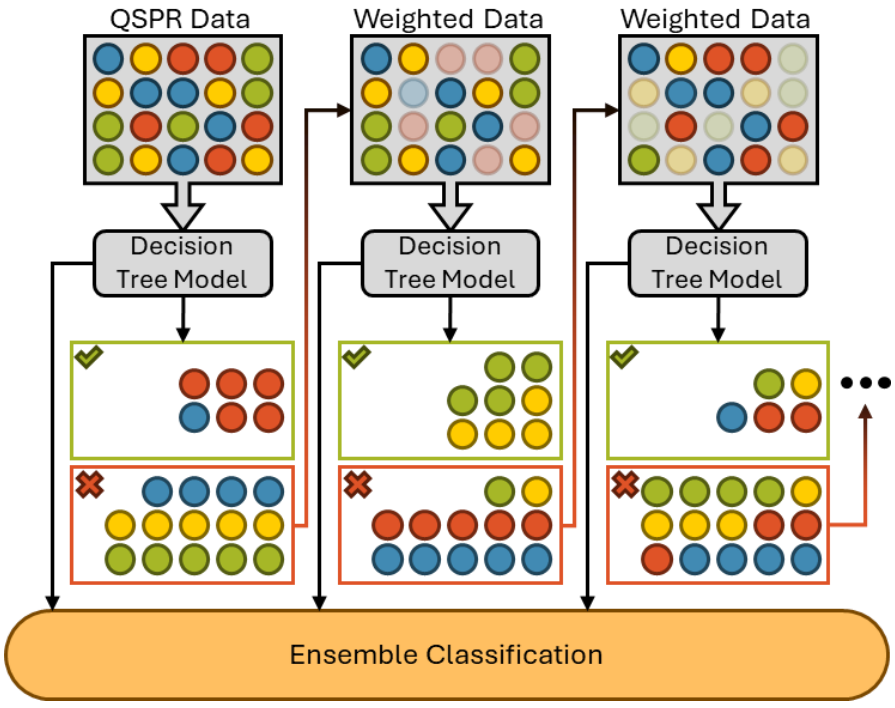


Figure 2.16. Visualisation of the gradient boosted decision tree machine learning procedure. An initial decision tree model is developed using the entire training set. Samples that are poorly modelled are given a greater weight in the development of a new model. The process is repeated, affording a series of trees modelling the errors of each previous iteration. The consensus of each tree yields the gradient-boosted classification.

While complex methods such as XGBoost are more difficult to understand and tune than other models, their superior performance and versatility can make their development a worthwhile investment. While XGBoost offers strong predictive performance on large, complex datasets, smaller datasets may be more appropriately analysed with simpler models.[86]

2.5.3 Support Vector Machines

Support vector machines (SVMs) are a well-established supervised learning model, which, along with the random forest algorithm, has proven to be one of the pillars of machine learning in recent decades.[81] The simplest form of this is the linear SVM, which uses training data to learn a boundary that separates the classes. This boundary takes the form of a hyperplane, which, in n-dimensional Euclidean space, is a flat n-1-dimensional plane that divides the space into two parts.

To identify the best hyperplane, not only must the SVM minimise the number of classification errors but also calculates the largest margin between the hyperplane and the data points closest to the boundary, known as the support vectors. Figure 2.17 makes it clear that by maximising the margin, a more optimised classification boundary is determined.

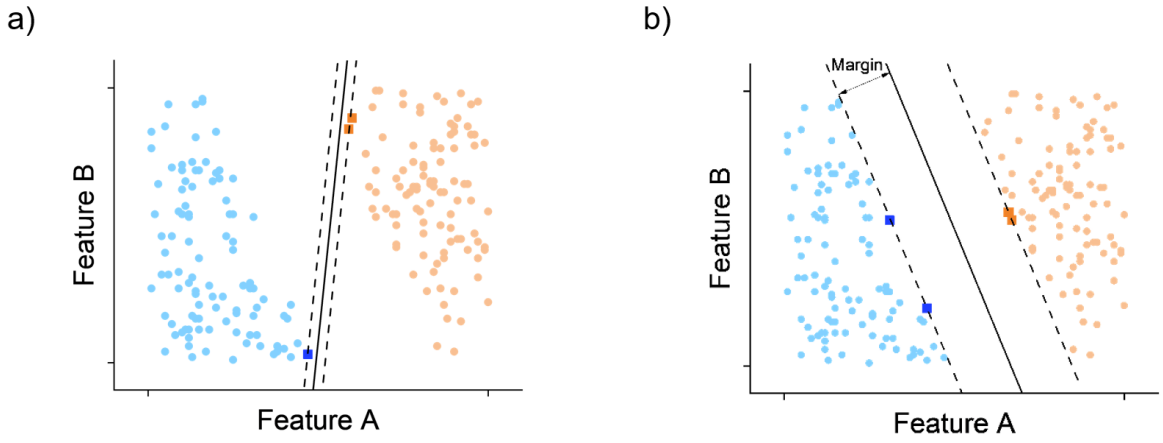


Figure 2.17. Visualisation of linear SVM model, with support vector datapoints represented with square markers in each case. a) A potential classification boundary with a small margin between the support vectors. b) The optimised classification boundary with maximised margin between support vectors.

While this model is easy to interpret, they are too restrictive to be meaningfully applied to most classification problems. While the construction of a non-linear boundary would rectify this issue, such a problem quicky become computationally infeasible. Kernel SVM circumvents this issue by mapping the data into a higher-dimensional descriptor space, where an optimal hyperplane can be determined with comparatively low computational cost (Figure 2.18). The radial basis function (RBF) kernel is one of the most commonly used as

it relies only on the Euclidean distance between each data point (x_j) and the decision hyperplane (x), as opposed their exact relative geometry to simplify the model, requiring the optimisation of a single hyperparameter (σ) which controls the complexity of the decision boundary (Equation 2.6).[81, 87]

$$RBF = \frac{\|x_j - x\|^2}{2\sigma^2} \qquad \text{(Equation 2.6)}$$

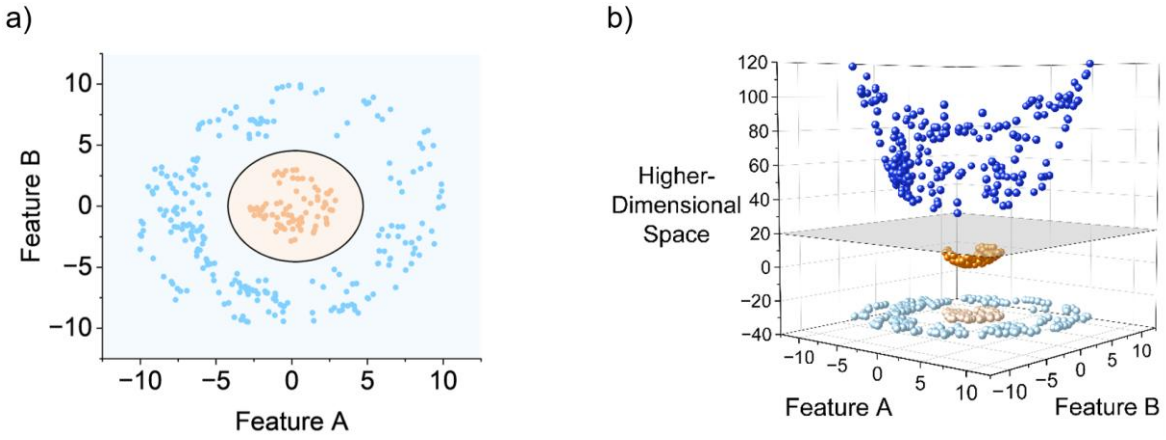


Figure 2.18. Visualisation of the SVM model. a) Data representing two classes, denoted by colour, in a 2D descriptor space wherein no adequate linear hyperplane can be constructed. b) An additional variable is created by calculating the squared sum of each descriptor value. A hyperplane can now easily separate the classes in the higher-dimensional space, representing a circular boundary with radius r = $\sqrt{20} \approx 4.5$ in the original 2D descriptor space.

Support vector machines are particularly popular due to their exceptional ability to generalise classification problems and build effective models even when the input dataset is small. However, as datasets become increasingly large, so too do the training time and cost of these models, making them prohibitively expensive to develop. Furthermore, SVM models are particularly sensitive to class imbalance, leading to poor predictive performance.[88]

2.5.4 Naïve Bayes

Naïve Bayes is a straightforward statistical model used to make classification predictions by considering the class frequency and the likelihood of observing each descriptor in every class. The classifier counts how often each descriptor appears in each class, then combines those counts to determine which class is most probable for a new molecule. It assumes that each descriptor contributes independently given the class, multiplies the resulting likelihoods across classes, and selects the class with the highest product (Equation 2.7). In

QSAR, descriptors may not be truly independent. For example, physicochemical descriptors depend on molecular structure, so the inclusion of both physicochemical and structural descriptors may overrepresent a specific feature of the molecule. Therefore, performance can diminish as the assumption is increasingly violated.[89] Despite their simplicity, naïve Bayes classifiers remain effective across classification problems, and have been found to work better than some complex models in the presence of noisy data.[79, 90] While naïve Bayes models are less prone to overfitting than other models, particularly for those trained on small datasets, their performance is particularly sensitive to class imbalance.[89]

$$P(Class \mid Descriptors) \propto P(Class) \cdot \prod_{i=1}^{n} P(Descriptor_i \mid Class) \qquad \text{(Equation 2.7)}$$

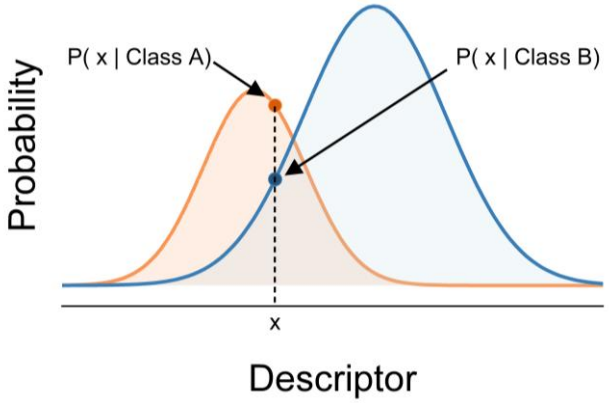


Figure 2.19. Visualisation of a Gaussian naïve Bayes classifier. The highlighted points describe the probabilities of observing a value of "x" for a given descriptor if it came from the Class A or Class B distributions.[91]

2.5.5 Logistic Regression

Logistic regression is another simple model that models the chance of an outcome based on the weighted values of each descriptor. The log odds ratio describing the probability that the molecule falls within the positive class (π) is defined by the sum of each descriptor (x_i) and its weighted coefficient (β_i) (Equation 2.8).

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m = z \qquad \text{(Equation 2.8)}$$

By algebraic manipulation, the probability that a molecule falls into the positive class can be expressed in the form of a sigmoid curve. The slope parameter (z) is defined by the descriptor values and their coefficients and describes the steepness of the sigmoid curve (Equation 2.9). Building a logistic regression relationship requires the estimation and the optimisation of the weighted coefficients (β_i) such that the slope parameter (z). These optimal values can be

identified by minimising the prediction error, as quantified by metrics such as the negative log likelihood (Figure 2.20).[92]

$$\pi = \frac{1}{1 + e^{-z}} \qquad \text{(Equation 2.9)}$$

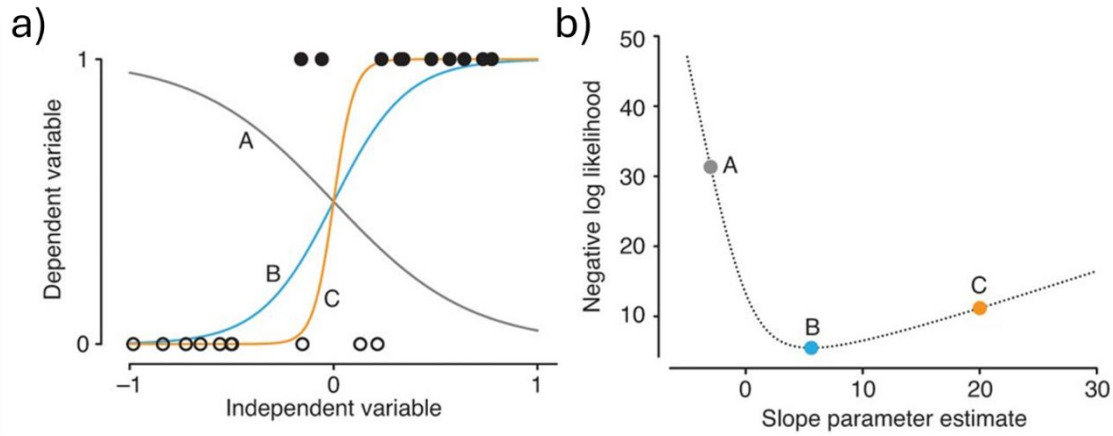


Figure 2.20. A non-separable dataset with different logistic curves using a single slope parameter. The slope parameter for each logistic curve (a) is indicated by the coloured point in (b), shown with its associated negative log likelihood. A minimum is found for the ideal curve (blue). Reproduced from Lever *et al.*[92]

Logistic regression has several advantages, including its simplicity, interpretability, and training efficiency. There are a few drawbacks that limit its application to QSPR. Primarily, logistic regression assumes linearity between the descriptors and the predicted value. Where there is no clear linear boundary, there is no solution that can be accessed by logistic regression. Given that linearly separable data are rare in real-world scenarios, there may be few QSPR problems for which such models perform well. Furthermore, while logistic regression is robust to overfitting in simple models, as the number of descriptors increases, logistic regression quickly becomes prone to overfitting.[93]

2.5.6 Hyperparameter Optimisation

Hyperparameter optimisation is a crucial step in QSPR, and it can be an intensive process. The goal of the optimisation procedure is to find the combination of hyperparameters which maximises the chosen scoring metric (discussed in Section 2.6.1) estimated by internal cross-validation (discussed in Section 2.5.7), which should lead to the most predictive model. While historically, manually searching and assessing hyperparameter combinations was used to identify effective models, such procedures have significant drawbacks. Not only is such an approach time-consuming, but it also requires significant expert intuition. Furthermore, because such investigations are guided by the individual's experience, independently

reproducing the hyperparameter optimisation results is difficult. It is clear, therefore, that automated methods are necessary to ensure the accessibility of our QSAR system.[94] There are several automated approaches to hyperparameter optimisation that help overcome these limitations. One of the most popular and simplest approaches is a brute-force grid search of the hyperparameter space. By specifying a start, end, and iteration step for each hyperparameter, every combination allowed by those values is investigated. This enables a thorough search of the parameters to identify the best combination within the predefined space, with reproducible results. However, such a search incurs a higher computational cost, with training time growing exponentially with the number of parameters. Much of the computation time is often spent evaluating large areas of the hyperparameter space which are ineffective, making the training inefficient.[94]

This can be improved by implementing a Bayesian optimisation strategy, one of the most popular automated approaches to hyperparameter optimisation.[95] Instead of evaluating every combination of parameters as in the grid search method, the parameters are randomly sampled in a preliminary search. Subsequently, a focused search on promising combinations yields the optimised hyperparameters. We found that when training a random forest model using the PLA2-IIA dataset, Bayesian hyperparameter optimisation could find models with predictive performance comparable to that of a more thorough grid search, with much fewer iterations, greatly increasing search efficiency (Table 2.2, Figure 2.21).

Table 2.2 Comparison of hyperparameter search methods. The maximum balanced accuracy statistics following 5-fold cross-validation of Random Forest models on the PLA2-IIA dataset are provided.

Optimisation Method	Hyperparameter Space Investigated				Hyperparameter Combinations Modelled	Optimal Balanced Accuracy
	Tree Depth	Step	Number of Models	Step		
Grid Search	5 - 50	1	25 - 500	5	5000	0.837
		5		25	200	0.815
Bayesian		-		-	50	0.806

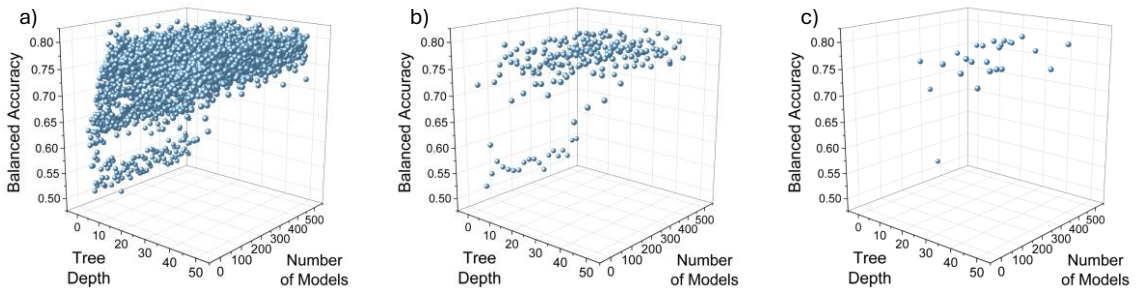


Figure 2.21. Plots showing the balanced accuracy achieved following 5-fold cross-validation for each investigated combination of hyperparameters, as described in Table 2.2. (a) Shows the grid search with smaller steps, (b) shows the grid search with larger steps, and (c) shows the Bayesian optimisation results.

While this early investigation indicated that Bayesian optimisation may be a suitable method for efficiently optimising hyperparameters, it was incompatible with the classification problems discussed in Chapters 3 and 4. Use of the method failed to yield optimised parameters, necessitating the use of the more costly grid search. Where Bayesian optimisation is problematic, future models may benefit from applying random search strategies. While they were not utilised in this project, randomly searching the hyperparameter space is more efficient, both theoretically and empirically, than grid-search strategies.[94]

2.5.7 Cross Validation

To optimise the parameters used in each model, the performance of each parameter selection needs to be evaluated. Modelling and testing performance using the same data from the training set is not acceptable; using the test set at this stage would make it no longer independent of the modelling procedure and inadequate for external model validation.[10] While an additional parameter-testing partition could be used, each partition of the training set reduces the information available for model building. To address this issue, cross-validation on the training set can be performed.

There are many cross-validation techniques; one of the most popular choices is k-fold cross-validation. In k-fold cross-validation, the training data is randomly partitioned into k sets. The model is built using $k - 1$ of these sets, and the held-out set is predicted by this model to estimate performance. The subset is returned to the training data, and each partition is iteratively held out and tested against models built by their respective $k - 1$ subsets (Figure 2.22). As k increases, the estimated performance values are expected to reflect the true performance values more closely. In the extreme, the number of folds can equal the number of samples, a method known as leave-one-out cross-validation (LOOCV). While LOOCV is also popular, as k increases, so too does the computational power required, making it less

practical. 5-fold and 10-fold cross-validation are common alternatives to LOOCV, with *k*-fold cross-validation showing comparable performance and being much more efficient.[67] While internal assessment allows us to estimate the model's performance, the true predictive power of a model can only be assessed by applying it to previously unseen data.

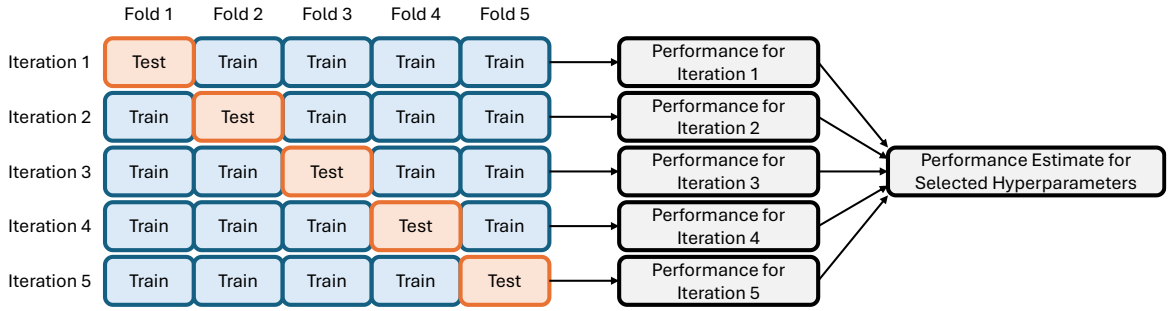


Figure 2.22. Cartoon illustrating 5-fold cross-validation for hyperparameter optimisation. The training set is split into 5 folds. Each fold is used iteratively to evaluate a model built on the remaining test data with the selected hyperparameters. The performance from each iteration is used to estimate the model performance for the given combination of hyperparameters.

2.6 Model Validation

Another key requirement set out by the OECD is the need for appropriate metrics to assess the model's performance.[55] The metrics used must be independent of the class distribution. In situations where a model could achieve high overall accuracy simply by favouring the dominant class, it is necessary to use metrics that ensure performance on each class is equally weighted, making it fairer and more informative when classes are skewed.

2.6.1 Performance Metrics

Balanced accuracy is a measure of the proportion of correctly classified molecules in each class (c = 2 for binary classification) based on values of a confusion matrix (TP, true positive; TN, true negative; FP, false positive; and FN, false negative) (Equation 2.10). It can range from 0 to 1, with an ideal model having a value of 1 and an acceptable value> 0.7.[3]

$$\text{Balanced Accuracy} = \frac{1}{c}\sum_{i=1}^{c}\frac{TP_i}{TP_i + TN_i} \qquad \text{(Equation 2.10)}$$

Cohen's kappa (κ) is a statistical measure that assesses the agreement between two raters, in this case, measured classifications compared to their predicted classifications. It can vary between -1 and 1, with an ideal model having a value of 1 and an acceptable value being > 0.4 (Equation 2.11).[3]

$$\kappa = \frac{2 \cdot (\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN})}{(TP + FP) \cdot (\mathrm{FP} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{FN} + \mathrm{TN})} \qquad \text{(Equation 2.11)}$$

In addition, the receiver operating characteristic (ROC) curve can be calculated by plotting the true positive rate against the false positive rate for every possible classification threshold (The cutoff probability used to decide whether a prediction should be classified as one class or another). It serves as an easy-to-visualise comparison of model performance, quantified by the area under the curve (AUC).

While early iterations of our model used Accuracy or Cohen's kappa as the target metric for cross-validation, we chose Matthew's correlation coefficient (MCC) as our metric in later models discussed in Chapters 3 and 4, as it has recently been shown that MCC should replace other popular assessment metrics like Cohen's kappa and ROC AUC as the standard metric for assessing binary classification.[96, 97]

The MCC considers all values of the binary classification confusion matrix to provide a balanced and reliable assessment, especially for imbalanced datasets as defined by (Equation 2.12. An ideal model has an MCC value of 1; a random model has a value of 0, and an acceptable value is> 0.5.

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(TP + FP) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}} \qquad \text{(Equation 2.12)}$$

While there are many other metrics for assessing the performance of classification models, many are only more effective when misclassification errors are disregarded. Given that misclassifications provided by chemical prediction tools may misinform chemists, leading to costly, time-consuming, and ultimately futile investigations, using metrics that account for classification errors is essential. For such problems, it has been shown that MCC is the best choice.[98]

2.6.2 Model Robustness

While cross-validation is effective at avoiding the overfitting of hyperparameters to maximise model performance, the relationship may be due to chance correlations in the data. If meaningful insights are to be extracted from the model, it is necessary to ensure that a true relationship between the chemical structures and properties has been identified. As the

number of descriptors considered for model building (M) and the number used (m) increases, the number of possible models grows quickly (Equation 2.13).[99]

$$^{M}C_{m} = \frac{M!}{m!\,(M-m)!} \qquad \text{(Equation 2.13)}$$

For example, if a model is built using 10 descriptors from a selection of 50 possible descriptors, there are more than 1×10^{10} possible models to investigate. With so many possible combinations, there are likely to be a few that fit the data well by chance. Subsequently, an additional metric is necessary to assess the likelihood that a chance correlation will be identified.

Y-randomisation is a popular approach for assessing the risk of chance correlation models. In contrast to cross-validation, it does not estimate the prediction performance, but instead the reliability of those predictions.[90] To investigate the model's robustness, the response variables, i.e. the property classifications, are randomly scrambled, and the effects on predictive performance are assessed. Since a particular permutation of the scrambled y-values may be similar to the unscrambled data, a few of the runs may result in a reasonable performance. Such permutations can make it difficult to assess the results of a y-randomisation experiment.[99] The Z-score is a metric that can be used to assess the statistical significance of the true model's performance relative to randomised models and dataset metrics. It is expected that, when building a model using randomised data or applying randomised test sets to the true model, predictive performance will decrease.

The chosen optimisation metric values (differing across chapters as discussed in Section 2.6.1) for test and in-domain test-set predictions were compared with those derived for random datasets on the optimised model, generated by random shuffling of molecule classifications. The statistical significance of QSAR models for each set was evaluated by determining the mean value of the chosen metric (μ) and its standard deviation (σ) for 100 random datasets applied to the optimised models. The Z-score is calculated by comparing the randomised average to the MCC value for the actual dataset (h) (Equation 2.14). The Z-score can then be compared with the critical values of Z_c at different levels of significance (α) (Equation 2.15, Table 2.3).[100] The robustness of the models discussed in Chapters 3 and 4 was further investigated. MCC values for the test and in-domain test set predictions were compared to those calculated by performing the same cross-validation hyperparameter optimisation on a randomised training set and making predictions with the new scrambled

models. The statistical significance of QSAR models for each set was evaluated by determining the mean value of MCC (μ) and its standard deviation (σ) for 10 random datasets applied to the optimised randomised models. The Z-score is calculated by comparing the randomised average to the MCC value for the actual dataset (*h*) using (Equation 2.14).

$$Z = \frac{h - \mu}{\sigma} \qquad \text{(Equation 2.14)}$$

$$\alpha = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}, z \leq 4 \qquad \text{(Equation 2.15)}$$

Table 2.3 Critical Z-scores (Zc) and their corresponding levels of significance (α), with a Z-score greater than 1.64 indicating a 95% confidence that a the resulting models' performance is not a result of chance correlations.

Z_c	α
1.28	0.10
1.64	0.05
2.33	0.01
3.10	0.001

Furthermore, for the models discussed in Chapters 3 and 4, the significance of the optimised models' predictive performance was assessed. A one-tailed one-sample t-test was performed to compare the MCC values for the optimised models with those obtained from cross-validation hyperparameter modelling on randomised training sets.

2.6.3 External Validation

While data partitioning provides a means of interrogating model performance on data independent of the training data, the final model configuration chosen will depend on the composition of the test set. For example, the appropriate domain of applicability is chosen based on the performance of the test set. While the test set metrics may indicate that the model performs well on unseen data, it is important to note that the test set is a stratified partition of the entire dataset. While a stratified partition affords a test set of data with class proportions representative of the overall data, it may be biased in other ways that are unclear to the modeller. Therefore, in Chapter 3, we discuss synthesising molecules with unknown properties to further support the model's predictions and highlight where additional training is required to improve the model's efficacy. In Chapter 4, we highlight the necessity of a

separate validation partition because the techniques required to characterise the property of interest are inaccessible.

2.7 Model Visualisation

The final OECD principle of effective QSPR development is to include a mechanistic interpretation where possible. While a QSPR framework may be designed and reported unambiguously, by their very nature, some machine learning methods are inherently uninterpretable. As models become increasingly complex, they become more difficult to probe. The development of a black box prediction system makes the use of QSPR to probe design principles difficult. Furthermore, without a rational relationship between the structural features of the molecules and the property, confidence in the model's ability will be limited. Therefore, it is necessary to incorporate methods to extract feature importance, which can be used to interrogate the structure-property relationship.[101, 102]

2.7.1 Descriptor Importance

There are many ways to probe the features underpinning machine learning models. Some, such as the information gain metric provided by the XGBoost model, provide a way to assess the contribution of each descriptor to the model, though relying on a single model may introduce significant bias in interpretation if the model is particularly biased toward including a certain substructural descriptor. Instead, previous models have used many shallow decision trees to generate importance metrics.[103] By quantifying how often each descriptor is selected as the optimal variable to partition the data into its appropriate classes, each descriptor's importance can be determined. If a descriptor is often selected as the best split across several trees, it is likely contributing more to the model than those that are not. This can be quantified as the ratio of the frequency of being selected as the optimum partition criterion to the frequency of being randomly selected as a candidate at each level of the decision tree. (Equation 2.16).

$$Importance = \frac{\#\ Split_0}{\#\ Candidate_0} + \frac{\#\ Split_1}{\#\ Candidate_1} + \frac{\#\ Split_2}{\#\ Candidate_2} \quad \text{(Equation 2.16)}$$

A significant limitation of such metrics is their inability to elucidate which class is favoured by the presence of a given descriptor. If QSPRs are to be utilised to realise the rational design and tailoring of properties of interest, it is essential that such knowledge can be extracted

from the model. Subsequently, a more informative metric of feature importance is essential to derive a model's mechanistic interpretation.

Shapley values are a game-theoretic technique for quantifying fair resource allocation based on a player's contribution to a cooperative game. The concept has been applied across fields, including in the development of explainable artificial intelligence. Lloyd Shalpey originally introduced these values, and it was later shown that they uniquely provide a range of properties that are incredibly useful for the problem of feature importance.[102, 104, 105] In terms of descriptor importance, the Shapley value formula describes the contribution of each descriptor averaged across all possible combinations of descriptors that could be used to make a prediction. The contribution of each of the descriptors (N) to a model can be determined by considering how adding an individual descriptor (i) influences the prediction ($v\ (S)$) when added to a subset of the total descriptors (S) used to make a prediction. The Shapley value is given by the average weighted sum of each of these marginal contributions of each descriptor when added to a coalition of other descriptors $[v\ (S \cup \{i\} - v\ (S)]$ (Equation 2.17). This yields a fair, permutation-averaged estimate of each descriptor's importance.

$$\phi_i(v) = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [v\ (S \cup \{i\}) - v(S)] \qquad \text{(Equation 2.17)}$$

Shapley values are a unique solution to the problem of value allocation, as the only method to satisfy the axioms of efficiency, symmetry, the null player, and additivity. These axioms state that the solution to the problem must ensure that the total value of the prediction is distributed among the descriptors, descriptors with equal contributions receive equal values, descriptors do not contribute and receive zero value, and distribute importance such that the linear combination of each descriptor results in the linear combination of their resulting importance.[105] The importance of a descriptor can be easily visualised if there is a robust relationship between the descriptor's value and the resulting prediction. Where the increasing value of the descriptor results in a prediction favouring the positive class, there will be a similar increase in the Shapley value for each molecule (and *vice versa*). This allows the features that promote or diminish the property of interest to be identified. Such relationships can be easily visualised in a beeswarm plot (Figure 2.23).

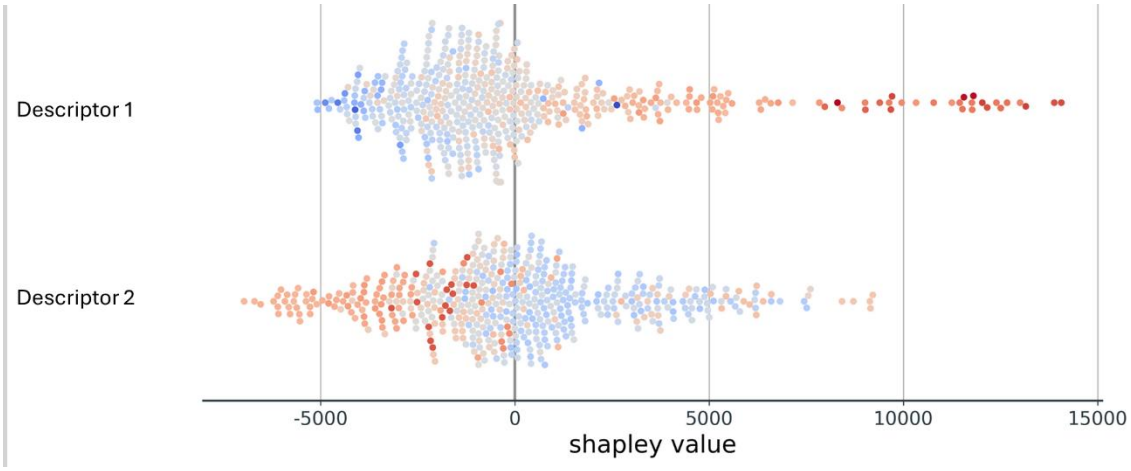


Figure 2.23. Beeswarm plot visualisation of Shapley values. Data colour representative of the descriptor value for each given molecule. Blue represents a low descriptor value, while red represents a greater descriptor value. The upper plot shows that, as the value of descriptor 1 increases, the Shapley value typically increases, indicating a positive relationship between descriptor 1 and its predicted positive class. The opposite relationship is shown in the lower plot for descriptor 2. Adapted from [106]

2.7.2 Substructure Identification

Quantifying the descriptor's importance is key to interpreting how the machine learning model is building a relationship between the descriptor and the properties of interest. However, simply identifying the abstract fingerprint bits which contribute to the model does not provide useful information for instructing future chemical design. Instead, the chemical substructure represented by each fingerprint bit of interest must be reverse-engineered to elucidate the chemical substructures that influence the desired property. Given that the Morgan fingerprints represent the mapping of infinite chemical space into a finite bit integer, the significant loss of information makes it impossible to directly extract the structure represented by each bit due to bit overlap (two independent substructures represented by the same fingerprint bit). To overcome this, we utilised a KNIME workflow previously prepared by Greg Landrum, which uses RDKit to track which atoms and their circular radii are used when generating a given molecular fingerprint bit from the input data to elucidate the structure.[107] It has previously been shown that as the size of the fingerprint is reduced, and the radii they represent increases, the number of bit collusions increases noticeably. Our choice of a 1024 bit fingerprint with a radii of 2 used across the project, while allowing for many more collisions that bigger fingerprints and smaller radii, the frequency of collision occuring is relatively small (Figure 2.24). On smaller datasets such as those utilised here, the likelihood of problematic quantities of overlap is further reduced. However, future projects may obtain more chemical insight by using larger radii fingerprints. To overcome the increase bit collision this would afford, larger fingerprints could be utilised.[108]

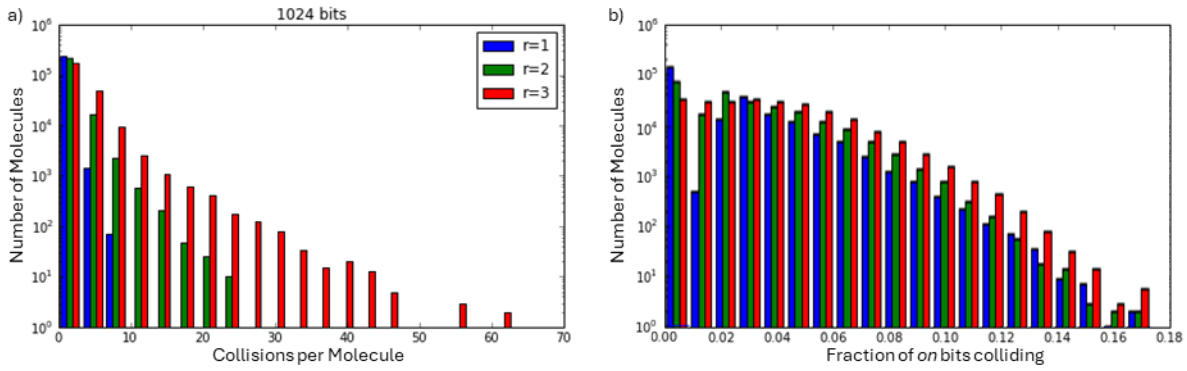


Figure 2.24. Histograms of proportion of bit collisions in 1024 bit Morgan fingerprints in a library of 230,000 molecule. a) shows the number of molecules with a given number of bit collisions. b) shows the number of molecules with a given proportion of *on* bits colliding. Adapted from reference [108].

In the case of the PLA2-IIA benchmark: by calculating descriptor importance using the ensemble of shallow trees, it was found that bits 140, 163, 416, 531, and 946 were the most influential to the decision tree ensemble model. Therefore, identifying the substructures related to these descriptors could help elucidate important structural features for the rational design of new PLA2-IIA inhibitors (Figure 2.25).

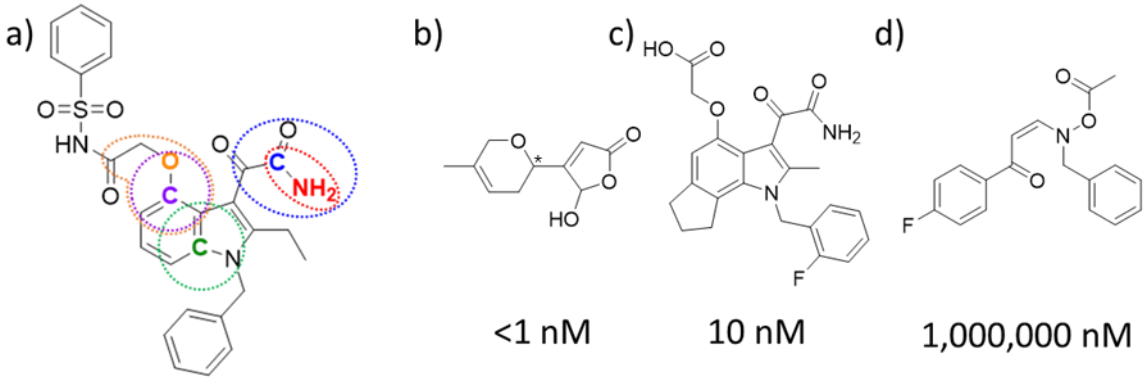


Figure 2.25. (a) Example molecule highlighting the substructures represented by bits 140, 163, 416, 531, and 946 (red, orange, green, blue, purple; respectively); with the central atom highlighted, and the extended connectivity of the substructure shown with a dashed line. (b-d) PLA2-IIA inhibitor candidates and their IC_{50} concentrations.

Many of the highly active molecules in the PLA2-IIA dataset have the indole, aryl ether, and oxoacetamide substructures represented by these important bits, indicating that the presence of such structures are key features in PLA2-IIA inhibition. However, while many of the structures lacking these key features showed an expected increase in IC_{50} concentration, a few exceptions, such as the example in Figure 2.25b, show significant activity. This indicates that the model has been biased towards the highlighted structures, with other areas of chemical space poorly represented. However, this result provides useful insight into the

focus of future iterations of PLA2-IIA datasets, highlighting areas of the dataset which should be selectively expanded to improve future predictive systems.

2.8 PLA2-IIA Model Performance

Before applying the QSAR framework to our material chemistry problems of interest, we benchmarked early iterations of our workflow using IC_{50} values for inhibition of phospholipase A2 group IIA (PLA2-IIA) (ChEMBL ID: CHEMBL3474).[19] This allowed us to assess the system's basic functionality on a more idealised dataset. The PLA2-IIA was an effective choice due to the reasonable dataset size and acceptable MODI, with 504 molecules yielding a MODI of 0.80. These indicated that the dataset should be amenable to modelling. This was supported by the work of Christopher Woodley, whose analogous statistics provide a useful comparison for the efficacy of our framework (Table 2.4).[24] Furthermore, the dataset exhibited data imbalance at the IC_{50} classification threshold of 100 nM, yielding an approximate 4:1 inactive:active molecule split. Showing that the model can handle data balance is necessary if the framework is to be applied to other chemical problems, which typically suffer from such an imbalance.

To prepare the data, the molecules were cleaned of erroneous structures and duplicate entries, reducing the 616 molecule dataset to 504 acceptable structures. Using the provided IC_{50} data, the classes were assigned, and the Morgan fingerprint was calculated from the structure. A 20% test set subsample was partitioned, stratified by assigned class to maintain the dataset's distribution. The fingerprint bits were filtered for low variance and high correlation, retaining 87 high-information descriptors. The dataset balance was treated in four ways: unaltered, maintaining the original dataset distribution; randomly downsampled, such that the training set achieved a 2:1 majority:minority class distribution, upsampled using SMOTE to achieve a 1:1 majority:minority class distribution; and a combination method, randomly downsampling to achieve a 2:1 majority:minority class distribution, followed by SMOTE to upsample to a 1:1 class distribution. Each of these datasets was used to train a random forest, a decision tree, support vector machines, and a k-nearest neighbour model. Hyperparameters were optimised using a grid search, and performance was internally assessed using a 5-fold cross-validation approach. The hyperparameters that yielded the highest accuracy after cross-validation were used to build the final models. The ROC AUC and Cohen’s kappa metrics were calculated for the training and test sets applied to each optimised model (Table 2.5). A y-randomisation robustness test was also performed by randomising the test set and applying the shuffled data to the optimised models built using the SMOTE balanced dataset

(Table 2.6). While the training set performance upon randomisation was used to generate a like-for-like robustness comparison with the work by Woodley, y-randomisation is typically performed using cross-validation performance metrics of the training set, or comparison of the test set against the randomised models is better used to assess the robustness.[109, 110] Therefore, the models in Chapters 3 and 4 utilise the change in test set performance upon randomisation to assess the model robustness.

Table 2.4 Model statistics for PLA2-IIA IC_{50} QSAR built in Python by Woodley.[24]

Balancing Procedure	Model	Training			Test	
		ROC AUC	κ	Z-score	ROC AUC	κ
Downsampled	Random Forest	0.766	0.533	-	0.924	0.788
	Support Vector Machines	0.842	0.685	5.86	0.935	0.816
	K Nearest Neighbour	0.750	0.500	-	0.902	0.755

Table 2.5 Model statistics across various sample class balancing procedures

Model	Balancing Procedure	Training Set		Test Set	
		ROC AUC	κ	ROC AUC	κ
Random Forest	Unbalanced	0.995	0.882	0.869	0.581
	Downsampled	0.995	0.875	0.898	0.581
	SMOTE	0.995	0.889	0.897	0.581
	Combination	0.993	0.878	0.869	0.508
Decision Tree Ensemble	Unbalanced	1.000	0.978	0.918	0.629
	Downsampled	1.000	0.978	0.928	0.665
	SMOTE	1.000	0.978	0.924	0.665
	Combination	0.996	0.970	0.843	0.322
Support Vector Machines	Unbalanced	0.998	0.978	0.853	0.116
	Downsampled	1.000	0.978	0.853	0.116
	SMOTE	0.995	0.978	0.853	0.116
	Combination	0.993	0.963	0.765	0.331
k Nearest Neighbour	Unbalanced	0.999	0.944	0.829	0.552
	Downsampled	0.999	0.964	0.813	0.462
	SMOTE	1.000	0.978	0.826	0.604
	Combination	0.999	0.970	0.744	0.418

Table 2.6 Model robustness statistics on training set across models using the SMOTE data balancing procedure.

Model	Z-score	α
Random Forest	9.48	$< 3 \times 10^{-15}$
Decision Tree Ensemble	1.85	0.032
Support Vector Machines	1.28	0.100
k Nearest Neighbour	5.37	$< 3 \times 10^{-5}$

The resulting models indicate that well-performing predictive models can be developed using KNIME, as evidenced by the PLA2-IIA models' performance metrics. Compared with Christopher Woodley's models, our models perform better on the training set but worse on the test set. This suggests that hyperparameter optimisation can be further improved to reduce overfitting to the training set.[24] As a result of overfitting, the models developed in this work describe relationships across the chemical space with poorer generalisability. This may, in part, be a result of the use of a single 1D chemical descriptor, which alone may not be capable of capturing the chemical information responsible for the inhibitory effect. The test metrics exceeding the performance of the training metrics by Woodley are unusual, as external test statistics are limited based on the accuracy of the dataset.[111] Given that, this suggests it's likely the test metrics have been overstated; whether by some chance correlation during partitioning, or erroneous reporting. With incomplete information, the cause cannot be determined but highlights the need for complete transparency in the reporting of QSPR models.

While the PLA2-IIA models appeared relatively robust to changes in the data balance, SMOTE was comparable to, or better than, the unbalanced and down-sampling approaches in each case. While it has been shown previously that combination balancing methods can outperform individual approaches, this balancing approach performed poorly, improving only the support vector machine model (though Cohen's kappa on the test set remains below the acceptable threshold).[112] Furthermore, while the overall accuracy of the model has been used as the parameter optimisation target, the severe class imbalance often found in chemical classification problems could be counteracted by instead optimising for the accuracy of the minority class in future models by selecting metrics such as MCC as the optimisation target. Some decision tree hyperparameters cannot be tuned within KNIME's RF learner node, such as the proportion of descriptors randomly sampled at each split or the fraction of data available for each tree to learn from. Subsequently, the less restrictive decision tree ensemble method was also investigated to tune the training process further. While the decision tree

ensemble has better performance on test sets, the model robustness Z-score indicates that the probability that chance correlation caused the performance increase is increased by several orders of magnitude. This is to be expected; however, as increasing the number of available hyperparameters makes it much more likely that random relationships in the data can be found. This is undesirable, as determining mechanistic interpretations from models with a high chance of being derived from chance correlations make it less likely to elucidate truly meaningful factors contributing to the desired properties.

The feature importance metric provided a mechanistic interpretation of the resulting predictions, suggesting that the presence of the indole, aryl ether, and oxoacetamide substructures is key to the development of PLA2-IIA inhibitors. The interpretation was rational for the dataset, though it indicated that the model did not explain some structures well (Figure 2.26).

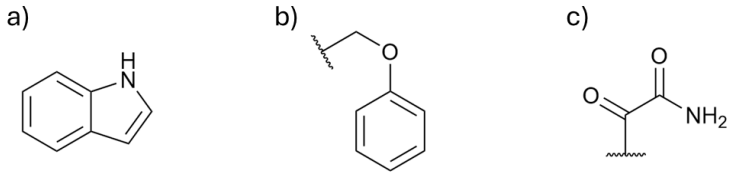


Figure 2.26. Chemical structures of the a) indole, b) aryl ether, and c) oxoacetamide substructures.

Together, these metrics show that the framework built in KNIME could establish effective QSAR relationships compared with those developed previously.[113] However, future models would require further adjustments and additional functionality to gain useful insights into our materials of interest. This includes the inclusion of additional machine learning models to make a more robust consensus based system, the incorporation of applicability domain metrics to screen out test set molecules that are expected to be poorly modelled due to their dissimilarity to the training set, the application of performance metrics robust to dataset imbalance and quantify characterisation errors, and finally more powerful model visualisation techniques that can identify how different descriptors influence the prediction afforded by each specific model.

2.9 Conclusion

In this chapter, we have introduced the concepts central to building our QSAR/QSPR framework and have shown how it satisfies the OECD guidelines for effective QSAR development. This has allowed us to develop models that have:

- A defined endpoint – aided by the inclusion of automated classification procedures.
- An unambiguous algorithm – afforded by the open-source KNIME platform, which is accessible to chemists without costly barriers of entry, and an interpretable and interrogatable visual programming interface.
- A defined domain of applicability – achieved using distance-based metrics and selectively chosen based on optimised test metrics to ensure optimised boundaries for the given dataset
- Appropriate assessment metrics – use of metrics robust to class imbalance for both internal and external validation. Predictivity and robustness can both be quantified to ensure that true and effective relationships are identified
- A mechanistic interpretation – feature importance metrics, along with substructure elucidation, give insight into the features governing the model's prediction, allowing mechanistic insights to be obtained.

The framework was built in KNIME and benchmarked using an imbalanced dataset of PLA2-IIA IC_{50} values. While the comparison of the models built in this work with those prepared previously indicates that further optimisation is required to reduce the impact of training-set overfitting, the performance metrics show that effective predictive relationships can be developed using our framework. This work lays the foundation for the development of the predictive models discussed in Chapters 3 and 4.

2.10 References

1. K. B. Lipkowitz, T. R. Cundari and ProQuest, *Reviews in computational chemistry*, Wiley, Hoboken, NJ, 2007.
2. A. van Teijlingen and T. Tuttle, *Journal of Chemical Theory and Computation*, 2021, **17**, 3221–3232.
3. J. K. Gupta, D. J. Adams and N. G. Berry, *Chemical Science*, 2016, **7**, 4713–4719.
4. Y. Liu, *ChemRxiv*, 2025, **2025**.
5. T. Le, V. C. Epa, F. R. Burden and D. A. Winkler, *Chemical Reviews*, 2012, **112**, 2889–2919.
6. T. K. Piskorz, L. Perez-Chirinos, B. Qiao and I. R. Sasselli, *ACS Omega*, 2024, **9**, 31254–31273.
7. P. Xu, X. Ji, M. Li and W. Lu, *npj Computational Materials*, 2023, **9**, 42.
8. OECD, *OECD Series on Testing and Assessment*, 2023, DOI: 10.1787/d96118f6-en.
9. A. Gissi, O. Tcheremenskaia, C. Bossa, C. L. Battistelli and P. Browne, *Computational Toxicology*, 2024, **31**, 100326.
10. A. Tropsha, *Molecular Informatics*, 2010, **29**, 476–488.
11. B. Khalid, H. Ghorab and A. Benkhemissa, *Journal of the Indian Chemical Society*, 2022, **99**, 100672.
12. D. Fourches, E. Muratov and A. Tropsha, *J Chem Inf Model*, 2010, **50**, 1189–1204.
13. D. Young, T. Martin, R. Venkatapathy and P. Harten, *QSAR & Combinatorial Science*, 2008, **27**, 1337–1345.
14. J. C. Dearden, M. T. D. Cronin and K. L. E. Kaiser, *SAR and QSAR in Environmental Research*, 2009, **20**, 241–266.
15. A. Rácz, D. Bajusz and K. Héberger, *Molecules*, 2021, **26**, 1111.
16. M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, *Chemical Society Reviews*, 2025, **54**, 1125–1150.
17. M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal and A. Valencia, *Chemical Reviews*, 2017, **117**, 7673–7761.
18. S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, Benjamin A. Shoemaker, Paul A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and Evan E. Bolton, *Nucleic Acids Research*, 2024, **53**, D1516–D1525.
19. A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit and A. R. Leach, *Nucleic Acids Research*, 2016, **45**, D945–D954.
20. G. Nicola, T. Liu and M. K. Gilson, *Journal of Medicinal Chemistry*, 2012, **55**, 6987–7002.
21. N. Bosc, F. Atkinson, E. Felix, A. Gaulton, A. Hersey and A. R. Leach, *Journal of Cheminformatics*, 2019, **11**, 4.
22. C. Nantasenamat, A. Worachartcheewan, S. Jamsak, L. Preeyanon, W. Shoombuatong, S. Simeon, P. Mandi, C. Isarankura-Na-Ayudhya and V. Prachayasittikul, in *Artificial Neural Networks*, ed. H. Cartwright, Springer New York, New York, NY, 2015, DOI: 10.1007/978-1-4939-2239-0_8, pp. 119–147.
23. O. A. Tarasova, A. F. Urusova, D. A. Filimonov, M. C. Nicklaus, A. V. Zakharov and V. V. Poroikov, *J Chem Inf Model*, 2015, **55**, 1388–1399.
24. C. Woodley, unpublished work.
25. G. A. Landrum and S. Riniker, *Journal of Chemical Information and Modeling*, 2024, **64**, 1560–1567.
26. G. v. Rossum and F. L. Drake, *The Python Language Reference Manual*, Network Theory Ltd., 2011.

27. S. U. Patil and S. U. Patil, *International Journal For Multidisciplinary Research. https://doi. org/10.36948/ijfmr*, 2024, v06i03.
28. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *the Journal of machine Learning research*, 2011, **12**, 2825–2830.
29. S. Nanz and C. A. Furia, 2015.
30. D. Sitaram and G. Manjunath, in *Moving To The Cloud*, eds. D. Sitaram and G. Manjunath, Syngress, Boston, 2012, DOI: https://doi.org/10.1016/B978-1-59749-725-1.00010-X, pp. 389–425.
31. *Journal*, 2024.
32. J. K. Gupta, D. J. Adams and N. G. Berry, *Chem Sci*, 2016, **7**, 4713–4719.
33. W. A. Warr, *J Comput Aided Mol Des*, 2012, **26**, 801–804.
34. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel and B. Wiswedel, Berlin, Heidelberg, 2008.
35. K. Mansouri, J. T. Moreira-Filho, C. N. Lowe, N. Charest, T. Martin, V. Tkachenko, R. Judson, M. Conway, N. C. Kleinstreuer and A. J. Williams, *J Cheminform*, 2024, **16**, 19.
36. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland and J. Laufer, *Journal of Chemical Information and Computer Sciences*, 1992, **32**, 244–255.
37. G. Landrum, RDKit: Open-source cheminformatics, https://www.rdkit.org, DOI: 10.5281/zenodo.7828379).
38. G. Landrum, P. Tosco, B. Kelley, R. Rodriguez, D. Cosgrove, Sriniker, R. Vianello, Gedeck, N. Schneider, G. Jones, E. Kawashima, D. Nealschneider, A. Dalke, B. Cole, M. Swain, S. Turk, A. Savelev, A. Vaucher and M. Wójcikowski, RDKit Release_2024.03.1, DOI: https://doi.org/10.5281/zenodo.10893044).
39. D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31–36.
40. P. Ambure and M. N. D. S. Cordeiro, in *Ecotoxicological QSARs*, ed. K. Roy, Springer US, New York, NY, 2020, DOI: 10.1007/978-1-0716-0150-1_5, pp. 97–109.
41. A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, Baskin, II, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard and A. Tropsha, *J Med Chem*, 2014, **57**, 4977–5010.
42. R. Batista, A. De Jesus Silva Júnior and A. B. De Oliveira, *Molecules*, 2009, **14**, 3037–3072.
43. F. Grisoni, D. Ballabio, R. Todeschini and V. Consonni, in *Computational Toxicology: Methods and Protocols*, ed. O. Nicolotti, Springer New York, New York, NY, 2018, DOI: 10.1007/978-1-4939-7899-1_1, pp. 3–53.
44. R. Todeschini, V. Consonni and P. Gramatica, in *Comprehensive Chemometrics*, eds. S. D. Brown, R. Tauler and B. Walczak, Elsevier, Oxford, 2009, DOI: https://doi.org/10.1016/B978-044452701-1.00007-7, pp. 129–172.
45. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Advanced Drug Delivery Reviews*, 2001, **46**, 3–26.
46. A. Capecchi, D. Probst and J.-L. Reymond, *Journal of Cheminformatics*, 2020, **12**, 43.
47. H. L. Morgan, *Journal of Chemical Documentation*, 1965, **5**, 107–113.
48. D. Rogers and M. Hahn, *Journal of Chemical Information and Modeling*, 2010, **50**, 742–754.

49. A Practical Introduction to the Use of Molecular Fingerprints in Drug Discovery, https://towardsdatascience.com/a-practical-introduction-to-the-use-of-molecular-fingerprints-in-drug-discovery-7f15021be2b1/, (accessed 19th April 2024).
50. R. E. Carhart, D. H. Smith and R. Venkataraghavan, *Journal of Chemical Information and Computer Sciences*, 1985, **25**, 64–73.
51. A. Golbraikh, E. Muratov, D. Fourches and A. Tropsha, *Journal of Chemical Information and Modeling*, 2014, **54**, 1–4.
52. D. Stumpfe, Y. Hu, D. Dimova and J. Bajorath, *Journal of Medicinal Chemistry*, 2014, **57**, 18–28.
53. F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni and R. Todeschini, *Molecules*, 2012, **17**, 4791–4810.
54. C. C. Aggarwal, A. Hinneburg and D. A. Keim, Berlin, Heidelberg, 2001.
55. OECD principles for the Validation, for Regulatory Purpose, of (Q)SAR Models, https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm, (accessed 02 March 2023).
56. D. Bajusz, A. Rácz and K. Héberger, *Journal of Cheminformatics*, 2015, **7**, 20.
57. C. L. Mellor, R. L. Marchese Robinson, R. Benigni, D. Ebbrell, S. J. Enoch, J. W. Firman, J. C. Madden, G. Pawar, C. Yang and M. T. D. Cronin, *Regulatory Toxicology and Pharmacology*, 2019, **101**, 121–134.
58. F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni and R. Todeschini, *Molecules*, 2012, **17**, 4791–4810.
59. M. Sivakumar, S. Parthasarathy and T. Padmapriya, *PeerJ Comput Sci*, 2024, **10**, e2245.
60. R. Silipo, I. Adae, A. Hart and M. Berthold, Seven Techniques for Dimensionality Reduction, https://www.knime.com/sites/default/files/inline-images/knime_seventechniquesdatadimreduction.pdf).
61. C. Spearman, *International Journal of Epidemiology*, 2010, **39**, 1137–1150.
62. A. Rácz, D. Bajusz and K. Héberger, *Mol Inform*, 2019, **38**, e1800154.
63. M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos and E. Tuzhilina, *Nature Reviews Methods Primers*, 2022, **2**, 100.
64. A. Tharwat, *International Journal of Applied Pattern Recognition*, 2016, **3**, 197.
65. P. Liu and W. Long, *Int J Mol Sci*, 2009, **10**, 1978–1998.
66. D. M. Hawkins, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 1–12.
67. M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer New York, NY, 1 edn., 2013.
68. J. Jiang, C. Zhang, L. Ke, N. Hayes, Y. Zhu, H. Qiu, B. Zhang, T. Zhou and G. W. Wei, *Chem Sci*, 2025, **16**, 7637–7658.
69. X. Gao, D. Xie, Y. Zhang, Z. Wang, C. Chen, C. He, H. Yin and W. Zhang, *Frontiers of Computer Science*, 2026, **20**, 2011622.
70. A. d. P. Rodrigues, A. S. Luna and L. Pinto, *Chemometrics and Intelligent Laboratory Systems*, 2023, **240**, 104933.
71. I. Mani and I. Zhang, 2003.
72. I. Tomek, 1976.
73. N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, *J. Artif. Intell. Res. (JAIR)*, 2002, **16**, 321–357.
74. D. Li, J. Liu and J. Liu, *Macromolecular Theory and Simulations*, 2021, **30**, 2100010.
75. G. LemaÃŽtre, F. Nogueira and C. K. Aridas, *Journal of machine learning research*, 2017, **18**, 1–5.
76. C. MacDonald, ADASYN object attribute error, https://forum.knime.com/t/adasyn-object-attribute-error/63238, (accessed 1st March 2026).

77. A. Morales-Hernández, I. Van Nieuwenhuyse and S. Rojas Gonzalez, *Artificial Intelligence Review*, 2023, **56**, 8043–8093.
78. J. Wang, What are KNN's advantages and disadvantages compared to the other popular machine learning models?, https://medium.com/@jing.wang.ds/what-are-knns-advantages-and-disadvantages-compared-to-the-other-popular-machine-learning-models-b171c48dd817, (accessed 23rd February 2026).
79. P. J. B. Pajila, B. G. Sheena, A. Gayathri, J. Aswini, M. Nalini and R. S. S, 2023.
80. T. Cover and P. Hart, *IEEE Transactions on Information Theory*, 1967, **13**, 21–27.
81. S. L. Brunton, J. N. Kutz and P. Cambridge University, *Data-driven science and engineering: machine learning, dynamical systems, and control*, Cambridge University Press, Cambridge, United Kingdom, 2019.
82. L. Breiman, *Classification and regression trees*, Wadsworth International Group, Belmont, Calif, 1984.
83. L. Breiman, *Machine Learning*, 2001, **45**, 5–32.
84. J. H. Friedman, *The Annals of Statistics*, 2001, **29**, 1189–1232, 1144.
85. T. Bailey, A. Jackson, R.-A. Berbece, K. Wu, N. Hondow and E. Martin, *Journal of Chemical Information and Modeling*, 2023, **63**, 4545–4551.
86. M. Wiens, A. Verone-Boyle, N. Henscheid, J. T. Podichetty and J. Burton, *Clin Transl Sci*, 2025, **18**, e70172.
87. J.-P. Vert, K. Tsuda and B. Schölkopf, in *Kernel Methods in Computational Biology*, The MIT Press, 2004, DOI: 10.7551/mitpress/4057.003.0004, p. 0.
88. J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua and A. Lopez, *Neurocomputing*, 2020, **408**, 189–215.
89. I. Wickramasinghe and H. Kalutarage, *Soft Computing*, 2021, **25**, 2277–2293.
90. W. Sippl and D. Robaa, in *Applied Chemoinformatics*, 2018, DOI: https://doi.org/10.1002/9783527806539.ch2, pp. 9–52.
91. R. Raizada and Y. Lee, *PloS one*, 2013, **8**, e69566.
92. J. Lever, M. Krzywinski and N. Altman, *Nature Methods*, 2016, **13**, 541–542.
93. P. Ranganathan, C. S. Pramesh and R. Aggarwal, *Perspect Clin Res*, 2017, **8**, 148–151.
94. J. Bergstra and Y. Bengio, *J. Mach. Learn. Res.*, 2012, **13**, 281–305.
95. J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, *Advances in neural information processing systems*, 2011, **24**.
96. D. Chicco and G. Jurman, *BioData Mining*, 2023, **16**, 4.
97. D. Chicco, M. J. Warrens and G. Jurman, *IEEE Access*, 2021, **9**, 78368–78381.
98. A. Luque, A. Carrasco, A. Martín and A. de las Heras, *Pattern Recognition*, 2019, **91**, 216–231.
99. C. Rücker, G. Rücker and M. Meringer, *Journal of Chemical Information and Modeling*, 2007, **47**, 2345–2357.
100. M. Shen, A. LeTiran, Y. Xiao, A. Golbraikh, H. Kohn and A. Tropsha, *J Med Chem*, 2002, **45**, 2811–2823.
101. E. Königsberger, *Journal of Chemical & Engineering Data*, 2019, **64**, 381–385.
102. L. S. Shapley, 1953.
103. S. L. Brunton and J. N. Kutz, *Data-driven science and engineering: machine learning, dynamical systems, and control*, Cambridge University Press, Cambridge, United Kingdom, New York, NY, Second edn., 2022.
104. H. P. Young, *International Journal of Game Theory*, 1985, **14**, 65–72.
105. L. Qin, Y. Zhu, S. Liu, X. Zhang and Y. Zhao, *Mathematics*, 2025, **13**, 1581.
106. A. Cooper, How Shapley Values Work, https://www.aidancooper.co.uk/how-shapley-values-work/, (accessed 1st March 2026).
107. G. Landrum, Displaying information about fingerprint bits., https://hub.knime.com/greglandrum/spaces/Public/RDKit-

Examples/Highlighting%20Fingerprint%20Bits~sfrKZIZ6iR4E_av5/current-state, (accessed 2nd March 2026).
108. G. Landrum, Looking at bit collisions in Morgan fingerprints, https://rdkit.blogspot.com/2014/02/colliding-bits.html#:~:text=Looking%20at%20bit%20collisions%20in,of%20some%20machine%2Dlearning%20algorithms., (accessed 19th April 2026).
109. S. Zhang, L. Wei, K. Bastow, W. Zheng, A. Brossi, K.-H. Lee and A. Tropsha, *Journal of Computer-Aided Molecular Design*, 2007, **21**, 97–112.
110. H. Tang, X. S. Wang, X.-P. Huang, B. L. Roth, K. V. Butler, A. P. Kozikowski, M. Jung and A. Tropsha, *Journal of Chemical Information and Modeling*, 2009, **49**, 461–476.
111. S. S. Kolmar and C. M. Grulke, *J Cheminform*, 2021, **13**, 92.
112. A. Newaz and F. S. Haq, *arXiv preprint arXiv:2208.09619*, 2022.
113. C. Woodley, personal communication.

Chapter 3: Predictive classification of organic solubility of functionalised naphthalenes and rylenes

This Chapter is adapted, in part, from the following publication:

"An accessible property classification framework to predict the solubility of functionalised naphthalenes and rylenes in organic solvents"

C. R. M. MacDonald, Alex S. Loch, Neil G.Berry, and E. R. Draper

CRMM was responsible for conducting the research, creating figures, and writing the publication. CRMM wrote the initial draft of the manuscript, to which all authors contributed for the final publication.

3.1 Introduction

Smart materials are gaining increasing popularity due to their broader range of applications compared to conventional materials. Smart organic films are of particular interest due to their range of responsive properties.[1-5] Napthalene dimides (NDIs) and perylene bisimides (PBIs) have previously been utilised to prepare such functional materials due to their ability to self-assemble to form films with a wide range of physical, chemical, optical, and electronic properties (Figure 3.1).[6-8]

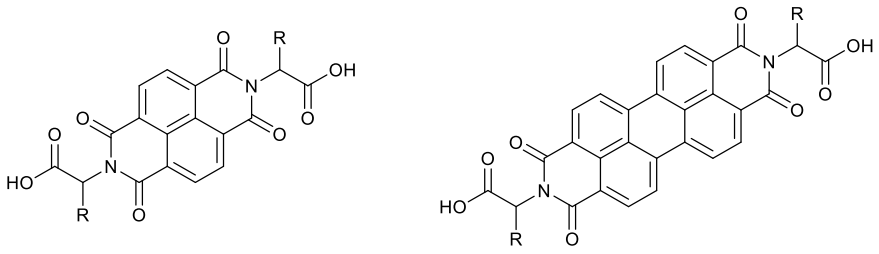


Figure 3.1. General structure of amino acid functionalised NDIs (left) and PBIs (right).

The properties of the self-assembled molecules are highly process-dependent, meaning a variety of materials can be prepared from a single molecule by tuning the bulk material preparation conditions.[9] The choice of solvent used to facilitate self-assembly can impact the molecular stacking. Yang *et al*. demonstrated that NDIs self-assemble into J-aggregated structures when precipitated from polar solvents, while H-aggregated structures are formed when precipitated from non-polar solvents.[10] The photoconductivity of self-assembled PBIs is dependent on the charge carrier mobility and exciton diffusion rate, both of which are superior in J-aggregated structures.[11] While π-π interactions facilitate the formation of primary aggregates, the amphiphilic nature of PBIs and NDIs enables these primary aggregates to self-assemble into secondary structures such as micelles, where supramolecular interactions cause the amphiphilic molecules to rearrange, utilising topology to sequester the non-polar core from the polar solvent (Chapter 1, Section 1.2.2). The morphology of micelle structures depends on the nature of the solvent and the solute concentration, amongst other factors. It has been shown that for other self-assembling molecules, there is a low critical micelle concentration (CMC), resulting in spherical micelles, and a higher CMC, corresponding to the formation of worm-like micelles that propagate in one dimension.[9] It is important, therefore, to be able to predict the solubility of NDIs and PBIs across a range of solvents to select candidate molecules and solvent systems better. For example, this would allow the prioritising of the formation of J-aggregates into one-dimensional micelles for optoelectronic applications, such as organic photovoltaics.

Enabling the prediction of solubility across solvents also provides key information which can enable the development of more complex tandem materials. The layer-by-layer assembly of multilayer films is a versatile method of material fabrication, enabling functionality that cannot be achieved with a single layer. These constructions benefit from simplicity, versatility, and precise control; however, intermixing between layers can diminish the film interfaces if the preceding layers cannot withstand erosion by the solvents used during subsequent depositions.[1] While films can be chemically altered after deposition to ensure orthogonality to subsequent solvents, changing the chemical structure may diminish the properties of interest. Instead, using chemically orthogonal solvents in each layer can ensure that the deposited material remains unaltered while minimising interfacial erosion (Figure 3.2).[1]

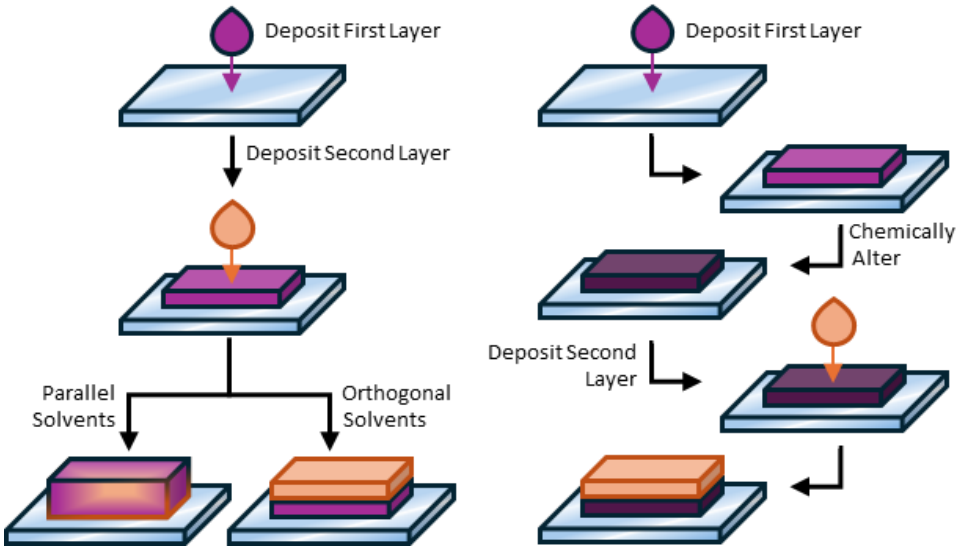


Figure 3.2. Cartoon depicting the fabrication of solution-processed multilayer films. When using parallel solvent systems, layers can intermix with the preceding layer. To avoid this, the solvents in adjacent layers must be orthogonal, or the first layer must be chemically altered after deposition to resist the second layer's solvent.

Derivatives of highly aromatic small molecules, such as rylenes, have been shown to afford films with unique properties.[3, 12, 13] These properties can be tuned by altering solubility and self-assembly.[14] However, given the innumerable number of structural changes that may be made to tune the molecule's properties of the molecule, testing the solubility of each candidate can be expensive and time-consuming. Instead, access to fast, accurate predictions of solvent compatibility could aid in the design of such materials by identifying candidates with ideal properties. Machine learning is expected to be a promising strategy to improve the discovery and optimisation of supramolecular materials.[15] Quantitative structure-property relationships have emerged as a useful tool for chemical prediction, linking descriptors of chemical structure to the properties of interest, with a breadth of literature

discussing their development and application.[16-18] While QSPR tools exist for solubility prediction, the data used in these models is limited and, in some cases, unreproducible, making their application problematic.[19, 20] To overcome this, we describe here the application of our open-source QSPR framework for property classification, previously described in Chapter 2, to the problem of solubility prediction for rylenes and other similar amino acid-functionalised molecules.

3.2 Results and Discussion

3.2.1 Existing Library

Before modelling can begin, a suitable dataset must be procured or prepared. The Draper group has extensively investigated the properties of amino acid-derivatised rylene derivatives.[8, 9, 21-24] This provided a useful basis for establishing a QSPR dataset. At the outset of the project, the group had 13 NDIs and 17 PBIs readily available to characterise (Table 3.1).

Table 3.1. Library of rylene molecules previously synthesised and available within the Draper group at the project outset.

NDI-A	NDI-F	NDI-FF	NDI-G	NDI-GA
NDI-GF	NDI-GL	NDI-I	NDI-L	NDI-M
NDI-V	NDI-W	NDI-Y	PBI-A	PBI-A 1-Py
PBI-D	PBI-DOPA	PBI-E	PBI-F	PBI-G
PBI-H	PBI-I	PBI-L	PBI-L 1-Py	PBI-M
PBI-S	PBI-V	PBI-V 1-Py	PBI-W	PBI-Y

While the previously prepared library serves as a useful starting point, it was expected that without further augmentation, the small dataset would lack sufficient information to establish a truly predictive model. We discuss later in this chapter how each of these molecules was classified. Using such classifications, the modelability index can be calculated as a computationally efficient indication of the dataset amenability to modelling. (Chapter 2: Section 2.4.3) Indeed, calculating such values for the initial QSPR dataset across five solvents of interest (acetone, dimethyl formamide (DMF), dimethyl sulfoxide (DMSO), ethanol, and tetrahydrofuran (THF)) affords indexes significantly below the ideal 0.65 threshold, especially in the case of DMF (Table 3.2).

Table 3.2. Modelability index of the initial QSPR datasets based on a solubility classification threshold of 10 mg mL^{-1} for each solvent.

Solvent	Acetone	DMF	DMSO	Ethanol	THF
MODI	0.459	0.372	0.400	0.646	0.544

Subsequently, it became clear that the first step in developing an effective predictive system was to expand the data library from which predictive relationships could be established. We discuss in detail in Chapter 1 (Section 1.3.1) existing *in silico* approaches for quantifying properties such as solubility, their limitations, and the need to build quantitative machine learning systems from experimental data. While the optimal approach is to develop a bespoke dataset to ensure data homogeneity, such endeavours can be costly. Therefore, it is worthwhile first to investigate data mining of literature sources.

3.2.2 Literature Datasets

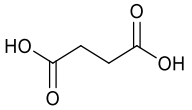
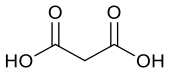
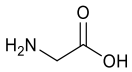
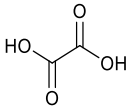
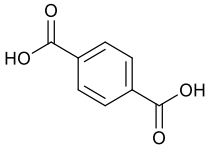
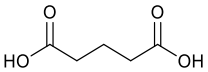
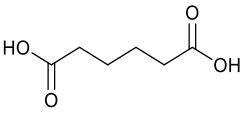
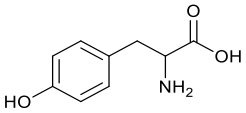
Given the importance of solubility across broad areas of chemistry, there is an increasing need for effective prediction tools. However, the development of such systems requires datasets of sufficient size and quality to realise effective predictions. This has led to the preparation of datasets such as AqSolDB, BigSolDB, and its recently expanded version, BigSolDB 2.0.[25-27] While the development of such databases is critical for the future development of solubility QSPRs, most reported databases and QSPRs focus solely on aqueous solubility. The advent of the initial BigSolDB enabled the emergence of organic solvent solubility models. BigSolDB 2.0 is one of the largest databases of organic solvent solubility data derived from peer-reviewed journal articles, serving as a starting point for developing more accurate predictive models. The manually curated dataset represents 1448 molecules across 213 solvents, providing a wealth of information that can be utilised to build predictive systems.[27] While such databases offer a significant increase in the number of molecules we can use to train a QSPR system (Table 3.3), it is important to consider the quality of data not only in terms of the robustness of the associated experiments, but also in terms of the amount of information it provides about our target chemical space.

Table 3.3. Summary of the number of molecules with solubility data provided for acetone, dimethyl formamide (DMF), dimethyl sulfoxide (DMSO), ethanol, and tetrahydrofuran (THF) in BigSolDB 2.0.[27]

Solvent	Acetone	DMF	DMSO	Ethanol	THF
Molecules	641	282	129	1016	172

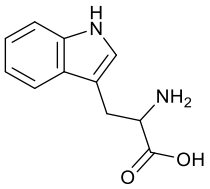
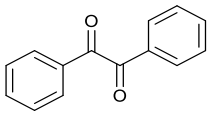
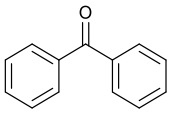
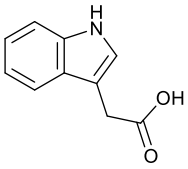
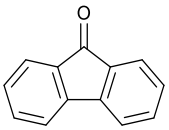
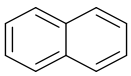
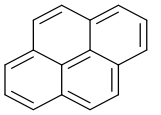
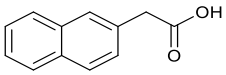
To illustrate the lack of information provided by BigSolDB 2.0 about the chemical space of interest, the molecules most chemically similar to the rylene molecules that comprised the initial QSPR dataset were identified by calculating the Manhattan distance between each pair of fingerprints across the two datasets. If we consider every molecule in BigSolDB 2.0, the molecules with the smallest distance, and therefore the greatest chemical similarity, are significantly different from the rylene structures, instead only showing a resemblance to the amino acid side chains (Table 3.4).

Table 3.4. The 8 molecules from the BigSolDB 2.0 with the smallest Manhattan distance from molecules reported in Table 3.1. The associated Manhattan distance is provided below each molecule.

			
19	19	20	20
			
20	21	21	21

Given the importance of the aromatic character of the rylene derivatives for their applications, a further inspection of the BigSolDB 2.0 dataset limited to molecules with at least 2 aromatic rings was undertaken (Table 3.5). While some molecules have a slight resemblance to the aromatic core, the majority of the dataset comprises chemicals from a very dissimilar chemical space. Subsequently, while much larger in quantity, datasets such as BigSolDB 2.0 likely provide little predictive information about our chemical space of interest. Subsequently, a targeted augmentation of the dataset with chemically similar derivatives was deemed to be necessary.

Table 3.5. The 8 molecules with at least 2 aromatic rings from the BigSolDB 2.0 with the smallest Manhattan distance from molecules reported in Table 3.1. The associated Manhattan distance is provided below each molecule.

			
21	21	22	22
			
23	24	25	25

3.2.3 Expanding the Dataset

The tuning of molecular properties afforded by amino acid derivatisation of rylenes has been key to accessing unique properties.[3] In addition, the chemical diversity of peptides and the tunability they afford for self-assembly through combinations of different amino acids make dipeptide-functionalised rylene derivatives a promising platform for tuning new functional materials. Subsequently, they were selected as target molecules to augment the QSPR dataset.

3.2.3.1 Preparing Candidate Structures

To identify the optimal molecules to expand the dataset, an *in silico* library of candidate structures was required. While such libraries can be manually prepared, even a small number of functionalisation options can yield a large library of candidates. Manually preparing such a dataset inevitably results in wds; therefore, we wrote a Python script using RDKit to install side-chain moieties onto core structures, enabling the automated generation of SMILES for each possible amino acid dipeptide candidate, based on previously published code.[28-31]

Using ChemDraw, each of the naturally occurring amino acid side chain structures (excluding proline) was prepared, using "*" to denote the attachment point. This can then be directly converted into the corresponding SMILES string (for example, alanine = [*]C, leucine = [*]CC(C)C, tyrosine = [*]=CC1=CC=C(O)C=C1), which can be easily interpreted by RDKit. Similarly, the SMILES of a core structure was prepared, denoting attachment points using [*:1] and [*:2] wildcards to indicate where different side chains are to be installed. Using RDKit, the core molecule is constructed, and a new bond is formed between

each attachment point and the relevant side chain structure to yield a new candidate molecule (Figure 3.3). The code then iterates over each permutation of sidechains, using the appropriate sidechain to install at the [*:1] and [*:2] wildcards, to generate the structures of each possible dipeptide functionalisation of the given core.

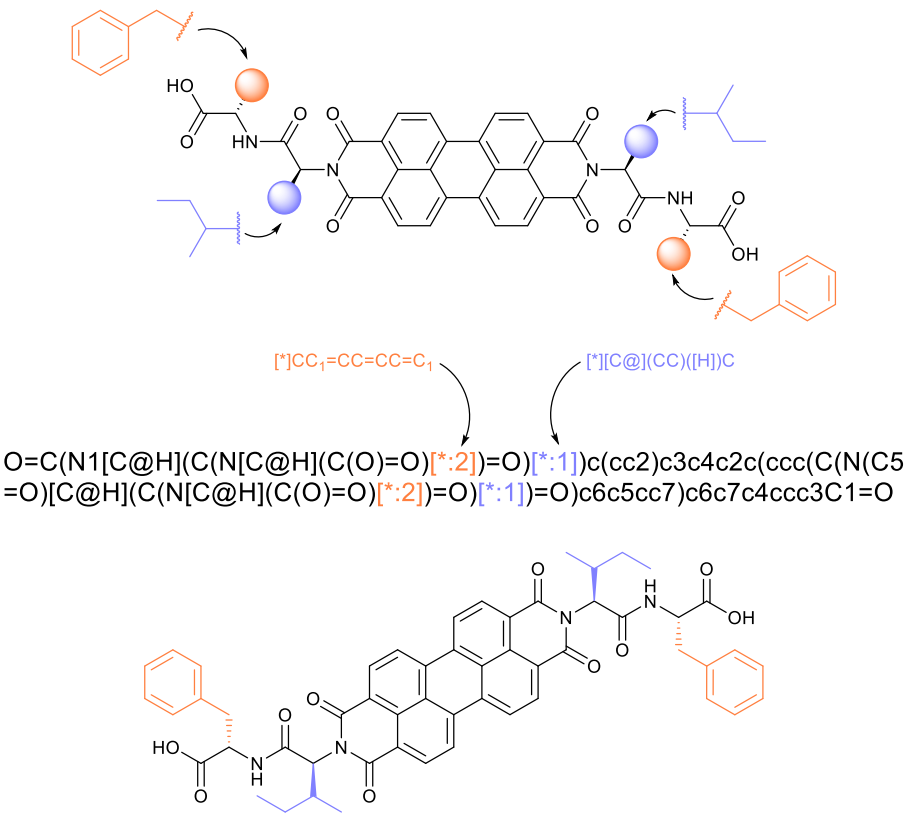


Figure 3.3. Simplified cartoon of *in silico* library generation using SMILES of core structure and side chain substituents.

One of the major issues plaguing QSPR datasets is the prevalence of incorrect chemical structure representations.[32] It has been shown previously that a lack of structure curation results in models with poor performance, whereas well-curated structures lead to improved modelling performance.[33] While this could have been achieved by manually determining the structure for each chemical candidate, applying each monopeptide and dipeptide permutation to the NDI and PBI rylene cores yielded 760 structures. Considering the scale of the task and the minute differences between structures, approaching it manually will certainly lead to errors. Reducing the manual component of the task to roughly 20 structures will yield fewer manual errors that are easier to curate.

Not only is structure curation important when manually defining structures, but it is also critical when assessing the output of computational tools. In this example, with a focus on natural amino acids, correctly capturing the chiral information in the final structures can be important. Still, it can be expressed incorrectly if care is not taken. While chemical structure

notation such as SMILES can represent chirality, the replacement of side chain strings to make new *in silico* candidates risks unintended stereochemical alterations that can be difficult to assess visually. Curation of the resulting structures after their use revealed significant stereochemical errors in the dataset, highlighting the importance of curating automated workflows and the care required when using such tools. A future system may make better use of chiral data by using Molfile representations that can better describe 3D geometry to generate *in silico* chiral candidates for investigation and prediction. However, the Morgan fingerprint used to calculate chemical similarity for molecule selection from these SMILES representations was configured to exclude stereochemistry; therefore R, S, and undefined stereochemistry yielded the same fingerprints. Because the fingerprints used were unaffected by the structural errors. This was deemed to have no impact on the resulting molecule selection.

3.2.3.2 Optimising the Training Library

The aim of synthesising dipeptide-functionalised rylenes was both to supplement the available dataset and to maximise model quality. However, by selecting the appropriate candidates, the adequately modelled chemical space could be greatly expanded. As discussed previously, an applicability domain can be calculated to describe a boundary of chemical similarity to the training set molecules. Those molecules that lie within the boundary are deemed to have sufficient chemical similarity such that the resulting QSPR model will describe them effectively. Therefore, maximising the proportion of candidate molecules considered to be within the domain represents an expansion of the appropriate chemical space. Therefore, from the *in silico* library of amino acid-derivatised rylenes, 30 new molecules were randomly chosen to augment the existing dataset. Using the augmented dataset, a new applicability domain could be calculated, and the proportion of in silico molecules deemed to be within it was evaluated. 10,000 iterations were performed and compared to identify the target combination of new molecules that best describes the chemical space. As the most pessimistic applicability domain, the dsz domain was chosen as the maximisation target to identify which molecules best expand it.

After calculating the proportion of the dataset that would be considered in-domain when adding 30 candidate molecules to those already available, the best selection was found to increase the proportion of molecules well represented by the model from 4.2% to 62.0% (Table 3.6 and Table 3.8). However, based on the wealth of knowledge on the synthesis of these molecules within the group, it was determined that several of the molecules would be

synthetically inaccessible due to the inclusion of moieties incompatible with pH-dependent purification regimes and reactive groups that would compete with the desired coupling reaction. To account for this, the scope of candidate dipeptides was limited to those which were expected to be much more synthetically accessible. The candidates were subsequently limited to the alanine, phenylalanine, glycine, isoleucine, leucine, valine, and tyrosine amino acid side chains. Considering only the accessible molecules, the top 30 candidates would increase the proportion of in-domain molecules from 21.5% to 80.8%, enabling the creation of a much more robust and useful QSPR model (Table 3.7 and Table 3.8). This process highlights a significant advantage of building models in-house over relying on dedicated computational scientists. By building such systems within the group, greater access to practical insights into the chemistries under investigation enables much more useful results.

Table 3.6. The optimal choice of 30 molecules drawn from an *in silico* library of dipeptide functionalised naphthalene diimide and perylene bisimides. Side chains are constrained to any natural amino acid side chain (excluding proline).

PBI-AK	PBI-CY	PBI-KH	PBI-KV	PBI-KW
PBI-MR	PBI-QF	PBI-QV	PBI-WS	PBI-WY
PBI-YA	NDI-HM	NDI-HY	NDI-DH	NDI-DW
NDI-EK	NDI-EW	NDI-TI	NDI-NG	NDI-QF
NDI-CY	NDI-VM	NDI-IT	NDI-IM	NDI-IF
NDI-LN	NDI-MW	NDI-FH	NDI-FQ	NDI-RR

Table 3.7. The optimal choice of 30 molecules drawn from an *in silico* library of dipeptide functionalised naphthalene diimide and perylene bisimides. Side chains are constrained to alanine, phenylalanine, glycine, isoleucine, leucine, valine, and tyrosine amino acid side chains.

NDI-AG	NDI-AI	NDI-AL	NDI-FL	NDI-IV
NDI-YI	NDI-LA	NDI-VF	NDI-VI	NDI-VY
NDI-YF	NDI-YG	NDI-YI	NDI-YL	NDI-II
NDI-VV	PBI-AY	PBI-FA	PBI-FG	PBI-FI
PBI-FV	PBI-GY	PBI-IF	PBI-IG	PBI-VA
PBI-YF	PBI-YI	PBI-YL	PBI-YV	PBI-II

Table 3.8. Improvement in model applicability domain afforded by targeted dataset expansion. Compares the existing dataset with those augmented by the molecules described in Table 3.6 and Table 3.7. For each, the proportion of molecules deemed to lie within the pessimistic domain is reported.

Dataset	Proportion of molecules considered within the dataset's dsz applicability domain	
	All functionalised rylenes	Synthetically accessible functionalised rylenes
Already available	4.2 %	21.5 %
Already available + 30 functionalised rylenes	62.0 %	86.2 %
Already available + 30 synthetically accessible functionalised rylenes.	22.0 %	80.8 %

3.2.3.3 Synthesis of Target Molecules

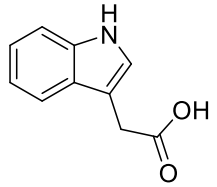
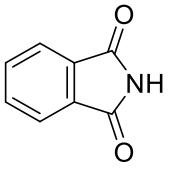
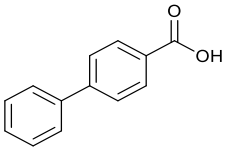
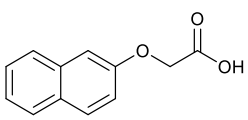
While many precursor molecules were successfully synthesised, most candidate molecules proved difficult to access due to several factors. Firstly, the dipeptides exhibited a range of physical properties that did not conform to those of previously synthesised dipeptides within the group. Notably, some dipeptides exhibited significant hygroscopicity, rendering the necessary filtration of impurities ineffective. It was hypothesised that this arose from the increased retention of excess TFA remaining in the purified products. However, the acidification with 4 M HCl in dioxane yielded a product with the same property, suggesting that it was an intrinsic limitation of some dipeptides, such as the dipeptide VA·HCL **D6**, which is susceptible to absorbing water and transition from a crystalline powder to a wet paste following purification and left at ambient conditions. Secondly, in some cases, characterisation could not be achieved due to the difficulty in separating impurities. For example, in the case of the deprotection of Boc-LA-O^{t}Bu **C4**, repeated attempts at purification proved unsuccessful, making characterisation by nuclear magnetic resonance (NMR) difficult, due to the cyclisation of the deprotected dipeptide. Further issues arose from the persistence of impurities after the cleanup procedure, which affected both naphthalene and perylene derivatives; the latter were much more difficult to clean. While excess acid to precipitate the desired product has proven robust for monopeptide derivatives, further investigation is needed to identify an effective purification procedure for the dipeptide. The results of these limitations were that, of the 30 target dipeptides expected to be synthetically accessible, only three were able to be sufficiently purified. While this result demonstrates the major limitation in experimentally developing QSPR datasets, it also highlights the necessity of accurate property prediction systems to guide the preparation of new materials. With a better understanding of the solubility and solvent compatibility, it

would be much easier to limit the selection of candidates to those with suitable orthogonal solvent solubilities that could enable their purification.

3.2.3.4 Introducing Similar Molecular Classes

A previous review by Tropsha suggests that a minimum of 40 molecules is required to establish a quantitative structure- property relationship (QSPR) with reasonable reliability.[34] Therefore, despite attempts to augment the dataset with the target rylene systems, it was deemed necessary to explore readily available molecules in a similar chemical space to augment the solubility dataset adequately. By considering the chemical similarity of select molecules in BigSolDB 2.0, promising moieties in a similar chemical space can be identified (Table 3.9). For example, 3-indolacetic acid is among the most similar molecules in the BigSolDB 2.0 dataset. While its chemical similarity likely stems from the inclusion of tryptophan moieties in the initial dataset, it may also identify amino acid-derivatised indoles as a promising candidate for future dataset expansion. Similar derivatives have previously been shown to form exceptionally strong hydrogels via self-assembly, akin to those of the rylenes of interest.[35] Likewise, phthalimide and biphenyl-4-carboxylic acid derivatives have been shown to undergo self-assembly.[36, 37] By incorporating such derivatives into future datasets, they may capture useful structure-property insights into rylenes and other classes of supramolecular molecules for the development of smart materials. Of particular interest is the similarity to 2-naphthoyloxyacetic acid, a precursor to self-assembling materials that have been extensively investigated within the Draper and Adams groups. This makes amino acid derivatised naphthalene molecules an ideal candidate for augmenting the QSPR dataset. Both owing to their well-known self-assembly properties akin to those of the rylenes of interest, which enable the development of smart materials, and to their ready access through a wealth of prior syntheses undertaken within the groups.

Table 3.9. Select molecules from the BigSolDB 2.0 with the minimum Manhattan distance from molecules reported in Table 3.1 provided below each molecule.

			
3-Indolacetic acid	Phthalimide	biphenyl-4-carboxylic acid	2-naphthoyloxyacetic acid
22	25	26	28

Using the readily available moleculess, the QSPR dataset was augmented with several 2-naphthoyloxyacetyl (2Nap) derivatives as well as a tetrahydronaphthalene and two para-substituted phenyl systems to afford a final QSPR dataset of 53 molecules following the general structure described in Figure 3.4.

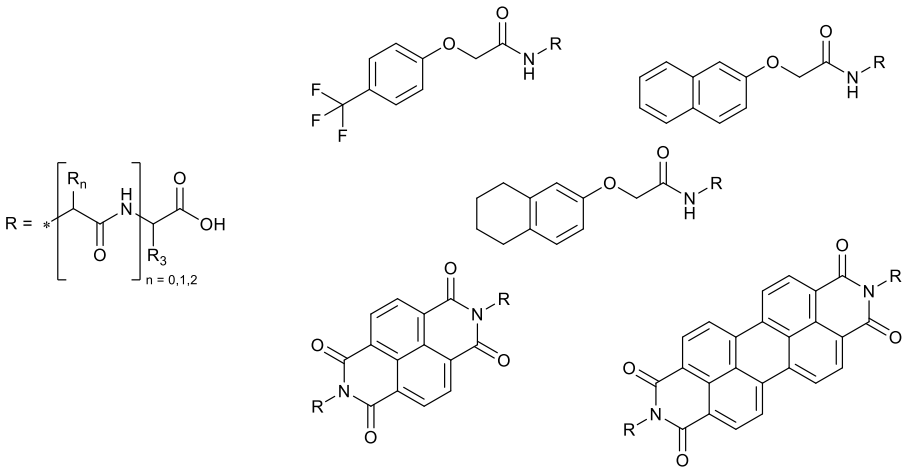


Figure 3.4. Generic chemical structures of materials library (R_n = amino acid side chains); see Appendix 3.3, molecules **1 – 53**, for specific structures.

While the alternative molecules ultimately used to augment the dataset are reasonably dissimilar from the target rylene dataset, the final dataset was sufficient to capture the desired chemical space. Surprisingly, the proportion of the synthetically accessible rylenes deemed to be within the dsz domain surpasses that of the target dataset (Table 3.10). Furthermore, calculating the MODI of the final QSPR datasets based on the solubility classifications discussed later in this chapter shows that expanding the dataset yielded the expected increase in the modelability of the solubility for each solvent compared to the initial dataset (Table 3.11, Chapter 2: Section 2.4.3). Table 3.11 suggests that further improvements to the dataset could enhance future models. However, given both the synthetic challenges of functionalising rylenes described here and their nearly complete inclusion of the candidate rylene targets within the applicability domain (as shown in Table 3.10), future attempts to expand the dataset should initially focus on expanding the applicability domain of the 2Nap candidates, which will not be as well described by the resulting models, but proved less problematic to access.

Table 3.10. Proportion of *in silico* library of mono and dipeptide functionalised rylene and 2-naphthoyloxyacetyl derivatives determined to be in-domain when constructing the dsz applicability domain with the final QSPR dataset.

	Rylenes	2Nap
Proportion of molecules considered within the dataset's dsz applicability domain	98.5 % (64 / 65)	14.3 % (8/56)

Table 3.11. Modelability index of the final QSPR datasets based on a solubility classification threshold of 10 mg mL^{-1} for each solvent.

Solvent	Acetone	DMF	DMSO	Ethanol	THF
MODI	0.506	0.460	0.438	0.663	0.626

With an adequate sample size and chemical space covered by the available library of molecules shown to be improved, their characterisation could be undertaken to build a classification dataset for QSPR modelling.

3.2.4 Characterising Solubility

Understanding the solvation properties of small molecules is crucial for predicting the morphology and characteristics of self-assembled structures. Accurate prediction of solubility and its thermodynamic factors is vital not only in soft-matter research but also across various chemical disciplines, especially in pharmaceutical research.[38] Subsequently, many tools have been developed to characterise solubility. As discussed in Chapter 1, such systems can offer powerful predictions of thermodynamic properties; however, these often come with significant limitations in computational cost and accessibility due to expertise requirements. While building machine learning systems on data procured from such systems may overcome these limitations, building such a system would introduce the biases afforded by the computational calculations into the machine learning model. Such biases could limit the predictive power of the final system. Instead, it is necessary to model the endpoint directly to build the necessary dataset.

As discussed previously, the first of the OECD requirements for effective QSPR development is that the model describes a well-defined endpoint.[39] While the solubility in a solvent of interest may seem like a simple property, it involves many considerations that must be controlled to ensure the dataset is sufficiently consistent.

The method by which solubility is measured, *e.g.*, how the insoluble material is separated from the supernatant, can have a significant impact on the reported solubility.[16] As such, we have focused solely on molecules synthesised and tested by ourselves, following a standardised protocol for each case (Section 3.4.7). Shake-flask methods are the classical approach to quantifying solubility, in which a solvent is saturated with a solute, allowed to equilibrate, and the excess solute is separated. While experimentally simple, they are time-consuming and lack a standard approach accepted across the literature.[40] A review of shake-flask solubility determination procedures by Baka *et al.* suggested that sedimentation was generally the best strategy for phase separation; however, centrifugation is the only method acceptable for materials that form micelles or aggregates.[40] It was found that centrifugation was ineffective for separating rylene materials using the available equipment. Subsequently, given the control and throughput required for effective measurements, centrifugation was deemed inadequate for separation. To ensure a consistent protocol was applied across measurements, data derived from the centrifugation protocol were disregarded and reanalysed using filtration as a means of separation. While separation by filtration may result in significant losses of low-solubility molecules due to adsorption onto filter membranes, the consistency and ease of analysis that such methods provide make them the optimal choice for a standardised protocol.

Once the supernatant is adequately separated from the undissolved solids, the amount of solubilised material needs to be quantified. Due to the small scale of shake-flask measurements, such quantities cannot be adequately determined gravimetrically by solvent removal. Subsequently, UV-visible spectrophotometric measurements (UV-vis) are a popular approach for quantifying the solute concentration in the supernatant using the Beer-Lambert relationship.[41] Significant changes in molecular aggregation and optoelectronic properties can occur simply by changing the solvent. These can result in a change in the absorbance spectrum of rylene solutions, as illustrated by a simple visual inspection (Figure 3.5). This highlights the inconsistency of the molar absorptivity across solvents, making the application of a single calibration curve to other solvent systems infeasible.

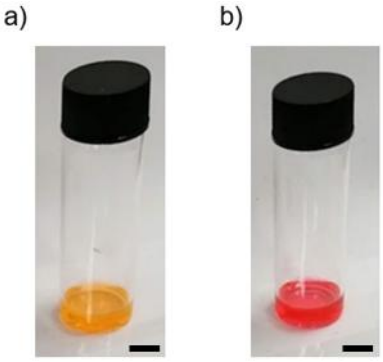


Figure 3.5. Solution of PBI-A **20** a) in THF before reconstitution and b) following reconstitution in DMSO as described in Section 3.4.7. Scale bars show 1 cm.

To establish the concentration by UV-vis absorption quantification, either a calibration curve is required for each sample in each solvent of interest, or samples must be quantified in the solvent used to prepare the calibration curve. To maximise throughput, the reconstitution method was selected to reduce the number of calibration curves required. Given that the rylenes were observed to be generally more soluble in DMSO, and the notably low vapour pressure of DMSO makes its removal by rotary evaporation unachievable, it was selected as the most appropriate solvent for determining the molar absorption coefficient (Figure 3.6, Table 3.13, Section 3.4.7 , Appendix 3.4). After determining the molar absorption coefficient (ε), the solubility classification was assigned to each molecule in the QSPR dataset in acetone, DMF, DMSO, ethanol, and THF, using a threshold of 10 mg mL^{-1} (Table 3.13, Section 3.4.7, Appendix 3.5).

Figure 3.6. Example UV-vis absorption a) dilution series and b) calibration curve in DMSO for PBI-A (**20**). The highest concentration spectrum is given in red (▬) and the lowest in purple (▬). c) Example solubility calculation for PBI-A **20**. The spectrum of the DMSO reconstituted samples is given for each solvent: acetone (▬), DMF (▬), DMSO (▬), ethanol (▬), and THF (▬).

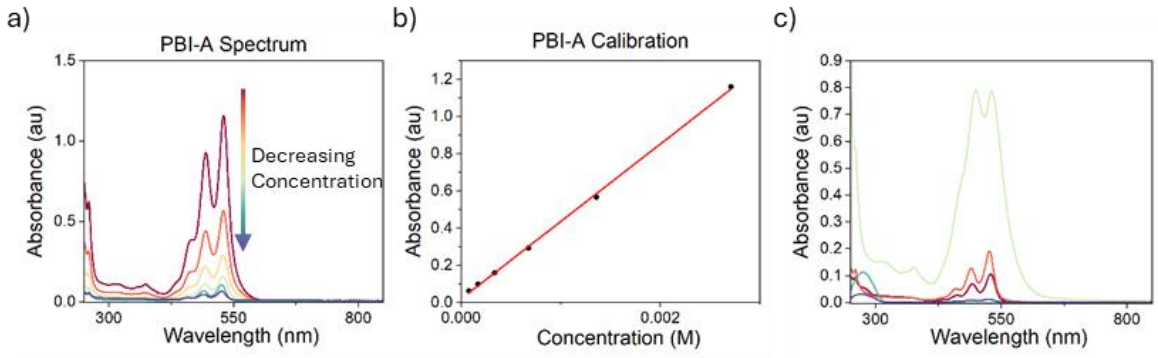


Table 3.12. Example calibration statistics for PBI-A (**20**) including calibration window for a measurement in a 0.1 mm cuvette, the molar absorption coefficient, and the R^2 goodness-of-fit metric are reported.

Min Conc. (mg mL^{-1})	0.045
Max Conc. (mg mL^{-1})	1.450
ε (M^{-1} cm^{-1})	41631
R^2 (au)	1.00

Table 3.13. Example solubility calculation for PBI-A **20**. Using the calibration curve, λ_{max}, associated absorbance, the dilution factor, and a known path length, the concentration of each sample can be determined from the Beer-Lambert relationship.

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone		0.10633	1	0.01	0.13651
DMF		0.17882	10	0.01	2.29578
DMSO	380.5	0.78319	1	0.001	10.05499
Ethanol		-0.00126	1	0.001	0.00
THF		0.012	1	0.001	0.15406

The threshold was chosen as it exceeds the typical concentration at which the self-assembly occurs for these families of materials, enabling their application in functional films.[21, 22, 42] Screening for molecules above this threshold would allow the wide chemical space of possible interesting candidates to be narrowed to those with a solubility sufficient to be cast into films. Furthermore, by understanding solubility across a range of solvents, we can not only identify useful candidates for multilayered film devices but also aid the synthesis of viable candidates by identifying acceptable solvents for purifying synthetic products. We have compiled datasets comprising >50 molecules to build the predictive models for each solvent of interest (Appendix 3.6). Class distributions varied across the solvents with 47, 83, 91, 44, and 74% of molecules being classified as soluble in acetone, DMF, DMSO, ethanol, and THF, respectively (Figure 3.7).

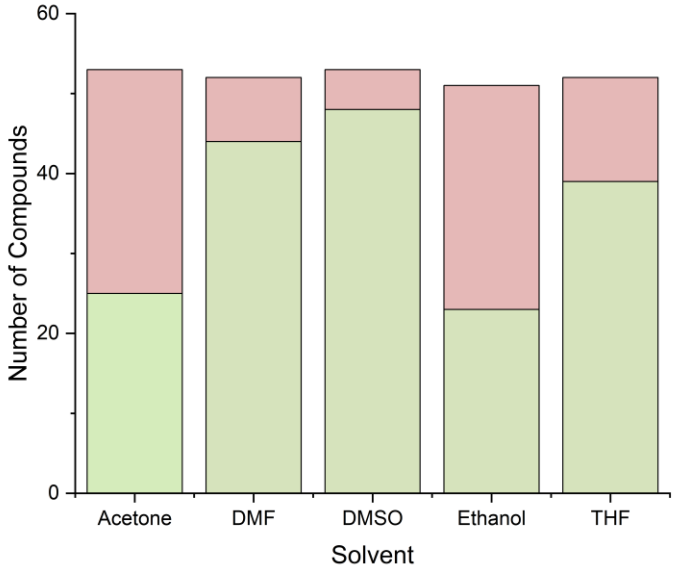


Figure 3.7. Class distribution of molecules **1 – 53,** which comprise the training and test sets across five solvents. The green segment represents the portion of molecules deemed soluble by the classification procedure described in Section 3.4.7, while red represents those deemed insoluble.

While spectrophotometric methods are generally much faster, enabling high-throughput characterisation, they cannot detect the presence or formation of impurities during solubility measurements. Subsequently, future characterisations should utilise high-performance liquid chromatography (HPLC) techniques in conjunction with spectrophotometer methods to ensure that the potential formation of more or less soluble analytes does not lead to overestimation or underestimation of the solubility of the molecule, respectively.[43] Another important consideration is whether the method used determines kinetic or thermodynamic solubility. Ideally, for a more consistent measurement, a thermodynamic equilibrium is reached. However, the solubility kinetics will vary for each molecule and therefore need to be determined experimentally. Subsequently, future protocols could be improved by periodically sampling the saturated mixture until a stable composition is achieved.[44] Furthermore, a major limitation of our protocol was the agitation method employed. While using a sample roller is a common method of agitation in shake-flask protocols, roller mixers typically lack active temperature control. Unless placed in a temperature-controlled environment, the measurement is influenced by ambient conditions, which, for solubility determination, will noticeably increase measurement error. For example, tryptophan (one of the most chemically similar molecules in our dataset in BigSolDB 2.0, Table 3.5) is shown to increase from 13.18 mg mL^{-1} to 14.08 mg mL^{-1} as the temperature is increased from 15°C to 30°C.[27] Future protocols should ensure a consistent temperature is maintained during equilibration. While such fluctuations would certainly hinder the development of regression

models, their impact on classification is reduced because the likelihood that they cause a measurement to cross the boundary should be relatively low.

While improvements to the quantification can be made for future analyses, the final QSPR dataset prepared here represents a reasonable starting point for predicting the solubility of rylenes and naphthalenes.

3.2.5 QSPR Data Preprocessing

To build a QSPR model capable of predicting a property from the chemical structure, the model must first interpret the molecule's structural features. Therefore, each molecule was generated *in silico* using ChemDraw and converted to SMILES format.[45, 46] From the SMILES representation, molecular descriptors were calculated in KNIME (Chapter 2: Section 2.4.2).[47] QSPRs tie molecular descriptors, data encoding the molecular structure and properties, to a measured property of interest by finding a relationship between them. This model can be as simple as a linear relationship or, more recently, more complex non-linear machine learning model.[48] Molecular descriptors range in complexity and information content. While a model may be built using an information-dense, three-dimensional (3D) electron-density representation for each molecule, such representations can be difficult to calculate and implement. Instead, simpler representations such as 1-dimensional (1D) counts of molecular fragments, or 0D molecular counts such as a count of hydrogen bond donors and acceptors or the molecular mass, can be used to represent the molecular structure. Indeed, simpler descriptors are crucial for predicting the relationship between molecular structure and oral bioavailability, as evidenced by their use in Lipinski's rule of five.[49] To establish such a relationship, descriptors capable of reliably segregating the property classifications are necessary (Figure 3.8).

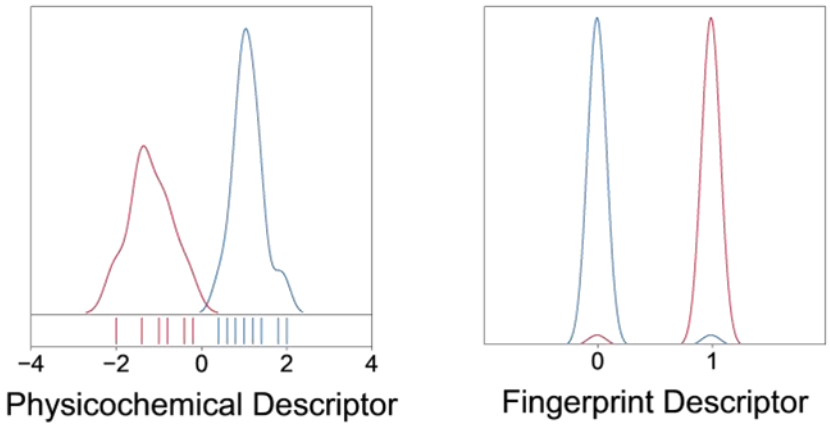


Figure 3.8. Idealised physicochemical and fingerprint descriptors. Descriptors show near-complete discrimination between the classes.

From our solubility results, no simple relationship could be found with the simplest physicochemical descriptors across each of the solvents (Figure 3.9, Appendix 3.7). Physicochemical descriptors included SlogP, atomic surface area, atomic polar surface area, molecular weight, rotatable bonds, hydrogen bond donors, hydrogen bond acceptors, number of atoms, and number of rings. While the number of rings was observed to be the best physicochemical descriptor at segregating the classifications for each solvent, it alone does not provide sufficient discrimination to afford an effective prediction. We also attempted to find a simple relationship using more information-rich Morgan fingerprint descriptors, a simple representation of the presence of different chemical substructures, but no relationship was found (Figure 3.10, Appendix 3.8).[50] The lack of such well-segregating descriptors indicates that developing Lipinski-like rules to predict solubility would be challenging for these molecules. This was expected, as the numerous interactions that affect the ability of a molecule to dissolve across a range of solvents make it difficult for a single descriptor to capture the nuances of solubility classification. Therefore, the development of QSPR classification models was deemed necessary to enable robust solubility prediction and a more robust approach to linking molecular descriptors to solubility.

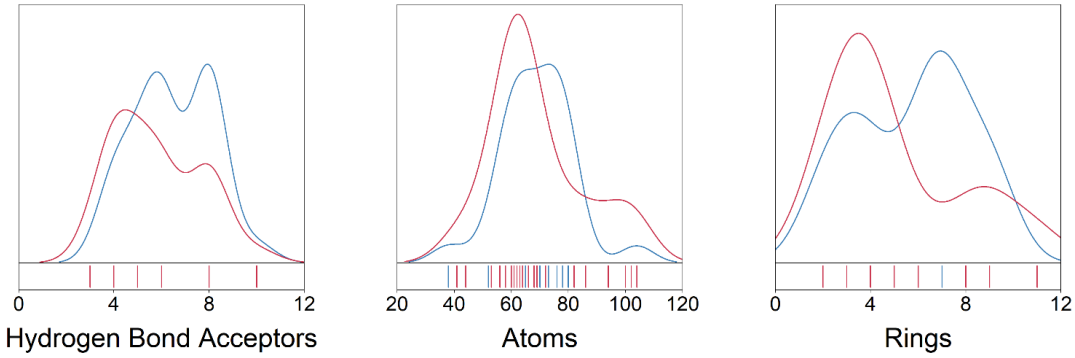


Figure 3.9. Select physicochemical descriptors calculated and their distribution, separated by acetone solubility classification (soluble, red; insoluble, blue) for the training set. The value of each molecule's descriptor is depicted by | marker.

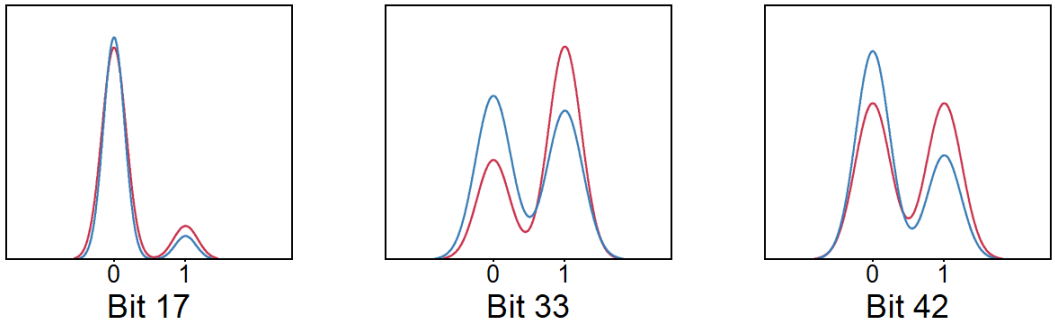


Figure 3.10. Select Morgan fingerprint descriptors and their distributions, separated by acetone solubility classification (soluble, red; insoluble, blue) for the training set.

The final step before modelling can begin is to prepare the datasets. An important step is partitioning the dataset to enable effective assessment of the performance of the model. As discussed in Chapter 2, it is necessary to segregate a portion of the data to test the performance of the machine learning models. Such models are likely to be biased by the information in the training set; thus, a 20% portion of the QSPR dataset was removed for independent testing after modelling to assess predictive performance on unseen data. However, on such limited datasets, removing even a small subset can significantly affect subsequent modelling due to shifts in the dataset's distribution and a potential loss of crucial predictive information. While the accepted models achieved MODI scores below the indicative threshold, limiting the accepted partitions to those with MODI scores above 0.65 did not yield improved models, indicating that this metric is merely an indication and not predictive of final performance. Therefore, while the MODI can indicate whether a dataset is amenable to modelling, alone it is not a good estimator of the resulting model performance (Table 3.14).

Table 3.14. MODI of training set partitions across 10 random stratified partitions, and the specific MODI values of the best-performing model.

Model	Training set MODI range	Optimal training set MODI
Acetone	0.44 – 0.61	0.51
DMF	0.43 – 0.63	0.63
DMSO	0.43 – 0.81	0.60
Ethanol	0.46 – 0.67	0.63
THF	0.52 – 0.74	0.57

To maximise the quality of the training data, it is important to screen and remove descriptors of low informational content. These include those with low variance (where the descriptors have near-constant values across samples) and high correlation (where values across the descriptors match) (Chapter 2: Section 2.4.5). The column with the most correlated columns is carried forward, and all correlated columns are filtered out. This procedure is repeated until no additional correlated columns remain. Here, we applied a low variance threshold of 0.1 and a high-correlation threshold of $r^2 = 0.9$ using Spearman's rho rank correlation to reduce the number of bit fingerprint descriptors from 1024 to between 38 and 64 descriptors of higher informational content.[51]

Utilising highly imbalanced data classes in a machine learning model can result in the minority class being poorly modelled because it is underrepresented in the training data.[52] To overcome this, class balancing techniques can be employed to ensure both classes are well represented and minimise the bias towards the majority class during the machine learning (Chapter 2: Section 2.4.6). The synthetic minority oversampling technique (SMOTE) was employed instead of simpler undersampling techniques due to the dataset's small size.[53] Using the average of a randomly selected data point and its nearest neighbours from the minority class, a new synthetic data point upsamples the minority class. The new synthetic samples are very similar to other samples in the data set, allowing the minority class to be augmented without further expanding the dataset. We highlight here that, while SMOTE was configured to synthesise new minority data points based on the 5 nearest neighbours (k) until the classes were balanced, no investigation was conducted into the optimal configuration for this dataset. SMOTE increase the number of minority-class instances by generating new ones based solely on their similarity to existing minority examples. This raises concerns about increased overfitting and noise as the number of neighbours considered increases. As discussed previously, in machine learning, there are no one-size-fits-all approaches. While k = 5 is a typical choice and the default in KNIME, if k is too small, the model may overfit and become highly sensitive to specific data points. At

the same time, if it is too large, the synthetic data points may be created in the chemical space of the majority class (Figure 3.11). Both outcomes can lead to reduced model performance.[54] Future optimisations of these models may be improved by investigating different SMOTE configurations to minimise data noise and maximise model quality. However, now with a balanced dataset, the machine learning models can be built and optimised.

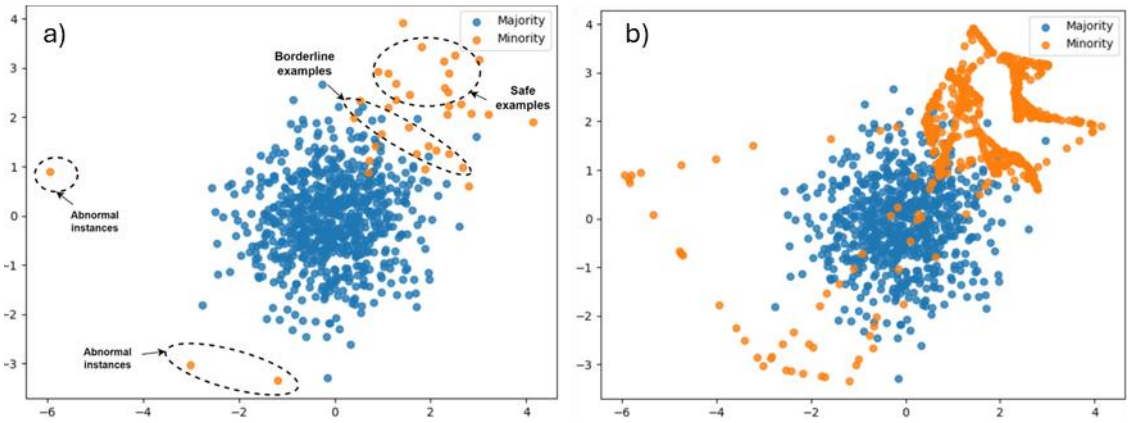


Figure 3.11. Example of an imbalanced dataset in a two-dimensional feature space. a) shows the initial dataset, highlighting areas of interest for the minority data. b) shows the SMOTE balanced dataset, illustrating how a poor configuration can introduce significant noise to the dataset. Adapted from Matharaarachchi *et al*.[54]

3.2.6 Machine Learning Model Training

The QSPR models were generated using a variety of machine learning methods: support vector machines (SVM),[55, 56] random forests (RF),[57] k nearest neighbours (kNN),[58] XGBoost (XG),[59] logistic regression (LR),[60] naive Bayesian (NB),[60] and C4.5 ensemble decision trees.[57] Each modelling method employed molecular fingerprints to capture molecular properties. As discussed in Chapter 2 (Section 2.5), each machine learning technique has its own strengths and weaknesses. By identifying multiple predictive models, our capacity to rationally design new materials with tailored properties would be significantly improved, as the average of multiple good prediction models should be closer to the actual measurement than any given individual model.

For each machine learning algorithm to successfully predict the properties of interest, its setup must be tuned to establish effective relationships. Hyperparameters are the variables of a machine learning model that can be modified to control its training process. One of the greatest barriers to efficient model development is selecting the appropriate hyperparameter values to investigate. As with other parts of machine learning, there is no universal approach that will optimally tackle each problem. Subsequently, discussions about the problem are often ambiguous, requiring one to perform a hyperparameter search with few ways to

identify the appropriate range other than by building the model and assessing the metrics. Subsequently, using the default metrics provided by KNIME and the limits of the hyperparameter ranges, a hyperparameter space was chosen (Table 3.15). The space was searched using a brute-force grid search with steps that provided a compromise between rigour and computational demand.

Table 3.15. Hyperparameter combinations optimised by a 5-fold cross-validation procedure.

Model	Parameter	Start	Stop	Step
RF	Maximum depth	10	100	10
	Number of trees	1000	5000	1000
C4.5	Minimum node size	2	10	2
	Number of trees	1000	5000	1000
	M_{Try}	0.2	0.8	0.2
	Sample fraction	0.2	0.8	0.2
	Maximum depth	10	100	10
kNN	Number of neighbours	3	20	1
SVM	Σ	-10	10	0.05
XG [a]	Depth	10	100	10
	Number of trees	1000	5000	1000
	Learning rate	0.2	0.8	0.2
LR	Learning rate	0.1	0.8	0.1
	Variance	0.1	0.8	0.1

[a] For simplicity, the log10 of the SVM σ hyperparameter input has been reported. Therefore, while all other parameters use a linear interval, the σ interval represents the logarithmic interval of each iteration.

The decision tree hyperparameter of depth is crucial: deeper trees can capture more complex relationships but are more prone to overfitting. The number of trees is less crucial: while more trees reduce prediction variance, they also increase training time, with performance plateauing with sufficient trees.[61] With the exception of the DMF RF model with a maximum depth of 20, all other models achieved optimal cross-validation metrics with maximum depths exceeding 50. However, splitting random forest trees until each leaf contains a single sample likely resulted in overfit trees, and generalised performance may have suffered as a result. Future decision tree models may benefit from restricting the maximum depth to a range of 5 – 20.[62] Support vector machines are later shown to perform poorly, typically optimising with a complexity parameter (σ) value of 10^5 or lower. When the complexity parameter is too small, the boundary's influence area is limited to the support vector, overfitting to that data point. Future models may be improved by restricting the σ parameter to a range of approximately 10^{-5} – 10^5.[63] For XGBoost, while the learning rate is 0.3 by default, lower learning rates may yield more effective models, again by reducing the risk of overfitting. While we show that the chosen space yielded effective models, reducing the

learning rate to 0.01–0.20 may be a more effective choice; this would increase computational cost, requiring more rounds of modelling to find effective models.[64] Similarly, while logistic regression offers a default learning rate of 0.1, a lower value may further improve model training by avoiding overcorrection between iterations, which can make identifying the optimal model difficult (Figure 3.12).

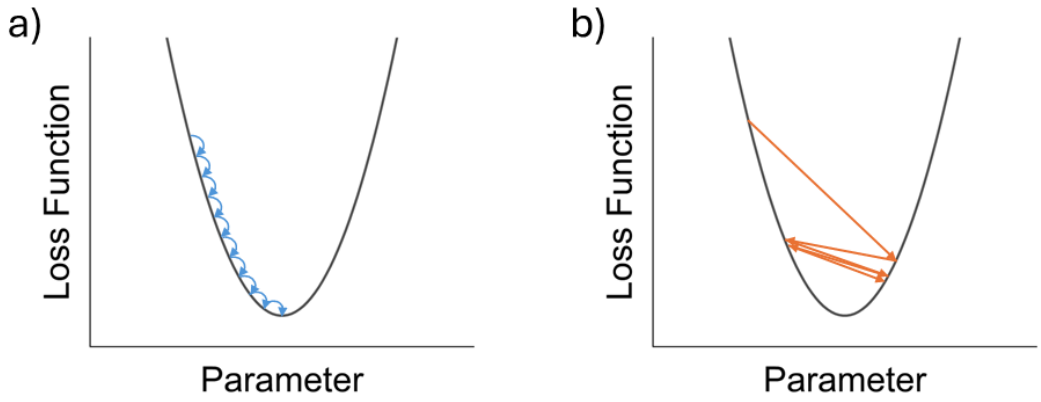


Figure 3.12. Illustration of the effect of learning rates that are a) too low or b) too high on model optimisation. Underestimating the learning rate may increase the training time required to reach the global minimum of the model; however, overestimating the learning rate may render such minima inaccessible.

A repeated 5-fold cross-validation approach was used to select the optimal QSPR model for each method based on the largest Matthew's correlation coefficient (MCC). An ideal model has an MCC value of 1, a random model has a value of 0, and an acceptable value is > 0.5. Using a cross-validated approach provides a good estimate of the models' predictive power. The models generated by each machine learning method, along with their associated statistics, are shown in Appendix 3.10. Across the different solvents, the class balance of the data varies markedly. Therefore, the metrics used must be independent of the class distribution. In situations where a model could achieve high overall accuracy simply by favouring the dominant class, balanced accuracy, MCC, and Cohen's kappa (κ) can be used when class sizes are skewed (Figure 3.7, Chapter 2: Section 2.6.1).

In any data set, model performance will vary with the dataset's composition. The negative effects of training and test set portioning can be significant, especially when modelling with small datasets. Therefore, it was expected that the changing partition of the training data would heavily impact model performance. Therefore, for each solvent model, 10 randomised partitions were investigated. The split was stratified to ensure that solubility classifications in the training set were representative of the entire dataset. The model with the best metrics and fewest out-of-domain molecules was carried forward for further investigation. However, the significant change in predictive quality as the partitions are randomised highlights a

significant limitation of the current dataset. Appendix 3.9 shows strong dataset dependence on prediction performance, which may indicate that the limited dataset employed here leads to poorer generalisation to unseen data in the final models. The training set performance of the optimal dataset split is summarised in Table 3.16.

Table 3.16. Overall quality of predictive performance of optimised QSPR models on the training set. Model is deemed to be good (G) when the following are satisfied: balanced accuracy > 0.7, MCC > 0.5, κ > 0.4. Models are otherwise deemed to be bad (B).

	Method						
Model	RF	C4.5	kNN	SVM	XG	NB	LR
Acetone	G	G	B	B	G	G	G
DMF	G	G	G	B	G	B	G
DMSO	G	G	G	B	G	B	G
Ethanol	G	G	G	G	G	B	G
THF	G	G	G	B	G	B	G

3.2.7 Applicability Domain

Assessing the model's quality by applying it only to the training set is insufficient to gauge its true predictive performance on unseen data. It is important to assess that the models provide competent predictions for molecules with sufficient chemical similarity to the training set and were not used to train the model. To assess if a candidate molecule has sufficient chemical similarity, distance-based metrics described by Sahigara *et al.*[65] were used to establish well-defined boundaries, known as "applicability domains". For each training partition, the Morgan fingerprint descriptors can be used to characterise well-defined distance-based applicability domains (Chapter 2: Section 2.4.4).

The Manhattan distance between each of the candidate descriptors and the average value for each descriptor in the training set was calculated. It is expected that the resulting QSPRs will better model molecules which are more similar to the training set. Therefore, it is important to choose the appropriate applicability domain to ensure the reliability of predictions is well-defined. By limiting test predictions to in-domain candidates within each applicability domain of interest, the impact of excluding out-of-domain molecules on predictive performance was assessed for each boundary. Typically, the maxdist was equivalent to the 2d, 3d, and p95 in performance metrics; therefore, as one of the more generous boundaries, the metrics obtained after applying the maxdist domain were compared with the much more pessimistic dsz boundary. By considering the models' performance metrics after applying the maxdist and dsz applicability domains, the impact of these domains can be identified. While maxdist performed slightly better on the DMSO models, all other models showed

comparable or better performance when employing the dsz applicability domain. In particular, for acetone prediction, only the XGBoost method showed acceptable performance using the maxdist domain, while the dsz domain yielded four predictive models with good metrics. The performance of the models on the in-domain test sets is reported in Appendix 3.10 and summarised in Table 3.17.

Table 3.17. Overall quality of predictive performance of optimised QSPR models on the in-domain dsz test set. Model is deemed to be good (G) when the following are satisfied: ROC A, balanced accuracy > 0.7, MCC > 0.5, and K > 0.4. Models are otherwise deemed to be bad (B).

	Method						
Model	RF	C4.5	kNN	SVM	XG	NB	LR
Acetone	G	G	B	B	G	B	G
DMF	G	G	G	B	G	B	G
DMSO	G	G	G	B	G	B	G
Ethanol	G	G	B	B	G	B	G
THF	G	G	B	B	G	B	G

The predictive performance of each model is further assessed using the receiver operating characteristic (ROC) curve (Appendix 3.11). ROC plots the true positive prediction rate against the false positive rate. An optimal ROC plot has a perfect true positive rate and no false positives at every threshold, with a random model showing equal rates across the curve. The "good" DMF, DMSO, and THF models showed ideal or near-ideal ROC relationships, suggesting a highly predictive model was developed. The acetone and ethanol models, while better than a random model, had poorer ROC curves, suggesting that their predictions would be less reliable (Figure 3.13).

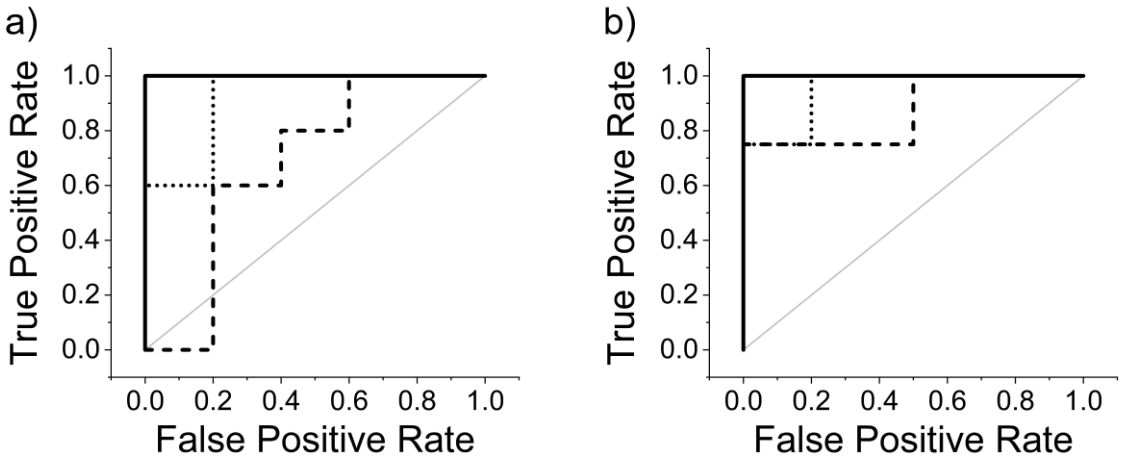


Figure 3.13. ROC curves for random forest models on a) maxdist and b) dsz in-domain test sets. The acetone model is represented by the dashed line, the ethanol by the dotted line; and the DMF, DMSO, and THF models by the solid line. The grey line indicates the ROC performance of a random predictive model.

Thus, the dsz domain was used as the domain of applicability to assess whether predicted molecules had sufficient chemical similarity to the model's training set. Given that the dsz is generally a more restricted domain of applicability, it was expected that in-domain molecules would exhibit greater chemical similarity and thus be more predictable, further rationalising this decision. For the 10 molecules in the acetone, DMF, DMSO, ethanol and THF test sets: between 6 and 9 molecules lay within the applicability domain of each model as defined by the molecular fingerprint descriptors used in the model building. Of these in-domain molecules, 4, 6, 5, 4, and 7 were soluble, and 4, 2, 1, 3, and 2 were insoluble in the respective solvents. While both classes were represented by the in-domain test set, the insoluble classes were underrepresented due to the limited number of examples of naphthalene and rylenes insoluble in DMF, DMSO, and THF. To further improve the predictive performance of future solubility models, insoluble materials within the chemical space should be targeted to better model the minority class.

3.2.8 Model Robustness

To further investigate the robustness of the optimised models, randomisation tests were conducted in which hyperparameter optimisation was conducted on a randomised training set (Chapter 2: Section 2.6.2). The performance metrics should be much poorer than those derived from the true data. These metrics identify the statistical significance of the QSPR model.[66] Indeed, the MCC metrics resulting from the randomisation tests were significantly poorer than those of the models found with true data (Table 3.18).

Table 3.18. Comparison of MCC scores of the in-domain test set on models built using the true data and models built on randomised data.

	Method							
Model	RF		C4.5		XG		LR	
Dataset	True	Randomised	True	Randomised	True	Randomised	True	Randomised
Acetone	0.78	-0.06 ± 0.15	0.78	0.16 ± 0.12	0.78	-0.14 ± 0.10	0.78	0.07 ± 0.08
DMF	1	0.00 ± 0.21	1	-0.11 ± 0.14	1	0.12 ± 0.24	1	-0.19 ± 0.25
DMSO	0.63	0.00 ± 0.15	1	-0.25 ± 0.40	1	-0.07 ± 0.34	0.63	-0.02 ± 0.20
Ethanol	0.55	-0.14 ± 0.11	0.79	0.16 ± 0.12	0.6	0.12 ± 0.12	0.6	-0.19 ± 0.10
THF	1	-0.03 ± 0.17	0.76	-0.05 ± 0.30	0.76	-0.07 ± 0.17	1	0.3 ± 0.16

This decrease in performance on random data increases our confidence in the strong predictive performance of random forest, C4.5, XGBoost, and logistic regression models. When applying the dsz applicability domain to the randomised models, all models showed a decrease in performance; however, four of these models did not show a significant decrease

(DMSO, RF; DMSO, LR; Ethanol, XG; THF, C4.5). This may indicate poorer generalisability for these models (Table 3.19).

Table 3.19. Z-scores for "good" models on the test and in-domain test sets applied to models built with randomised training data. Scores above the 0.05 confidence threshold are highlighted in green, and scores below are highlighted in red.

Model	Partition	Method			
		RF	C4.5	XG	LR
Acetone	Test	1.471	0.849	2.653	1.605
	Dsz	2.144	1.759	2.902	2.471
DMF	Test	2.250	3.109	2.663	2.696
	Dsz	2.202	2.982	1.805	2.375
DMSO	Test	1.752	2.297	2.017	2.511
	Dsz	1.624	1.959	1.821	1.446
Ethanol	Test	1.551	1.333	1.132	2.203
	Dsz	2.030	1.792	1.385	2.444
THF	Test	2.138	0.997	2.610	1.634
	Dsz	2.457	1.473	2.008	1.717

To further show that the prediction performance of the optimised models outperformed randomised models, a one-tailed one-sample t-test was performed comparing the MCC values for the optimised models with those obtained from the cross-validation hyperparameter modelling on randomised training sets (Table 3.20). The t-test shows that the improved performance of each model was at least statistically significant, with the highest p-values corresponding to the poorest Z-scoring models.

Table 3.20. One-tailed one-sample t-test p-values for "good" models on the test, and in-domain test sets applied to models built with randomised training data. Scores below the 0.05 p-value threshold are highlighted in green.

Model	Partition	Method			
		RF	C4.5	XG	LR
Acetone	Test	1.127×10^{-3}	1.255×10^{-2}	7.560×10^{-6}	3.348×10^{-4}
	Dsz	1.009×10^{-4}	1.747×10^{-4}	3.629×10^{-6}	3.744×10^{-5}
DMF	Test	2.790×10^{-5}	2.058×10^{-6}	7.331×10^{-6}	6.627×10^{-6}
	Dsz	3.297×10^{-5}	9.686×10^{-6}	6.955×10^{-4}	4.974×10^{-5}
DMSO	Test	1.802×10^{-4}	2.369×10^{-5}	6.403×10^{-5}	3.352×10^{-5}
	Dsz	6.218×10^{-4}	1.855×10^{-4}	6.611×10^{-4}	2.324×10^{-3}
Ethanol	Test	4.227×10^{-4}	1.124×10^{-3}	2.967×10^{-3}	3.284×10^{-5}
	Dsz	6.126×10^{-5}	1.529×10^{-4}	8.806×10^{-4}	1.452×10^{-5}
THF	Test	4.116×10^{-5}	5.825×10^{-3}	8.609×10^{-6}	2.932×10^{-4}
	Dsz	1.396×10^{-5}	5.938×10^{-4}	6.652×10^{-5}	2.084×10^{-4}

3.2.9 Model Validation

With the 4 models that were demonstrated to perform well at predicting solubility classification in test sets across the 5 solvents, we wished to predict the solubility of naphthalenes and rylenes from a large *in silico* library. Selected molecules were synthesised and tested to serve as a validation set and to demonstrate our approach's ability to identify both soluble and insoluble molecules.

As discussed previously, while an initial *in silico* library of dipeptide-functionalised rylenes was generated, restricting the candidates to those expected to be viable synthetic targets yielded limited success. While the synthesis of the dipeptide itself was trivial, the installation of the dipeptide onto the rylene's core proved problematic. Therefore, the literature was screened for naphthalene and rylene molecules with the general structure shown in Figure 3.4, for which synthetic pathways were reported in Reaxys®.[67] This limited the synthesis of validation molecules to those that are more likely to be synthetically accessible. The library contained 1207 molecules, each with the same set of descriptors calculated as for the training set. When predicting properties, a molecule closer to the training set typically yields more accurate predictions. For candidates within the applicability domain, defined using chemical similarity measures based on Manhattan fingerprint distances, the prediction is more trustworthy.[65] For the library of 1207 molecules, those molecules which lay outside the model applicability domain were removed, leaving only 41 – 78 molecules. For each in-domain molecule, solubility predictions were made using the accepted models. Three in-domain candidate molecules were chosen for synthesis and testing based on the combined likelihood from the four machine learning models. However, two of the molecules could not be accessed using the reported syntheses (Appendix 3.1.7). In addition, an out-of-domain candidate was synthesised for investigation. Its high prediction confidence across models, despite being out-of-domain, illustrates the model's capacity to probe chemical spaces underrepresented in the training data. The remaining validation molecule predictions agreed with the DMF, DMSO, ethanol, and THF measurements; however, the acetone measurement did not agree with its prediction for the in-domain candidate (Table 3.21).

Table 3.21. Validation prediction for each QSPR model with correct predictions highlighted in green, and incorrect predictions highlighted in red. Confidence is the average and standard deviation across the RF, C4.5, XG, and LR models.

Molecule	Applicability	Acetone Prediction (Confidence)	DMF Prediction (Confidence)	DMSO Prediction (Confidence)	Ethanol Prediction (Confidence)	THF Prediction (Confidence)
44	In-Domain	Soluble (23 ± 4.4%)	Soluble (95 ± 4.7%)	Soluble (96 ± 1.6%)	Insoluble (27 ± 19.5%)	Soluble (78 ± 1.9%)
45	Out-of-Domain	Soluble (49 ± 10.2%)	Soluble (98 ± 2.8%)	Soluble (94 ± 3.3%)	Soluble (71 ± 8.2%)	Soluble (77 ± 1.9%)

When considering the consensus of prediction confidence across the four accepted models, the misclassified prediction had the lowest confidence. This indicates that consensus confidence can help identify candidates most likely to be correctly predicted. The validation predictions indicate that, while predictive QSPR models have been developed for DMF, DMSO, and THF, further optimisation is required to improve the prediction of the solubility of amino acid functionalised rylenes and naphthalenes in acetone and ethanol.

3.2.10 Feature Importance

As discussed in Chapter 2 (Section 2.7.1), a key feature of a high-quality model is the inclusion of a mechanistic interpretation, if possible. Having identified models with predictive ability, chemical substructure importance can be calculated to extract useful insights for future investigation and rational design. Previous models have used many shallow decision trees to generate importance metrics.[68] By quantifying how often each descriptor is selected as the optimal variable to partition the data into its appropriate classes, each descriptor's importance can be determined (Chapter 2: Section 2.7.1). However, identifying the abstract fingerprint bits which contribute to the model does not provide useful information for instructing future chemical design. Instead, the chemical substructure represented by each fingerprint bit of interest was identified and visualised to probe the models' mechanisms. We have reported the first structure of the 9 most important descriptors for each model (Appendix 3.12). An example of the feature importance output is given in Figure 3.14.

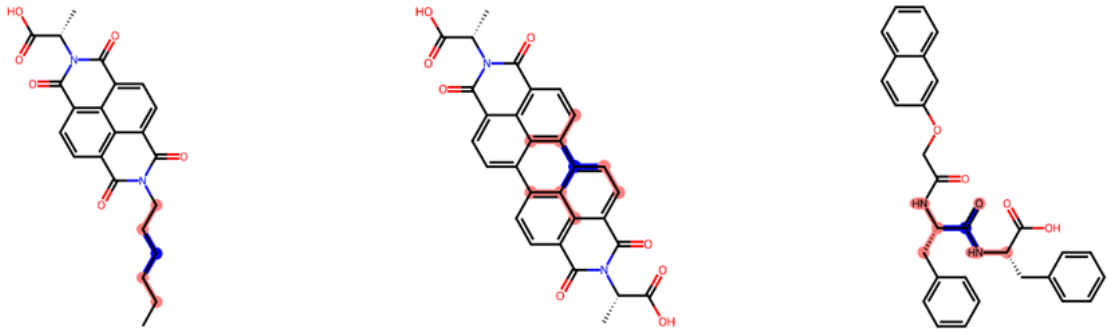


Figure 3.14. Example descriptor feature importance of the 3 most important substructures for the acetone model. The core atom represented by the fingerprint bit of interest is highlighted in blue, with the extended connectivity captured by the structural descriptor highlighted in red. The quantified importance of each descriptor is also provided.

As would be expected, descriptors relating to the core structures are among the most important, exerting a dominant influence on solubility. In addition, the peptide-link substructure is often highlighted, indicating that changing the length of the peptide attached to the core structure has a significant effect on its propensity to dissolve. Lastly, amino acid side chains are highlighted, particularly those that are highly aromatic, indicating that the presence or absence of aromatic amino acids has a greater effect on the molecule's solubility than that of aliphatic amino acids. The ethanol model, which included physicochemical descriptors, shows the number of rotatable bonds, atomic surface area, and number of atoms as the most important descriptors, which may align with the importance of the peptide link in the other models, indicating molecular size as one of the most important factors. Similarly, the number of rings is shown to be highly important, which may align with the importance of the size of the aromatic core, as well as the presence of aromatic functionalities, in governing solubility in ethanol.

While the simpler importance metric provides some insight into the important features, a metric that indicates the direction in which the presence or absence of a descriptor shifts the prediction would better aid the selection of appropriate features when rationally designing candidates. Shapley values describe the contributions of each descriptor across each possible coalition of descriptors, providing an estimate of each descriptors importance (Chapter 2: Section 2.7.1). While metrics such as Shapley values are a promising method for obtaining even deeper insight, allowing the descriptors impact on the outcome prediction to be elucidated, it was observed to be an ineffective approach on our models (Figure 3.15).[69] Shapley values can be highly unstable on randomly drawn datasets, a problem that is amplified by the limited amount of training data.[70] Subsequently, to obtain more detailed

insights, a larger data library is required. This would aid both the stability of the Shapley values and the model robustness, synergistically improving the insights that can be obtained.

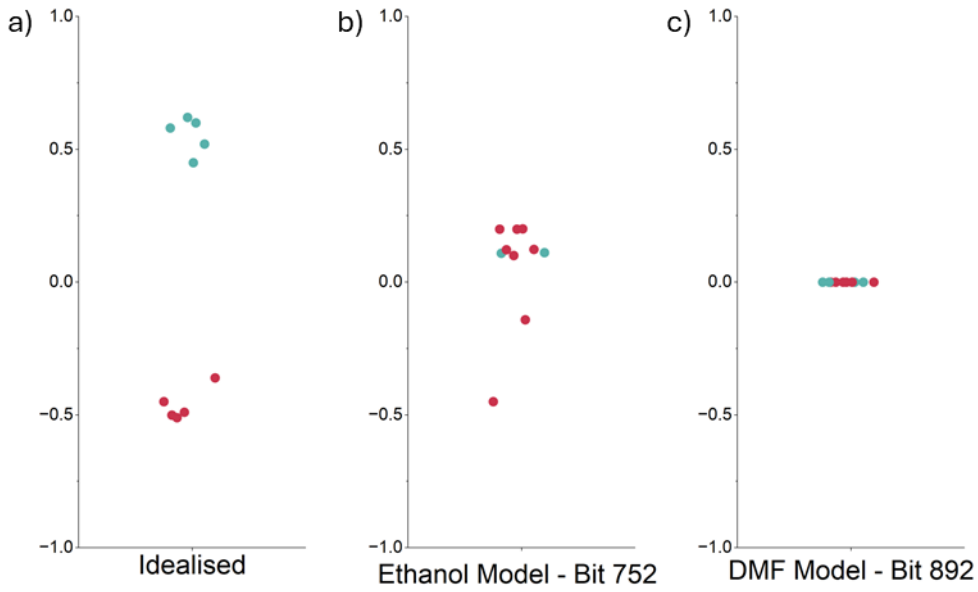


Figure 3.15. Beeswarm plots of Shapley values for test set molecule predictions on the optimised models. Green data points indicate a molecule for which the descriptor is present, and red data points indicate its absence. a) shows an idealised Shapley value plot that shows a clear relationship between the presence of a descriptor and its classification. b) and c) are examples of the typical Shapley value beeswarm plots obtained with the reported models and limited data with an incoherent relationship or an absence of any correlation between the descriptor and classification, respectively.

3.3 Conclusions

In conclusion, we have demonstrated that the solubility of naphthalene and rylene derivatives can be modelled using QSPR models. We have shown that investigating physicochemical descriptions alone is insufficient to elucidate the complexity governing solubility, and that more robust machine learning approaches are required to model how the nuances of chemical structure impact bulk properties. The open-source framework we have developed provides solubility predictions and can filter out molecules deemed too chemically dissimilar to make adequate predictions. If a molecule lies outside of the applicability domain, it does not guarantee that the prediction is incorrect, but it helps inform the likelihood of a molecule's solubility in the given solvent. While we have used this framework to investigate the solubility of naphthalenes and rylenes, the workflow has been developed such that it can be adapted to build models for any chemical classification problem or further improved with the addition of further solubility data.

While we have used this framework to investigate the solubility of naphthalenes and other aromatics, the workflow is designed to be adaptable to building models for any chemical

classification problem. It can be further improved with additional solubility data. While machine learning is expected to be a promising strategy for improving the discovery and optimisation of supramolecular materials, there is currently little effort to unify characterisation across the field. We stress here that the development of more complex QSPR models to aid supramolecular material design would be enhanced by standardised characterisation protocols that enable high-quality training data preparation from literature review methods.

We invite researchers to use this open-source tool to predict the solubility of molecules similar to those discussed in this paper, retrain the models with their own solubility data, or tackle another chemical classification problem entirely. The classification workflow can be found freely available for use at the following address:

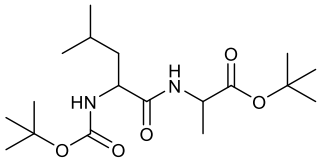
(https://hub.knime.com spaces/Public/Classification%20QSPR).

3.4 Experimental

3.4.1 Materials and Synthetic Procedures

All chemicals and solvents were purchased from Alfa Aesar, Fischer Scientific, Honeywell, Sigma-Aldrich, Thermo Scientific, and VWR Chemicals and used as received. Deionised water was used throughout. Many of the molecules discussed were previously synthesised within the group; however, several dipeptide-substituted rylenes were synthesised for this project. A comprehensive list of molecules used in this chapter is described in Appendix 3.1. In the interest of brevity, a representative synthesis has been reported in full, and the subsequent syntheses tabulated in Table 3.22 – Table 3.32 and the data of all additional syntheses have instead been tabulated in Appendix 3.1.

3.4.2 Representative Synthesis of Protected Dipeptides



C4

To a 250 mL round bottom flask, 2.71 g (11.70 mmol) of Boc-$_L$-leucine **A5** was dissolved in 100 mL of chloroform. 2.93 mL (26.60 mmol, 2.3 eq.) of N-methyl morpholine**,** 1.58 mL (12.11 mmol, 1.0 eq.) of isobutyl chloroformate, and 2.39 g (13.10 mmol, 1.1 eq.) $_L$-Alanine tert-butyl ester hydrochloride **B1**, was added to the solution and allowed to stir for 5 minutes between each addition. Solution is then allowed to stir overnight. The solution was then transferred to a separating funnel, to which 200 mL of hydrochloric acid and 100 mL of chloroform was also added. The mixture was shaken thoroughly and partitioned. The organic phase was washed with 150 mL of saturated brine and partitioned into a clean vessel, dried with magnesium sulfate, and gravity filtered. The filtrate was then rotary evaporated to yield crude product **C4**. A subsample of the crude product was purified by column chromatography (2:8 ethyl acetate:hexane) in order to characterize protected dipeptide **C4**.

Yield 3.42g (9.54 mmol, 81%). δ_H (400 MHz, DMSO-d_6) 8.07 (1H, d, *J* 7.0, N$\underline{H}$), 6.80 (1H, d, *J* 8.6, N$\underline{H}$), 4.09 (1H, dq, *J* 7.2, 7.2, NH-C$\underline{H}_A$), 3.98 (1H, td, *J* 8.9, 6.0, NH-C$\underline{H}_L$), 1.62 (1H, dt, *J* 13.5, 6.8, C$\underline{H}_L$), 1.43 – 1.39 (2H, m, C$\underline{H}_{2L}$), 1.37 (18H, d, *J* 4.5, C(CH_3)$_3$), 1.24 (3H, d, *J* 7.3, C$\underline{H}_{3A}$), 0.86 (6H, dd, *J* 6.6, 5.1, C$\underline{H}_{3V}$). δ_C (101 MHz, DMSO-d_6) 172.4, 171.7, and 155.3 (C=O), 80.2 and 77.8 (C(CH$_3$)$_3$); 52.3 (NH-$\underline{C}H_A$), 48.2 (NH-$\underline{C}H_A$); 40.9 ($\underline{C}H_{2L}$),

28.2 and 27.6 ($C(CH_3)_3$), 24.2 ($\underline{C}H_V$), 23.1 and 21.4 ($\underline{C}H_{3V}$), 16.9 ($\underline{C}H_{3A}$). HRMS (ESI) m/z: [M+H]+ calculated 359.2540, found 359.2550.

Table 3.22. Reagent proportions and product yields for the synthesis of protected dipeptides prepared by Connor MacDonald.

Boc reagent	^{t}Bu reagent	Product	Crude yield (%)
Boc-F **A2**, 5.60 g (21.10 mmol)	V-O^{t}Bu·HCl **B6**, 4.64 g (22.10 mmol)	Boc-FV-O^{t}Bu **C1**, 7.97 g (18.94 mmol)	90
Boc-I **A4**, 4.74 g (20.48 mmol)	F-O^{t}Bu·HCl **B2**, 5.53 g (21.47 mmol)	Boc-IF-O^{t}Bu **C2**, 8.48 g (19.50 mmol)	105
Boc-I **A4**, 5.24 g (22.66 mmol)	V-O^{t}Bu·HCl **B6**, 5.02 g (23.93 mmol)	Boc-IV-O^{t}Bu **C3**, 8.69 g (21.41 mmol)	95
Boc-L **A5**, 5.46 g (25.13 mmol)	A-O^{t}Bu·HCl **B1**, 4.79 g (26.37 mmol)	Boc-LA-O^{t}Bu **C4,** 7.65 g (23.73 mmol)	88
Boc-V **A6**, 2.65 g (12.20 mmol)	A-O^{t}Bu·HCl **B1,** 2.53 g (13.92 mmol)	Boc-VA-O^{t}Bu **C5**, 4.97 g (14.43 mmol)	118
Boc-Y-^{t}Bu **A7**, 5.68 g (16.83 mmol)	F-O^{t}Bu·HCl **B2**, 4.55 g (17.64 mmol)	Boc-YF-O^{t}Bu **C6**, 8.60 g (15.90 mmol)	94
Boc-Y-^{t}Bu **A7**, 6.00 g (17.93 mmol)	L-O^{t}Bu·HCl **B5**, 4.19 g (18.73 mmol)	Boc-YL-O^{t}Bu **C7**, 8.58 g (19.03 mmol)	94

Table 3.23. Reagent proportions and product yields for the synthesis of protected dipeptides prepared by Dr. Rebecca Ginesi

Boc reagent	^{t}Bu reagent	Product	Crude yield (%)
Boc-A **A1**, 2.29 g (12.10 mmol)	L-O^{t}Bu·HCl **B5,** 2.84 g (12.69 mmol)	Boc-AL-O^{t}Bu **C8,** 4.12 g (11.49 mmol)	95
Boc-F **A2**, 2.97 g (11.19 mmol)	A-O^{t}Bu·HCl **B1,** 2.16 g (11.89 mmol)	Boc-FA-O^{t}Bu **C9,** 3.32 g (8.46 mmol)	76
Boc-F **A2**, 5.42 g (20.43 mmol)	L-O^{t}Bu·HCl **B5,** 4.81 g (21.50 mmol)	Boc-FL-O^{t}Bu **C10,** 7.12 g (16.38 mmol)	80
Boc-I **A4**, 5.89 g (25.47 mmol)	G-O^{t}Bu·HCl **B3,** 4.44 g (26.49 mmol)	Boc-IG-O^{t}Bu **C11,** 7.32 g (21.25 mmol)	83
Boc-I **A4**, 5.11 g (22.09 mmol)	I-O^{t}Bu·HCl **B4,** 5.17 g (23.11 mmol)	Boc-II-O^{t}Bu **C12,** 10.44 g (26.06 mmol)	118
Boc-V **A6**, 5.10 g (23.47 mmol)	V-O^{t}Bu·HCl **B6,** 5.17 g (24.65 mmol)	Boc-VV-O^{t}Bu **C13,** 7.45 g (20.00 mmol)	85
Boc-Y-^{t}Bu **A7**, 6.17 g (18.29 mmol)	V-O^{t}Bu·HCl **B6,** 4.03 g (19.22 mmol)	Boc-YV-O^{t}Bu **C14,** 5.23 g (12.17 mmol)	67

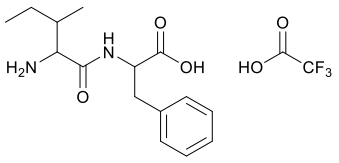
Table 3.24. Reagent proportions and product yields for the synthesis of protected dipeptides prepared by Dr. Bart Dietrich

Boc reagent	^{t}Bu reagent	Product	Crude yield (%)
Boc-A **A1**, 3.24 g (17.10 mmol)	G-O^{t}Bu·HCl **B3,** 3.01 g (17.96 mmol)	Boc-AG-O^{t}Bu **C15,** 4.12 g (19.15 mmol)	112
Boc-A **A1**, 3.56 g (18.80 mmol)	I-O^{t}Bu·HCl **B4,** 4.42 g (19.74 mmol)	Boc-AI-O^{t}Bu **C16,** 6.80 g (18.99 mmol)	101
Boc-A **A1**, 1.37 g (7.22 mmol)	Y-O^{t}Bu·HCl **B7,** 2.50 g (7.58 mmol)	Boc-AY-O^{t}Bu **C17,** 3.30 g (7.10 mmol)	98

Boc-F **A2**, 4.40 g (16.60 mmol)	G-O^{t}Bu·HCl **B3,** 2.92 g (17.43 mmol)	Boc-FG-O^{t}Bu **C18,** 6.25 g (16.43 mmol)	99
Boc-F **A2**, 5.02 g (18.90 mmol)	I-O^{t}Bu·HCl **B4,** 4.45 g (19.85 mmol)	Boc-FI-O^{t}Bu **C19,** 7.77 g (17.90 mmol)	95
Boc-G **A3**, 1.33 g (7.56 mmol)	Y-O^{t}Bu·HCl **B7,** 2.62 g (7.94 mmol)	Boc-GY-O^{t}Bu **C20,** 3.23 g (7.18 mmol)	95
Boc-I **A4**, 1.86 g (8.03 mmol)	Y-O^{t}Bu·HCl **B7,** 2.78 g (8.43 mmol)	Boc-IY-O^{t}Bu **C21,** 4.06 g (8.03 mmol)	100
Boc-V **A6**, 3.62 g (16.70 mmol)	F-O^{t}Bu·HCl **B2,** 4.51 g (17.54 mmol)	Boc-VF-O^{t}Bu **C22,** 6.62 g (15.70 mmol)	94
Boc-V **A6**, 4.19 g (19.30 mmol)	I-O^{t}Bu·HCl **B4,** 4.53 g (20.27 mmol)	Boc-VI-O^{t}Bu **C23,** 7.32 g (18.91 mmol)	98
Boc-V **A6**, 1.65 g (7.61 mmol)	Y-O^{t}Bu·HCl **B7,** 2.56 g (7.76 mmol)	Boc-VY-O^{t}Bu **C24,** 3.79 g (7.69 mmol)	101
Boc-Y-^{t}Bu **A7**, 5.15 g (15.30 mmol)	G-O^{t}Bu·HCl **B3,** 2.69 g (16.07 mmol)	Boc-YG-O^{t}Bu **C25,** 6.99 g (15.45 mmol)	101
Boc-Y-^{t}Bu **A7**, 5.52 g (16.30 mmol)	I-O^{t}Bu·HCl **B4,** 3.84 g (17.12 mmol)	Boc-YI-O^{t}Bu **C26,** 8.46 g (16.63 mmol)	102

3.4.3 Representative Deprotection of Peptides

3.4.3.1 Representative Trifluoroacetic Acid Deprotection



D1

8.32g (19.14 mmol) of crude product Boc-IF-O^{t}Bu **C2** was dissolved in 50 mL of chloroform and 25 mL (326 mmol) of trifluoroacetic acid, and allowed to stir overnight. The solution was then rotary evaporated to near dryness and dissolved in a further 50 mL of chloroform and 25 mL of trifluoroacetic acid, and allowed to stir overnight. The solution was rotary evaporated to near dryness, dissolved in 200 mL of diethyl ether while stirring to yield a precipitate. Precipitate was filtered under vacuum and dried in an oven at 50°C under vacuum overnight yielding dipeptide salt **D1**.

Yield 5.32 g (16.81 mmol, 82%). δ_H (400 MHz, DMSO-d_6) 8.69 (1H, d, *J* 7.6, N$\underline{H}$), 8.50 – 7.50 (2H, bs, N$\underline{H}$) 7.34 – 7.18 (5H, m, $\underline{H}_{ArF}$), 4.55 – 4.45 (1H, m, NH-C$\underline{H}_F$), 3.64 (1H, d, *J* 5.1, NH-C$\underline{H}_I$), 3.09 (1H, dd, *J* 14.1, 5.2, C$\underline{H}_{2F}$), 2.95 (1H, dd, *J* 14.1, 8.8, C$\underline{H}_{2F}$), 1.92 – 1.77 (1H, m, C$\underline{H}_I$), 1.46 (1H, ddd, *J* 13.4, 7.5, 3.4, C$\underline{H}_{2I}$), 1.08 (1H, ddd, *J* 16.5, 14.4, 7.2, C$\underline{H}_{2I}$), 0.93 (3H, d, *J* 6.9, C$\underline{H}_{3I}$), 0.84 (3H, t, *J* 7.4, C$\underline{H}_{3I}$). δ_C (101 MHz, DMSO-d_6) 172.3 and 168.2 ($\underline{C}$=O), 158.6, 158.3, 158.0, and 157.7 (F_3C-$\underline{C}$=O), 137.3, 129.1, 128.3, and 126.6 ($\underline{C}_{ArF}$),

121.7, 118.8, 115.8, 112.8 ($\underline{C}F_3$), 56.5 (NH-$\underline{C}H_F$), 53.9 (NH-$\underline{C}H_I$), 36.5 ($\underline{C}H_{2F}$), 36.4 ($\underline{C}H_I$), 23.5 ($\underline{C}H_{2I}$), 14.5 and 11.2 ($\underline{C}H_{3I}$). HRMS (ESI) m/z: [M+H]+ calculated 279.1703, found 279.1712.

Table 3.25: Reagent proportions and product yields for TFA dipeptide deprotection prepared by Connor MacDonald.

Crude reagent	Product	Yield (%) [a]
Boc-IF-O^{t}Bu **C2**, 8.32 g (19.14 mmol)	IF·TFA **D1**, 5.32 g (16.81 mmol)	82
Boc-LA-O^{t}Bu **C4,** 3.32 g (9.26 mmol)	LA·TFA **D2**, 1.06 g (3.35 mmol)	29
Boc-YF-O^{t}Bu **C6**, 8.41 g (15.55 mmol)	YF·TFA **D3**, 4.87 g (13.30 mmol)	79

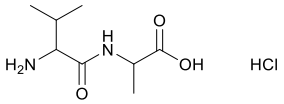
Table 3.26: Reagent proportions and product yields for TFA dipeptide deprotection prepared by Dr. Rebecca Ginesi

Crude reagent	Product	Yield (%) [a]
Boc-AL-OtBu **C8**, 4.02 g (11.21 mmol)	AL·TFA **D7**, 2.29 g (7.24 mmol)	60
Boc-FA-O^{t}Bu **C9**, 3.22 g (8.21 mmol)	FA·TFA **D8**, 2.53 g (7.22 mmol)	65
Boc-FL-O^{t}Bu **C10**, 7.02 g (16.15 mmol)	FL·TFA **D9**, 5.97 g (15.21 mmol)	74
Boc-IG-O^{t}Bu **C11**, 7.22 g (20.96 mmol)	Could not be isolated	-
Boc-II-O^{t}Bu **C12**, 10.34 g (25.81 mmol)	II·TFA **D10**, 1.19 g (3.32 mmol)	15
Boc-YV-O^{t}Bu **C14**, 5.13 g (11.94 mmol)	YV·TFA **D11**, 4.72 g (11.97 mmol)	65

Table 3.27: Reagent proportions and product yields for TFA dipeptide deprotection prepared by Dr. Bart Dietrich

Crude reagent	Product	Yield (%) [a]
Boc-AG-O^{t}Bu **C15**, 5.13 g (17.00 mmol)	AG·TFA **D12**, 3.71 g (14.28 mmol)	84
Boc-FG-O^{t}Bu **C18**, 6.12 g (16.20 mmol)	FG·TFA **D13**, 4.83 g (14.42 mmol)	89
Boc-VY-O^{t}Bu **C24**, 3.65 g (7.41 mmol)	VY·TFA **D14**, 1.63 g (4.15 mmol)	56

3.4.3.2 Representative Hydrochloric Acid Deprotection



D6

7.65 g (23.73 mmol) of crude product Boc-VA-O^{t}Bu **C5** was dissolved in 50 mL (200 mmol) of 4M HCl in dioxane and allowed to stir overnight before rotary evaporating to near dryness. 200 mL of diethyl ether was added to the mixture and allowed to stir yielding a precipitate. Precipitate was filtered under vacuum and dried in an oven at 50°C under vacuum overnight yielding dipeptide salt **D6**.

Yield 4.87g (21.69 mmol, 86%). δ_H (400 MHz, DMSO-d_6) 8.71 (1H, d, *J* 8.4, N$\underline{H}$), 7.37 – 7.21 (2H, bs, N$\underline{H}$), 7.37 – 7.21 (5H, m, $\underline{H}_{ArF}$), 4.18 (2H, ddd, *J* 8.5, 5.6, 3.2, NH-C$\underline{H}_V$ and NH-C$\underline{H}_F$), 3.15 (1H, dd, *J* 14.1, 5.6, C$\underline{H}_{2F}$), 2.98 (1H, dd, *J* 14.1, 7.5, C$\underline{H}_{2F}$), 2.16 – 2.03 (1H, m, C$\underline{H}_V$), 0.92 (6H, dd, *J* 6.8, 1.6, C$\underline{H}_{3V}$). δ_C (101 MHz, DMSO-d_6) 172.2 and 168.2 ($\underline{C}$=O), 134.8, 129.6, 128.4 and 127.1 ($\underline{C}_{ArF}$), 57.5 (NH-$\underline{C}H_F$), 53.0 (NH-$\underline{C}H_V$), 36.8 ($\underline{C}H_{2F}$), 29.9 ($\underline{C}H_V$), 19.1 and 18.0 ($\underline{C}H_{3V}$). HRMS (ESI) m/z: [M+H]+ calculated 265.1547, found 265.1553.

Table 3.28: Reagent proportions and product yields for HCl dipeptide deprotection prepared by Connor MacDonald.

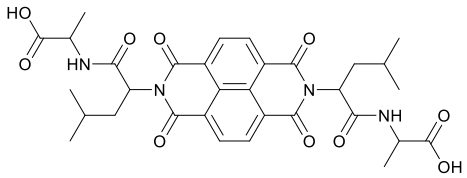
Crude reagent	Product	Yield (%) [a]
Boc-FV-O^{t}Bu **C1**, 7.76 g (18.45 mmol)	FV·HCl **D4**, 5.32 g (17.68 mmol)	84
Boc-IV-O^{t}Bu **C3,** 8.08 g (20.89 mmol)	IV·HCl **D5**, 4.35 g (16.30 mmol)	29
Boc-VA-O^{t}Bu **C5**, 8.65 g (23.73 mmol)	VA·HCl **D6**, 4.87 g (21.69 mmol)	86
Boc-YL-O^{t}Bu **C7** 6.00 g (17.93 mmol)	Could not be isolated	-

Table 3.29: Reagent proportions and product yields for HCl dipeptide deprotection prepared by Dr. Bart Dietrich

Crude reagent	Product	Yield (%) [a]
Boc-AI-O^{t}Bu **C16**, 6.70 g (18.70 mmol)	AI·HCl **D15**, 3.82 g (16.08 mmol)	86
Boc-AY-O^{t}Bu **C17**, 3.30 g (7.10 mmol)	AY·HCl **D16**, 1.70 g (5.89 mmol)	83
Boc-FI-O^{t}Bu **C19**, 7.77 g (17.90 mmol)	FI·HCl **D17**, 5.30 g (16.83 mmol)	94
Boc-GY-O^{t}Bu **C20**, 3.12 g (6.92 mmol)	GY·HCl **D18**, 1.53 g (5.536 mmol)	80

Boc-IY-O^{t}Bu **C21**, 3.92 g (7.74 mmol)	IY·HCl **D19**, 2.32 g (7.01 mmol)	91
Boc-VF-O^{t}Bu **C22**, 6.45 g (15.30 mmol)	VF·HCl **D20**, 4.25 g (14.08 mmol)	92
Boc-VI-O^{t}Bu **C23**, 7.12 g (18.40 mmol)	VI·HCl **D21**, 3.74 g (13.98 mmol)	76
Boc-YG-O^{t}Bu **C25**, 6.82 g (15.10 mmol)	YG·HCl **D22**, 3.86 g (14.04 mmol)	93
Boc-YI-O^{t}Bu **C26**, 8.17 g (16.10 mmol)	YI·HCl **D23**, 4.32 g (13.04 mmol)	81

3.4.4 Representative Synthesis of Imide Functionalised Rylenes



48

In a 100 mL Schlenk flask, 0.46 g (1.71 mmol) of 1,4,5,8-naphthalenetetracarboxylic dianhydride (NTCDA **E1**), 0.63 g (1.88 mmol) of LA·TFA **D2**, and 1.0301 g (15.15 mmol) of imidazole were added. These were stirred and purged with nitrogen for 10 minutes. The mixture was then heated to 120°C whilst under nitrogen and left stirring overnight. The reaction was cooled to 90°C and 50 mL of deionised water was added. The reaction was left for an hour and then filtered to remove unreacted NTCDA. 100 mL 2 M HCl was then added after which the precipitate was collected by vacuum filtration and allowed to dry. The precipitate was added to 2 M HCl and heated to 80°C overnight. The precipitate was collected by vacuum filtration and washed thoroughly with deionised water yielding NDI-LA **48**.

Yield 0.60 g (0.55 mmol, 58%). δ_H (400 MHz, DMSO-d_6) 12.49 (2H, s, COO$\underline{H}$), 8.71 (4H, d, *J* 2.5, Core $\underline{H}_{Ar}$), 8.34 – 8.15 (2H, m, N$\underline{H}$), 5.55 – 5.36 (2H, m, N-C$\underline{H}_L$), 4.39 – 4.17 (2H, m, N-C$\underline{H}_A$), 2.27 – 2.12 (2H, m, C$\underline{H}_{2L}$), 2.10 – 1.81 (2H, m, C$\underline{H}_{2L}$), 1.56 – 1.43 (2H, m, C$\underline{H}_L$), 1.24 (1H, dd, *J* 7.3, 2.4, C$\underline{H}_{3A}$), 1.08 (5H, d, *J* 7.3, C$\underline{H}_{3A}$), 0.91 (6H, d, *J* 6.5, C$\underline{H}_{3L}$), 0.84 (6H, dd, *J* 6.8, 1.8, C$\underline{H}_{3L}$).δ_C (101 MHz, DMSO-d_6) 174.3, 168.3, and 162.8 ($\underline{C}$=O), 130.5, 126.7, 126.5 (Core C_{Ar}), 52.5 (N-$\underline{C}H_L$), 47.6 (N-$\underline{C}H_A$), 37.2 ($\underline{C}H_{2L}$), 24.9 ($\underline{C}H_L$), 23.1 and 22.1 ($\underline{C}H_{3L}$), 17.0 ($\underline{C}H_{3A}$). HRMS (ESI) m/z: [M+H]+ calculated 637.2504, found 637.2503.

Table 3.30: Reagent proportions and product yields for imide functionalised rylenes prepared by Connor MacDonald.

Dipeptide reagent	Rylene Reagent	Product	Yield (%)
IF·TFA **D1**, 1.26 g (3.98 mmol)	PTCDA **E2**, 0.60 g (1.52 mmol)	Could not be isolated	-
LA·TFA **D2**, 0.63 g (1.88 mmol)	NTCDA **E1**, 0.46 g (1.71 mmol)	NDI-LA **48**, 0.60 g (0.55 mmol)	58
YF·TFA **D3**, 1.43 g (3.91 mmol)	NTCDA **E1**, 0.50 g (1.87 mmol)	Could not be isolated	-
YF·TFA **D3**, 1.12 g (3.05 mmol)	PTCDA **E2**, 0.59 g (1.50 mmol)	Could not be isolated	-
FV·HCl **D4**, 1.12 g (3.72 mmol)	PTCDA **E2**, 0.70 g (1.78 mmol)	Could not be Isolated	-
IV·HCl **D5**, 1.25 g (4.68 mmol)	NTCDA **E1**, 0.60 g (2.23 mmol)	Could not be Isolated	-
VA·HCl **D6**, 0.96 g (4.28 mmol)	PTCDA **E2**, 0.80 g (2.03 mmol)	Could not be Isolated	-

Table 3.31: Reagent proportions and product yields for imide functionalised rylenes prepared by Dr. Rebecca Ginesi

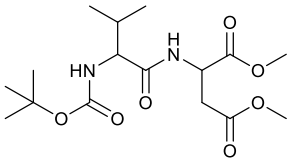
Dipeptide reagent	Rylene reagent	Product	Yield (%)
AI·HCl **D15**, 2.53 g (8.00 mmol)	NTCDA **E1**, 1.00 g (3.72 mmol)	Could not be isolated	-
AL·TFA **D7**, 2.36 g (7.46 mmol)	NTCDA **E1**, 1.00 g (3.73 mmol)	Could not be isolated	-
FA·TFA **D8**, 2.12 g (6.05 mmol)	PTCDA **E2**, 1.01 g (3.77 mmol)	Could not be isolated	-
FL·TFA **D9**, 2.93 g (7.46 mmol)	NTCDA **E1**, 1.01 g (3.77 mmol)	Could not be isolated	-

Table 3.32: Reagent proportions and product yields for imide functionalised rylenes prepared by Dr. Bart Dietrich

Dipeptide reagent	Rylene reagent	Product	Yield (%)
FG.TFA **D13**, 2.33 g (15.39 mmol)	PTCDA **E2**, 2.99 g (7.62 mmol)	Could not be isolated	-
VY·TFA **D14**, 1.49 g (3.78 mmol)	NTCDA **E1**, 0.49 g (1.84 mmol)	Could not be isolated	-
AY.HCl **D16**, 1.54 g (5.33 mmol)	PTCDA **E1**, 1.02 g (2.60 mmol)	Could not be isolated	-
FI.HCl **D17**, 2.23 g (7.08 mmol)	PTCDA **E2**, 1.36 g (3.46 mmol)	Could not be isolated	-
GY.HCl **D18**, 1.46 g (5.31 mmol)	PTCDA **E2**, 1.02 g (2.59 mmol)	Could not be isolated	-
IY.HCl **D19**, 1.10 g (3.33 mmol)	NTCDA **E1**, 0.44 g (1.62 mmol)	Could not be isolated	-
VF.HCl **D20**, 2.07 g (6.88 mmol)	NTCDA **E1**, 0.90 g (3.36 mmol)	Could not be isolated	-
VI·HCL **D21**, 1.81 g (6.79 mmol)	NTCDA **E1**, 0.89 g (3.31 mmol)	NDI-VI **F1**, 0.31 g (0.46 mmol)	14

YG·HCl **D22**, 1.68 g (6.12 mmol)	NTCDA **E1**, 0.80 g (2.98 mmol)	Could not be isolated	-
YI.HCl **D23**, 2.02 g (6.11 mmol)	NTCDA **E1**, 0.82 g (3.04 mmol)	NDI-YI **49**, 1.21 g (1.49 mmol)	49

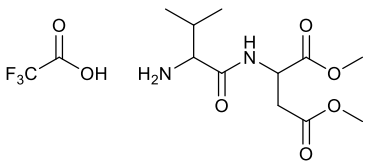
3.4.5 Synthesis of Validation Molecules



G1

To a solution of Boc valine **A6** (3.04 g, 15.4 mmol) in chloroform (50 mL) was added iso-butyl chloroformate (2.00 mL, 15.4 mmol) followed by N-methylmorpholine (3.72 mL, 33.9 mmol). After stirring for 10 min, aspartic acid dimethyl ester hydrochloride (3.41 g, 15.7 mmol) was added and the reaction stirred overnight. It was then diluted with chloroform, washed in turn with aqueous hydrogen chloride (1 M), water, and brine. The layers were separated and the organic layer dried over anhydrous magnesium sulphate, filtered, and evaporated under reduced pressure. The crude oil was purified by column chromatography using an ethyl acetate/dichloromethane mixture (5:95) as eluant to afford **G1** as a colourless oil which solidified on standing (2.70 g, 49%). The presence of rotamers (ca. 85:15) is inferred from the proton NMR.

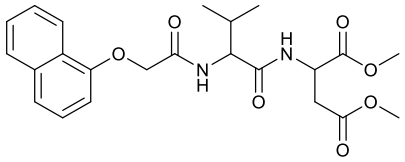
δ_H (400 MHz, DMSO-d_6) 8.30 (1H, d, *J* 7.7, NH), 6.66 (0.85H, d, *J* 8.9, NH), 6.30 (0.15H, d, *J* 7.0, NH), 4.66–4.61 (1H, m, CH-NH), 3.81–3.77 (1H, m, CH-NH), 3.61 (3H, s, O-CH_3), 3.60 (3H, s, O-CH_3), 2.82 (1H, dd, *J* 16.5, 5.9, CH_2), 2.71 (1H, dd, *J* 16.5, 7.3, CH_2), 1.96–1.87 (1H, $CH(CH_3)_2$), 1.38 (9H, s, $C(CH_3)_3$), 0.84 (3H, d, *J* 6.7, $CH(CH_3)_2$), 0.80 (3H, d, *J* 6.7, $CH(CH_3)_2$). δ_C (100 MHz, DMSO-d_6) 171.3, 171.0, 170.3, and 155.3 (C=O), 78.0 ($C(CH_3)_3$), 59.4 (CH-NH), 52.0 and 51.6 (O-CH_3), 48.4 (CH-NH), 35.5 (CH_2), 30.4 ($CH(CH_3)_2$), 28.1 ($C(CH_3)_3$), 19.0 and 17.9 ($CH(CH_3)_2$). HRMS (ESI/QTOF) m/z: $[M+Na]^+$ calculated 383.178872; found: 383.180287.



G2

To a solution of **G1** (647 mg, 1.46 mmol) in tetrahydrofuran (15 mL, heating necessary, solution allowed to cool back to ambient temperature before next step) was added a solution of lithium hydroxide (139 mg, 5.84 mmol) in water (15 mL). The solution was stirred for 5 h, then poured into aqueous hydrochloric acid (1 M, 75 mL) and stirred overnight. The solids were filtered off, washed with water on the filter, and dried on a rotary evaporator to give **G2** as a white solid (548 mg, 90%) with *ca.* 0.4% residual tetrahydrofuran (NMR), which could not be removed after drying in a vacuum oven.

δ_H (400 MHz, DMSO-d_6) 8.90 (1H, d, *J* 7.6, N<u>H</u>), 8.09 (3H, br s, N<u>H</u>$_3^+$), 4.75-4.70 (1H, m, C<u>H</u>*), 3.66-3.63 (1H, m, C<u>H</u>*), 3.64 (3H, s, OC<u>H</u>$_3$), 3.63 (3H, s, OC<u>H</u>$_3$), 2.84 (1H, d, *J* 1.0, C<u>H</u>$_2$), 2.82 (1H, d, *J* 2.1, C<u>H</u>$_2$), 2.14-2.05 (1H, m, C<u>H</u>(CH$_3$)$_2$), 0.96 (3H, d, *J* 7.1, CH(C<u>H</u>$_3$)$_2$), 0.94 (3H, d, *J* 7.1, CH(C<u>H</u>$_3$)$_2$). δ_C (100 MHz, DMSO-d_6) δ: 170.56, 170.27, and 167.99 (<u>C</u>=O), 158.16 (q, *J* 30.5, O<u>C</u>-CF$_3$), 117.24 (q, *J* 300.05, <u>C</u>F$_3$), 57.10 (<u>C</u>H*), 52.26 and 51.77 (O<u>C</u>H$_3$), 48.54 (<u>C</u>H*), 35.21 (<u>C</u>H$_2$), 29.83 (<u>C</u>H(CH$_3$)$_2$), 18.02 and 17.39 (CH(<u>C</u>H$_3$)$_2$). HRMS (ESI/QTOF) m/z: $[M\text{-}TFA+H]^+$ calculated 261.144498; found 261.144725.

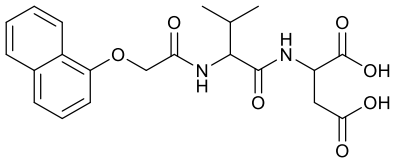


G3

1-Naphthaleneoxyacetic acid (436 mg, 2.16 mmol) was suspended in chloroform (15 mL). *N*-methylmorpholine (522 µL, 4.75 mmol) and *iso*-butyl chloroformate (280 µL, 2.16 mmol) were added and the mixture was stirred for 3 min, upon which all solids dissolved. **G2** (816 mg, 2.18 mmol) was added and the mixture stirred overnight. The resulting mixture was diluted with chloroform, washed in turn with aqueous hydrochloric acid (1 M), water, and brine. The separated organic layer was dried over anhydrous magnesium sulphate, filtered, and evaporated to dryness under reduced pressure. The crude solid was purified over column chromatography using an ethyl acetate/dichloromethane mixture (1:9) as eluent to afford **G3** as an off-white solid (705 mg, 73%).

δ_H (400 MHz, DMSO-d_6) 8.63 (1H, d, *J* 7.7, N<u>H</u>), 8.24–8.19 (1H, m, <u>H</u>$_{Ar}$), 8.00 (1H, d, *J* 9.0, N<u>H</u>), 7.92–7.86 (1H, m, <u>H</u>$_{Ar}$), 7.59–7.48 (3H, m, <u>H</u>$_{Ar}$), 7.40 (1H, t, *J* 8.0, <u>H</u>$_{Ar}$), 6.91 (1H, d, *J* 7.4, <u>H</u>$_{Ar}$), 4.82 (1H, d, *J* 14.6, O-C<u>H</u>$_2$), 4.78 (1H, d, *J* 14.5, O-C<u>H</u>$_2$), 4.69–4.63 (1H, m, C<u>H</u>-NH), 4.34 (1H, dd, *J* 9.0, 6.2, C<u>H</u>-NH), 3.62 (3H, s, O-C<u>H</u>$_3$), 3.60 (3H, s, O-C<u>H</u>$_3$), 2.84 (1H, dd, *J* 16.5, 5.9, CH-C<u>H</u>$_2$), 2.73 (1H, dd, *J* 16.5, 7.5, CH-C<u>H</u>$_2$), 2.09–1.97 (1H, m,

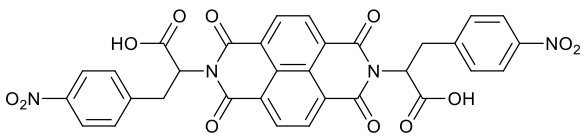
C<u>H</u>(CH$_3$)$_2$), 0.87 (3H, d, *J* 6.8, CH(C<u>H</u>$_3$)$_2$), 0.83 (3H, d, *J* 6.8, CH(C<u>H</u>$_3$)$_2$). δ_C (100 MHz, DMSO-d$_6$) 170.9, 170.6, 170.3, and 167.3 (<u>C</u>=O), 153.1, 134.1, 127.5, 126.6, 126.0, 125.5, 124.8, 121.4, 120.6, and 105.6 (<u>C</u>$_{Ar}$), 66.7 (O-<u>C</u>H$_2$), 56.7 (<u>C</u>H-NH), 52.1 and 51.7 (O-<u>C</u>H$_3$), 48.5 (<u>C</u>H-NH), 35.3 (CH-<u>C</u>H$_2$), 31.1 (<u>C</u>H(CH$_3$)$_2$), 19.0 and 17.6 (CH(<u>C</u>H$_3$)$_2$). HRMS (ESI/QTOF) m/z: $[M+Na]^+$ calculated 467.178872; found: 467.180725.



54

To a stirred solution of **G3** (647 mg, 1.46 mmol) in tetrahydrofuran (15 mL, heating necessary, solution allowed to cool back to ambient temperature before next step) was added a solution of lithium hydroxide (139 mg, 5.84 mmol) in water (15 mL). After 5 h the reaction mixture was poured into aqueous hydrochloric acid (1 M, 75 mL) and stirred overnight. The solids were collected at the filter, washed with water on the filter, and dried on a rotary evaporator to give **54** as a white solid (548 mg, 90%) with *ca.* 0.4% residual tetrahydrofuran (NMR) which could not be removed after drying in a vacuum oven.

δ_H (400 MHz, DMSO-d$_6$) 12.59 (1H, br, COO<u>H</u>), 8.47 (1H, d, *J* 7.8, N<u>H</u>), 8.25-8.19 (1H, m, <u>H</u>$_{Ar}$), 7.98 (1H, d, *J* 9.1, N<u>H</u>), 7.92-7.86 (1H, m, <u>H</u>$_{Ar}$), 7.58-7.49 (3H, m, <u>H</u>$_{Ar}$), 7.40 (1H, t, *J* 8.0, <u>H</u>$_{Ar}$), 6.92 (1H, dd, *J* 7.6, 0.5, <u>H</u>$_{Ar}$), 4.82 (1H, d, *J* 14.5, OC<u>H</u>$_2$), 4.77 (1H, d, *J* 14.5, OC<u>H</u>$_2$), 4.58-4.52 (1H, m, C<u>H</u>*), 4.37 (1H, dd, *J* 9.1, 5.9, C<u>H</u>*), 2.72 (1H, dd, *J* 16.6, 5.7, CH*C<u>H</u>$_2$), 2.59 (1H, dd, *J* 16.6, 7.4, CH*C<u>H</u>$_2$), 2.11-1.97 (1H, m, C<u>H</u>(CH$_3$)$_2$), 0.87 (3H, d, *J* 6.8, CH(C<u>H</u>$_3$)$_2$), 0.82 (3H, d, *J* 6.8, CH(C<u>H</u>$_3$)$_2$). δ_C NMR (101 MHz, DMSO) 172.2, 171.5, 170.4, and 167.2 (<u>C</u>=O), 153.1, 134.1, 127.5, 126.6, 126.0, 125.5, 124.8, 121.4, 120.7, and 105.7 (<u>C</u>$_{Ar}$), 67.0 (O-<u>C</u>H$_2$), 56.6 (<u>C</u>H-NH), 48.7 (<u>C</u>H-NH), 35.8 (CH-<u>C</u>H$_2$), 31.2 (<u>C</u>H(CH$_3$)$_2$), 19.1 and 17.6 (CH(<u>C</u>H$_3$)$_2$). HRMS (ESI/QTOF) m/z: $[M+H]^+$ calculated 417.1656; found: 417.1670.



55

A mixture of NTCDA **E1** (2.79 g, 10.4 mmol), 4-nitrophenylalanine (4.37 g, 20.8 mmol), and imidazole (14.2 g, 208 mmol) was heated to 120 °C in an oil bath under nitrogen and

exclusion of light overnight. The temperature was then reduced to 90 °C, the mixture diluted approx. 1:1 with water and stirred for 1 h. After cooling to ambient temperature, the reaction mixture was filtered, then the filtrate poured into aqueous hydrochloric acid (1 M, 100 mL) and stirred for 2 h. The precipitate was filtered off, returned into aqueous hydrochloric acid (0.25 M), and stirred for 1 h. The solids were filtered again, transferred into water (1 L), and stirred overnight. The solid was collected at the filter and freeze-dried to afford **55** as a light brown solid (5.45 g, 80%).

δ_H (400 MHz, DMSO-d_6) 13.15 (2H, br, COO<u>H</u>), 8.66 (4H, s, <u>H</u>$_{Ar}$), 8.03 (4H, d, *J* 8.7, <u>H</u>$_{Ar}$), 7.52 (4H, d, *J* 8.7, <u>H</u>$_{Ar}$), 5.92 (2H, dd, *J* 8.7, 5.7, C<u>H</u>-NH), 3.76 (2H, dd, *J* 14.1, 5.6, Ph-C<u>H</u>$_2$), 3.44 (2H, dd, *J* 14.1, 8.8, Ph-C<u>H</u>$_2$). δ_C (101 MHz, DMSO) 169.9 and 162.1 (<u>C</u>=O), 146.4, 146.2, 131.3, 130.4, 126.2, 125.7, and 123.3 (<u>C</u>$_{Ar}$), 54.2 (<u>C</u>H-NH), 34.3 (<u>C</u>H_2). (ESI/QTOF) m/z: $[M-H]^-$ calculated 417.1656; found: 417.1670.

3.4.6 Determination of Molar Absorption Coefficient

Up to 10 mg of the molecule of choice was dissolved in 2 mL of dimethyl sulfoxide (DMSO) and briefly vortex mixed using a VWR Lab Dancer. 1 mL of this solution was diluted with 1 mL of DMSO to halve the concentration and this dilution procedure was repeated recursively to generate the dilution series. Absorption spectra were obtained using a Cary 60 UV-visible spectrophotometer (Agilent Technologies). Solutions were measured in a 0.1 mm pathlength (l) quartz cuvette (Hellma Analytics or Starna Scientific). The wavelength of maximum absorption (λ_{max}) between 260 and 800 nm for the highest concentration solution in the calibration curve was determined. The λ_{max} absorption was measured across the dilution series and plotted against the associated concentrations to establish a linear Beer-Lambert relationship. The part of the plot where absorption (A) was proportional to concentration (c), the calibration window, was used to calculate the molar absorption coefficient (ε) (Equation 3.1). For each sample, with the exception of PBI-G **25**, the bounds of the calibration curve were sufficient to quantify the absorption at 1 mg mL^{-1}. The quality of the linear fit was quantified using the coefficient of determination (R^2). The spectra and resulting calibration curve have been provided for each molecule, along with key metrics of the calibration window, molar absorption coefficient and coefficient of determination are summarised in Appendix 3.3.

$$A = \epsilon c l \qquad \text{(Equation 3.1)}$$

3.4.7 Determination of Organic Solubility

Approximately 40 mg portions of each molecule were combined with 1.5 mL of acetone, dimethyl formamide (DMF), dimethyl sulfoxide (DMSO), ethanol, or tetrahydrofuran (THF). Each sample was agitated overnight using an SLS lab basics DigiRoll 6 before filtering through a 0.45 μm PTFE syringe filter (Fisherbrand). The filtered solutions were evaporated to dryness under vacuum and reconstituted in DMSO (with the exception of those already solvated in DMSO). Absorption spectra were obtained using a Cary 60 UV-visible spectrophotometer (Agilent Technologies). Solutions were measured in a 0.01 mm or 0.1 mm pathlength quartz cuvette (Hellma Analytics or Starna Scientific). Solubility classification was determined from molar absorption coefficients and dilution history. Molecules were classified as soluble when solubility exceeded 10 mg mL^{-1}. All measurements in the 0.1 mm cuvettes fell withing the calibration window. Many measurements in the 0.01 mm cuvettes fell outside the calibration window, however, as the absorption of 1 mg mL^{-1} in a 0.1 mm cuvette was within the calibration window, the absorbance of the classification threshold of 10 mg mL^{-1} was known for a 0.01 mm cuvette. Therefore, solubility samples in a 0.01 mm cuvette with an absorbance at λ_{max} exceeding the absorbance afforded by 1 mg mL^{-1} in a 0.1 mm cuvette were classified as soluble, and those with an absorbance below were classified as insoluble. PBI-G **25** was the only exception where the calibration window fell below 1 mg mL^{-1}, with classification measurements falling within the calibration window or below. Therefore, PBI-G could be classified as insoluble for each solvent. The spectra and resulting Beer-Lamber calculations for each molecule and solvent has been summarised in Appendix 3.4, with the resulting binary classifications provided in Appendix 3.5.

3.5 References

1. Z. Hu, F. Huang and Y. Cao, *Small Methods*, 2017, **1**, 1700264.
2. Y. Zhao, L. Liu, F. Zhang, C.-a. Di and D. Zhu, *SmartMat*, 2021, **2**, 426–445.
3. V. Adams, J. Cameron, M. Wallace and E. R. Draper, *Chemistry – A European Journal*, 2020, **26**, 9879–9882.
4. H. Torres-Pierna, D. Ruiz-Molina and C. Roscini, *Materials Horizons*, 2020, **7**, 2749–2759.
5. M. R. Cavallari, L. M. Pastrana, C. D. F. Sosa, A. M. R. Marquina, J. E. E. Izquierdo, F. J. Fonseca, C. A. Amorim, L. G. Paterno and I. Kymissis, *Materials (Basel)*, 2020, **14**.
6. E. Castaldelli, E. R. Triboni and G. J.-F. Demets, *Chemical Communications*, 2011, **47**, 5581–5583.
7. N. R. Murray, DOI: 10.5525/gla.thesis.85387PhD, University of Glasgow, 2025.
8. E. R. Draper, J. J. Walsh, T. O. McDonald, M. A. Zwijnenburg, P. J. Cameron, A. J. Cowan and D. J. Adams, *Journal of Materials Chemistry C*, 2014, **2**, 5570–5575.
9. E. R. Draper and D. J. Adams, *Chem*, 2017, **3**, 390–410.
10. X. Yang, X. Xu and H.-F. Ji, *The Journal of Physical Chemistry B*, 2008, **112**, 7196–7202.
11. E. R. Draper, B. J. Greeves, M. Barrow, R. Schweins, M. A. Zwijnenburg and D. J. Adams, *Chem*, 2017, **2**, 716–731.
12. N. R. Murray, T. J. D. McCabe, M. Reid and E. R. Draper, *Journal of Materials Chemistry C*, 2024, **12**, 12483–12490.
13. Q. Liu, T. Liu and Y. Fang, *Langmuir*, 2020, **36**, 2155–2169.
14. K. Balakrishnan, A. Datar, R. Oitker, H. Chen, J. Zuo and L. Zang, *Journal of the American Chemical Society*, 2005, **127**, 10496–10497.
15. S. A. H. Jansen, G. Vantomme and E. W. Meijer, *Angewandte Chemie International Edition*, 2025, DOI: https://doi.org/10.1002/anie.202509122, e202509122.
16. K. Roy, S. Kar, R. N. Das and SpringerLink, *A primer on QSAR/QSPR modeling: fundamental concepts*, Springer, Cham, 2015.
17. A. A. Toropov, M. Raskova, I. Raska and A. P. Toropova, in *QSPR/QSAR Analysis Using SMILES and Quasi-SMILES*, eds. A. P. Toropova and A. A. Toropov, Springer International Publishing, Cham, 2023, DOI: 10.1007/978-3-031-28401-4_1, pp. 3–24.
18. J. C. Dearden, M. T. D. Cronin and K. L. E. Kaiser, *SAR and QSAR in Environmental Research*, 2009, **20**, 241–266.
19. A. Sun, C. Turner, B. Lin, A. Kabin, S. Kowtha, D. Mohanraj and W. E. Acree Jr, *Physics and Chemistry of Liquids*, DOI: 10.1080/00319104.2025.2523754, 1–23.
20. A. Gissi, O. Tcheremenskaia, C. Bossa, C. L. Battistelli and P. Browne, *Computational Toxicology*, 2024, **31**, 100326.
21. E. R. Draper, L. J. Archibald, M. C. Nolan, R. Schweins, M. A. Zwijnenburg, S. Sproules and D. J. Adams, *Chemistry – A European Journal*, 2018, **24**, 4006–4010.
22. R. I. Randle, L. Cavalcanti, S. Sproules and E. R. Draper, *Materials Advances*, 2022, **3**, 3326–3331.
23. J. J. Walsh, J. R. Lee, E. R. Draper, S. M. King, F. Jäckel, M. A. Zwijnenburg, D. J. Adams and A. J. Cowan, *The Journal of Physical Chemistry C*, 2016, **120**, 18479–18486.
24. T. A. Welsh, J. G. Egan, B. Dietrich, N. Rafferty, R. E. Ginesi, J. Doutch, R. Schweins and E. R. Draper, *Journal of Physics: Materials*, 2024, **7**, 015004.
25. M. C. Sorkun, A. Khetan and S. Er, *Scientific Data*, 2019, **6**, 143.
26. L. Krasnov, S. Mikhaylov, M. Fedorov and S. Sosnin, *ChemRxiv*, 2023, **2023**.

27. L. Krasnov, D. Malikov, M. Kiseleva, S. Tatarin, S. Sosnin and S. Bezzubov, *Scientific Data*, 2025, **12**, 1236.
28. G. v. Rossum and F. L. Drake, *The Python Language Reference Manual*, Network Theory Ltd., 2011.
29. G. Landrum, P. Tosco, B. Kelley, R. Rodriguez, D. Cosgrove, Sriniker, R. Vianello, Gedeck, N. Schneider, G. Jones, E. Kawashima, D. Nealschneider, A. Dalke, B. Cole, M. Swain, S. Turk, A. Savelev, A. Vaucher and M. Wójcikowski, RDKit Release_2024.03.1, DOI: https://doi.org/10.5281/zenodo.10893044).
30. G. Landrum, RDKit: Open-source cheminformatics, https://www.rdkit.org, DOI: 10.5281/zenodo.7828379).
31. D. Huang, Add R-groups to RDKIT Molecules, https://oloren-ai.github.io/blog/add_rgroup.html, (accessed 10 July 2023, 2023).
32. L. Zhao, W. Wang, A. Sedykh and H. Zhu, *ACS Omega*, 2017, **2**, 2805–2812.
33. D. Fourches, E. Muratov and A. Tropsha, *J Chem Inf Model*, 2010, **50**, 1189–1204.
34. A. Tropsha, *Molecular Informatics*, 2010, **29**, 476–488.
35. A. D. Martin, A. B. Robinson, A. F. Mason, J. P. Wojciechowski and P. Thordarson, *Chemical Communications*, 2014, **50**, 15541–15544.
36. A. Mitrović, N. Todorović, A. Žekić, D. Stanković, D. Milić and V. Maslak, *European Journal of Organic Chemistry*, 2013, **2013**, 2188–2193.
37. T. Zhang, Y. Zhou, J. Shan, J. Hou, C. Gu and B. Wu, *Journal of Chemical Sciences*, 2025, **137**, 61.
38. J. H. Moore, D. J. Cole and G. Csanyi, *arXiv preprint arXiv:2405.18171*, 2024.
39. OECD, *OECD Series on Testing and Assessment*, 2023, DOI: 10.1787/d96118f6-en.
40. E. Baka, J. Comer and T.-N. Krisztina, *Journal of pharmaceutical and biomedical analysis*, 2008, **46**, 335–341.
41. M. Vertzoni, J. Alsenz, P. Augustijns, A. Bauer-Brandl, C. A. S. Bergström, J. Brouwers, A. Müllerz, G. Perlovich, C. Saal, K. Sugano and C. Reppas, *European Journal of Pharmaceutical Sciences*, 2022, **168**, 106043.
42. L. Chen, S. Revel, K. Morris, L. C. Serpell and D. J. Adams, *Langmuir*, 2010, **26**, 13466–13471.
43. C. Saal and A. C. Petereit, *European Journal of Pharmaceutical Sciences*, 2012, **47**, 589–595.
44. E. Königsberger, *Journal of Chemical & Engineering Data*, 2019, **64**, 381–385.
45. *Journal*.
46. D. Weininger, *Journal of Chemical Information and Computer Sciences*, 1988, **28**, 31–36.
47. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel and B. Wiswedel, Berlin, Heidelberg, 2008.
48. S. Yousefinejad and B. Hemmateenejad, *Chemometrics and Intelligent Laboratory Systems*, 2015, **149**, 177–204.
49. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Advanced Drug Delivery Reviews*, 2001, **46**, 3–26.
50. H. L. Morgan, *Journal of Chemical Documentation*, 1965, **5**, 107–113.
51. C. Spearman, *International Journal of Epidemiology*, 2010, **39**, 1137–1150.
52. L. Wang, M. Han, X. Li, N. Zhang and H. Cheng, *IEEE Access*, 2021, **9**, 64606–64628.
53. N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, *J. Artif. Intell. Res. (JAIR)*, 2002, **16**, 321–357.
54. S. Matharaarachchi, M. Domaratzki and S. Muthukumarana, *Machine Learning with Applications*, 2024, **18**, 100597.
55. J. C. Platt, *Advances in kernel methods*, 1999, 185–208.

56. S. S. Keerthi, S. K. Shevade, C. Bhattacharyya and K. R. K. Murthy, *Neural Computation*, 2001, **13**, 637–649.
57. L. Breiman, *Classification and regression trees*, Wadsworth International Group, Belmont, Calif, 1984.
58. A. Mucherino, P. J. Papajorgji and P. M. Pardalos, in *Data mining in agriculture*, Springer, 2009, pp. 83–106.
59. T. Chen and C. Guestrin, 2016.
60. V. Nasteski, *HORIZONS.B*, 2017, **4**, 51–62.
61. J. D. Hedengren, Random Forest, https://apmonitor.com/pds/index.php/Main/RandomForest, (accessed 25th March 2026).
62. Random Forest Hyperparameters & Tuning Strategies, https://baekholab.com/2025/03/22/part-6-random-forest-hyperparameters-tuning-strategies/, (accessed 25th March 2026).
63. RBF SVM parameters, https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html, (accessed 26th March 2026).
64. P. Banerjee, A Guide on XGBoost hyperparameters tuning, https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning, (accessed 25th March 2026).
65. F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni and R. Todeschini, *Molecules*, 2012, **17**, 4791–4810.
66. M. Shen, A. LeTiran, Y. Xiao, A. Golbraikh, H. Kohn and A. Tropsha, *Journal of Medicinal Chemistry*, 2002, **45**, 2811–2823.
67. Reaxys®, https://www.reaxys.com/, (accessed 03 June 2025).
68. S. L. Brunton, J. N. Kutz and P. Cambridge University, *Data-driven science and engineering: machine learning, dynamical systems, and control*, Cambridge University Press, Cambridge, United Kingdom, 2019.
69. S. M. Lundberg and S.-I. Lee, *Advances in neural information processing systems*, 2017, **30**.
70. Y. Kwon, M. A. Rivas and J. Zou, presented in part at the International Conference on Artificial Intelligence and Statistics, 2021.

Chapter 4: Predictive identification of self-assembled aggregates in naphthalene and rylene solutions

This Chapter is adapted, in part, from the following publication:

"Applications of microscopy and small angle scattering techniques for the characterisation of supramolecular gels"

C. R. M. MacDonald and E. R. Draper

CRMM was responsible for conducting the research, creating figures, and writing the publication. CRMM wrote the initial draft of the manuscript, to which all authors contributed for the final publication.

4.1 Introduction

In Chapter 3, we focused on developing solubility models to identify solute-solvent combinations that provide sufficient solubility to enable processing of naphthalenes and pyrenes for device preparation. While soluble materials can now be predicted and identified, it is assumed that soluble materials afford uniform molecular dispersions.[1] However, previous work in the group has highlighted that these soluble dispersions of naphthalenes and rylenes are actually self-assembled aggregates in solution, and more that, under different conditions they form different aggregates with differing properties. For example, characterisation by UV-Vis absorption can probe molecular-scale assembly, showing that while the same compound can be dissolved in two different solvents, significant changes in the UV-Vis absorption profile can be observed, arising from changes in monomer aggregation. Solutions of such materials have been characterised using various techniques to show the presence of micellar structures, and films formed by drying such solutions have been shown to form fibrous networks that afford responsive functional properties (Figure 4.1).[2-4] Therefore, while predicting solubility is key to identifying solvent-solution combinations sufficiently soluble to enable molecular rearrangement, it alone is insufficient to identify combinations that may subsequently aggregate and yield such functional materials.

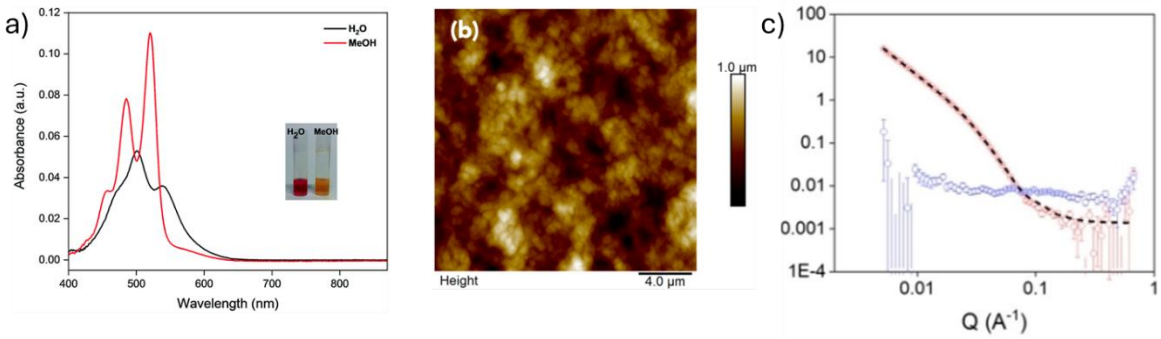


Figure 4.1. a) UV-Vis spectroscopy of solutions showing differences in molar absorptivity and in the degree of aggregation as a function of solvent. Inset: Equimolar (10^{-4} M) soaking solutions of PBI-A. Reproduced from Lundy *et al.*[2] b) AFM images of films of PBI-A on treated plastic showing bunched, thin fibrous structures upon drying. Reproduced from Adams *et al.*[3] c) Small angle neutron scattering from PBI-A solutions at pD 6 (red) and pD 9 (blue). Reproduced from Ginesi *et al.*[4]

As previously discussed, this self-assembly across length scales allows for the control of properties through not only chemical modifications of the self-assembling molecule but also changes to the preparation process.[5] While this is highly advantageous for the development and tuning of new materials, it makes predicting material properties extremely challenging. Currently, there are few ways to predict how self-assembly processes change as the

environment and preparation process are altered, despite these changes having a significant effect on the resulting material. Thomson *et al.* investigated the effects of varying the concentration of 2Nap-FF (**1**) solutions and their impact on the nanostructural characterisation. It was shown that measuring these changes could be used to probe bulk behaviours and identify materials that can be processed into threads, films, and noodles, each with unique properties.[10] This shows the direct link between micellar changes tuned by the preparation process and the bulk properties of interest. Similarly, Naranjo *et al*. have demonstrated solvent-dependent changes in the aggregation of *N*-annulated perylene bisimides. In solvents with a greater solvating ability for the compounds investigated, such as chloroform, the PBI core will favour the formation of dimers. However, switching to apolar methylcyclohexane induces anti-cooperative supramolecular polymerisation, in which dimers stack to form larger oligomers, demonstrating the influence of solvent–solute interactions. These changes to the solution-state assembly significantly affect the resulting materials when cast into films. Analysing by atomic force microscopy shows that chloroform affords small nanoparticles; however, the extended oligomerisation in methylcyclohexane affords longer rope-like fibres (Figure 4.2).[6]

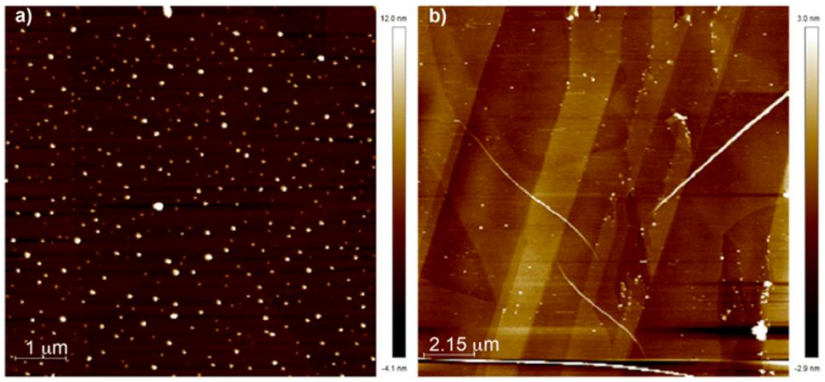


Figure 4.2. Height AFM images of the nanoparticles and isolated fibres formed upon self-assembly in a) chloroform and b) methylcyclohexane, respectively. Reproduced from Naranjo *et al*.[6]

As well as changes to the molecular structure and preparation, properties may also be modulated by combining the properties of different materials. Multicomponent systems allow for enhanced control over supramolecular structures and provide a straightforward way to introduce new properties and functions. By combining two components that exhibit different properties, it is possible to enhance existing features or develop entirely new ones.[7] These systems can form various structures: they may self-sort into assemblies made of a single component or co-assemble, creating structures that contain a mixture of both components. These co-assemblies can be organised into ordered social arrangements, like donor-acceptor systems, or arranged randomly.[8] Each configuration offers opportunities to

develop novel systems that cannot be achieved with a single component alone. (Figure 4.3).[7, 8] While multicomponent systems can enable significant changes in the resulting materials, controlling supramolecular co-assembly is not trivial. The degree of interaction between multicomponent systems is governed by multiple interactions. The main two factors are the formation kinetics and the interfiber interaction. Firstly, it has been shown that the order of component formation can substantially change the resulting structure, with the kinetics of network formation for each component determining the resulting network patterns. If interactions between the components are negligible, they can diffuse and self-assemble independently in the void space of the network, resulting in an orthogonal network. If the LMW components interact with each other, the supramolecular fibre formation preferentially occurs to afford the interactive network[9] However, the strong directional interactions are generally exhausted in the formation of the main chain of the supramolecular system, only weak non-directional interactions remain to contribute to the higher-order organisation[10]

Many tools exist to offer insight into such a process, but complete control and prediction of the assembly mechanism remains difficult.[11, 12] In addition to the process-dependent complexity afforded by supramolecular systems generally, the preparation and combination method of the individual components can significantly change the structures and properties of the resulting multicomponent materials.[13] Furthermore, without distinct identifiable differences between individual components, it becomes difficult to characterise how combining components does or does not change the structures measured.[12] A review by Li *et al*. summarised numerous examples of multicomponent self-assembling systems, highlighting cases in which none, some, or all components can self-assemble individually. Across these examples, emergent properties included the controlling of chirality, enhanced mechanical strength, and optimised stimulus response, amongst other features.[14] Indeed, it has been shown that preparing multicomponent systems of naphthalenes and aromatics induces changes in molecular packing and aggregation in solution, as well as in the resulting materials.[13] By enabling prediction of the structural changes afforded by multicomponent tuning, the number of useful candidates for application increases dramatically, potentially offering greater functionality than any single-component system can provide. With appropriate characterisation libraries for multicomponent systems, QSPRs may provide another approach to understanding and predicting their complex assembly.

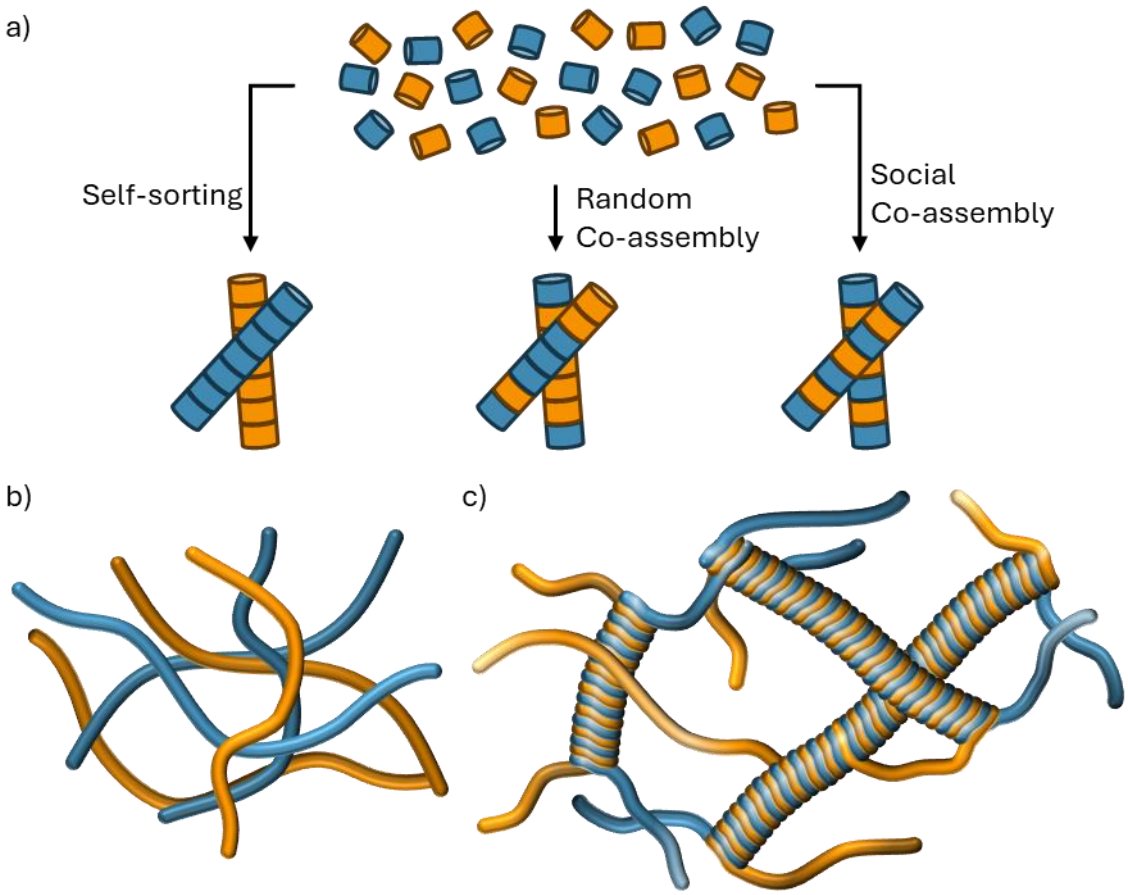


Figure 4.3. (a) Cartoon depicting varying modes of self-sorting and co-assembly. Primary fibres can further assemble with (b) few interactions or (c) intimately self-assemble, for example.

To fully utilise the preparation dependence of self-assembling systems and the materials they can afford, it is crucial that we develop tools capable of predicting how the desired properties can be tuned *via* such process control. However, such process-dependent changes show that it is not possible to use molecule structure alone to predict the properties of supramolecular aggregates. Given the necessity of tuning aggregate morphology in solution to afford desirable properties in bulk material, it may be possible to use morphology to make a useful predictive system of material properties. For example, we have previously highlighted the work by Adams *et al*., which showed that the structural functionalisation of amino acid-derivatised PBIs led to changes in the self-assembled aggregates. When cast into films, the longer micelles afforded by alanine functionalisation enabled the preparation of mechanoresponsive films. Building relationships between descriptors of these structural descriptors and the desirable properties is precisely the type of problem QSPRs aim to solve. With sufficient high-quality information about molecular structures, the morphologies that can be accessed for each, the impact of preparation changes, and properties of interest, a predictive system that probes properties across the hierarchical scales of assembly may be achieved. Subsequently, building systems to obtain and predict greater insights into the

nanostructural and microstructural assembly could enable more complex property prediction in future models. However, to measure such properties, robust characterisation techniques and protocols are required.

4.1.1 Characterising Self-assembly

A comprehensive review by Weiss and Térech discussed in detail the plethora of chemical systems, characterisation techniques, and applications of molecular gels, which remains deeply insightful despite the rapid advancement of the field.[15] A more recent review by Yu *et al*. summarised several characterisation techniques used to study supramolecular materials more generally.[16] Two of the most commonly used methods for characterising the larger structures present are microscopy and scattering techniques.[17]

4.1.2 Microscopy Techniques

Imaging is the most widely used and accessible technique for characterising these gel materials. Researchers are looking for fibrous structures, coils, entanglements, spherulites, and other such features that can provide insight into how the gel forms within the network through the assembly of interlocking nano- and microstructures. Imaging is often inexpensive and accessible for both data acquisition and interpretation. These images are attractive and give an instant visual summary of the network. Depending on the imaging technique used, multiple length scales can be probed to identify how individual fibres form and then entangle to form the supramolecular networks. In addition to showing the shape of the self-assembled structures, software tools such as ImageJ can be used to quantify features such as fibre width, fibre length, and pore sizes.[18, 19] However, measurements are limited to the snapshot captured in the image and require manual measurement of the fibres. This can be error-prone, inconsistent, and time-consuming, requiring significant expertise to perform and limiting the number of samples that can be feasibly analysed.[20] Our recent review discusses the most popular microscopy techniques for characterising supramolecular gels, highlighting both the significant advantages and limitations of microscopy in the context of characterising supramolecular systems.[21]

Among microscopy techniques, electron microscopy (EM) is among the most popular and informative for investigating supramolecular systems. Scanning electron microscopy (SEM) and transmission electron microscopy (TEM) together can enable the multiscale quantification of nanoscale and microstructural features.[22] However, the sample preparation required by SEM and TEM can lead to significant changes in the materials' morphology. EM

techniques commonly, but don't always, require the addition of ultrathin coatings, which alter the material's surface. Furthermore, TEM is limited to the investigation of ultrathin films, which are typically achieved via mechanical abrasion or blotting.[19] Supramolecular materials, however, are particularly sensitive to shear forces and will be particularly sensitive to morphological changes from the stresses of abrasive thinning. One of the most important considerations is that EM techniques may require a fundamental change in the material under investigation, as their high-vacuum environment necessitates drying samples. This is problematic for gels in particular, as gels are mostly solvent by composition, it is to be expected that the drying procedures required may lead to a change in morphology due to effectively concentrating the sample, leading to more aggregation, any salts present in the solution could also crystallise and disrupt the network, or change the stability and solubility of any charged structures present. While supramolecular films can be prepared by removing the solvent from solutions of self-assembling aggregates, this is often done under ambient conditions, where some solvent may remain after drying.[23] Subsequently, EM drying techniques may affect the characterisation even for these dried films. Indeed, it has been shown that the characterisation of a dipeptide derivatised naphthalene xerogel will change significantly when comparing air-dried and freeze-dried analogues.[24] It has also been shown that the choice of fixation technique required for the cryogenic variants of these techniques, meant to reduce the drying artefacts, can impact morphology significantly. Aston *et al.* demonstrated that changing cryo-fixation alone can alter macropore and micropore sizes by an order of magnitude.[19] All of these issues could result in statistics which are not representative of the structures *in situ*.[25] Recently, Katrantzi *et al*. have shown that the preparation route prior to vitrification can also significantly impact the microstructure of hydrogels. By altering the preparation protocol, sample thickness can vary, affecting the ice crystal formation and the resulting structure. The subsequent imaging may not be of the native network, but rather of the boundaries of ice crystal nucleation and growth.[26] Therefore, techniques capable of *in situ* characterisation are required to build a robust QSPR database from which we can probe and predict properties of interest. While the robust characterisation protocols described by Katrantzi *et al*. suggest an effective route for the characterisation of supramolecular materials using imaging techniques, higher throughput *in situ* techniques may be necessary for QSPR dataset development with sufficient scale.[26]

4.1.3 Scattering Techniques

Small angle scattering techniques have been used to investigate soft matter systems for decades, with pioneering work by Térech and coworkers showed how these techniques can

probe both local and long-range structures in 3D networks.[27] In the decades since there has been a wealth of reviews on soft matter characterisation as interest in such systems increases, and techniques develop at a rapid pace.[15, 16, 28, 29] Small angle X-ray scattering (SAXS) is achieved by firing a monochromatic beam of radiation through a sample, with the majority of the radiation being transmitted (and blocked by a beamstop) without interacting with any structures in the sample. A small fraction of the radiation will interact with the electron density of the sample's structures, resulting in Thomson scattering that can be detected. The resulting scattering is characterised by the scattering vector (q), which is the result of a photon of wavelength (λ) scattering off the sample at an angle of 2θ (Equation 4.1 and Figure 4.4).

$$q = \frac{4\pi \, sin\theta}{\lambda} \qquad \text{(Equation 4.1)}$$

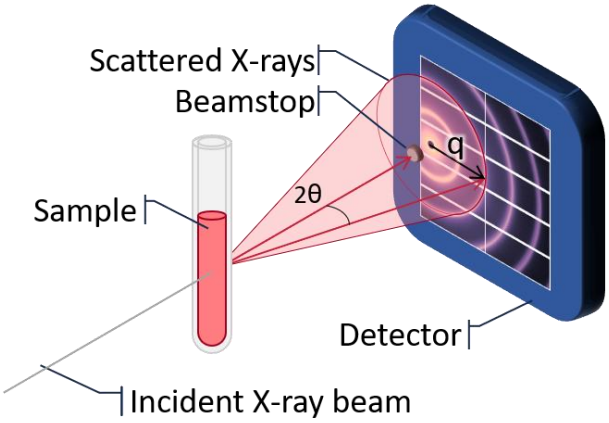


Figure 4.4. Cartoon of a general small angle X-ray scattering setup. Reproduced from MacDonald *et al*.[30]

Typically, a 2D detector measures the scattered radiation to determine the scattering intensity (I(q)), which is then plotted as a function of q. The scattering vector has units of reciprocal space, so the smaller low-q measurements represent larger real-space distances (d) of repeating structures (Equation 4.2).

$$d = \frac{2\pi}{q} \qquad \text{(Equation 4.2)}$$

The scattering pattern characterises the electron density throughout the sample, which is directly affected by the arrangements of molecules and atoms throughout the structure. Therefore, the observed scattering carries bulk-averaged information, a statistical average of the structures across the macroscopic irradiated volume.[15] Depending on the q-range measured, different structural scales can be investigated. As q increases, smaller structures are probed, allowing the topology to be determined.

Many models exist that describe the expected scattering arising from a wide range of structural topologies. This fitting allows multiple parameters to be determined. For example, self-assembled materials form worm-like micelles that various cylindrical model parameters, including structure length, Kuhn length, structure radius, and axis ratio can describe.[31] Scattering is representative of the overall system in comparison to the snapshots of individual micelles observed in microscopy techniques.[24] To achieve this, the fitted model must capture several parameters that describe the scattering intensity (I(q)). The intensity is governed not only by the squared difference in scattering length density ($\Delta\rho^2$) between the particle and solvent, which can be calculated using online resources.[32] The number of scattering particles (c) and the squared volume of a single scattering particle also contribute. The squared contribution of particle volume highlights the ability for larger particles to exhibit greater scattering, making scattered intensity proportional to the volume of the largest scattering object in solution. The topology and distribution of the particles also affects the scattering significantly, with contributions from the form factor (P(q)) which describes the shape and size of the individual particles, and the structure factor (S(q)) which describes the spatial arrangement and interactions between particles (Equation 4.3).[33]

$$I(q) = c\,\Delta\rho^2\,V^2\,P(q)\,S(q) \qquad \text{(Equation 4.3)}$$

One of the foremost benefits of SAS is the capability to perform *in situ* measurements, wherein samples can be analysed with minimal change to the materials environment. This allows for structures that are more representative of their typical environment. Sample environments can also be as simple as a capillary at ambient conditions. Without a need for the sample environment to be under vacuum, as is required for some microscopy techniques, SAS allows for volatile solvent systems to be characterised. Therefore, by providing a better understanding of how self-assembled systems are structured, SAXS can be a more accurate tool for identifying structural characteristics that govern the material properties of interest.

4.1.4 Literature Data

As discussed previously in Section 2.1, the quality of training data is dependent entirely on the reliability of its sources.[34] As a result, data curation is among the most important steps in QSPR development.[35] A fundamental assumption in cheminformatics studies is that chemical descriptors are calculated to represent underlying chemical structures accurately.[34] It has been demonstrated that a small number of erroneous structures represented by erroneous descriptors will have a significant detrimental effect on the predictivity of the QSPR model.[36, 37]

Although cheminformatics databases exist for many chemical properties, the application of cheminformatics in the field of materials science has historically received limited attention, and even less so for supramolecular materials chemistry in particular. As such, there are no well-curated databases focused on the supramolecular properties of small organic molecules suitable for QSPR modelling. Therefore, to build QSPR models, either data mining or a direct measurement of the properties of interest is required.

Data mining involves the investigation of primary literature sources to identify properties of interest within previously measured systems. As the interest in supramolecular self-assembling compounds increases, so too does the quantity of related scientific literature. The growing volume of literature makes it increasingly difficult to find and compile meaningful datasets through manual curation, as has been the case with analogous problems in drug discovery and QSPR model development.[38] With the majority of chemical data being reported in unstructured natural language, the advent of increasingly powerful large language models (LLMs) represents a potential solution for efficient automated data extraction that could enable QSPR dataset curation.[39] However, both manual and automated literature mining require comprehensive data documentation to mine from. Data reporting procedures are lacking for techniques commonly used to characterise supramolecular materials.[21] Furthermore, a problem across many research fields is that the reporting of research findings is often biased, with the preferential reporting of positive results.[40] The same bias extends to the dissemination of such results, with citation biases potentially leading to an overrepresentation of positive results.[41] To make a QSPR model capable of identifying the positive cases of a property of interest, it must also be able to learn what chemical structures do not exhibit those same properties.

Where data is reported, the standards for reporting scattering profiles and fitted model parameters vary significantly, making the extraction of necessary information difficult and often impossible. For example, the work by Colquhoun *et al*., like many examples in the literature, reported their scattering data with very limited information.[42] The paper itself details the fitting parameters of the accepted model, with a supplemental figure simply overlaying the scattering plot with the model fit. While this suggests that cylindrical models appear suitable, it does not adequately explain why or even whether this is the most acceptable model. In contrast, McDowall *et al*. comprehensively describe their acceptance criteria by highlighting a few decision criteria and including more detailed fitting parameters.[31] Firstly, the types of models investigated were described and rationalised. Secondly, the complexity of the models used is limited to minimise the risk of overfitting.

Thirdly, the selection criteria for choosing the best model are defined. Finally, in addition to visually comparing the scattering data and the modelled fit, the parameters and accepted metrics are tabulated, including the parameter fit errors, which indicate how well the parameters represent the observed scattering, to provide a complete description of the model.[31]

These limitations make the mining of primary sources for adequate information problematic, with the resulting dataset of supramolecular properties likely to be biased, inaccurate, and incomplete. The current limitations in literature mining and characterisation reporting for supramolecular materials constrain the potential applications of QSPR modelling. With sufficient data, QSPR systems could address the current lack of design principles for many promising classes of supramolecular compounds under investigation across the field. Such limitations necessitated the preparation of a bespoke dataset for QSPR modelling ourselves, ensuring data quality through the application of standardised data collection principles.

4.1.5 Characterisation Techniques

As discussed, there are many techniques available for the characterisation of supramolecular materials, the most common of which are microscopy and scattering techniques. Some of the key benefits and drawbacks of microscopy and scattering techniques are summarised in Table 4.1.

Table 4.1. Comparison of microscopy and scattering techniques with advantageous traits highlighted in bold. Table 4.1 was reproduced from C. R. M. MacDonald *et al*.[21]

	Microscopy	Scattering
Local Structure and Detail	**Easily detected**	Details lost
Average Structure	Difficult to obtain	**Always obtained**
Preparation Artefacts	Difficult to prevent	**None**: Measurement in-situ
Resulting Data	**Direct pictures** with poor statistics	Difficult to interpret, **good statistics**

Table 4.1 highlights the complementary nature of these techniques and how, when used together, they can overcome the shortfalls each exhibits on its own. It has previously been shown that combining such complementary techniques is key to obtaining a complete picture of supramolecular systems.[21] The properties of supramolecular materials are controlled by hierarchical structures which assemble across multiple length scales. Consequently, characterisation is required across multiple length scales to understand these systems fully. The scale disparity between the nanoscale fibre diameter and the microscale structure spans multiple orders of magnitude. No single microscopy or scattering technique can accurately

characterise both scales, and so a combination of techniques is optimal.[17] One of the limitations of QSPR modelling is the high volume of data required to begin finding meaningful relationships between the structure and property. Subsequently, building adequate datasets demands techniques that are both high-quality and high-throughput. To ensure that consistent property classifications are applied to each compound from the characterisation data, a consistent measurement protocol is required. Given the scale of data ideally required to build robust models, it was determined that gathering the highest-quality data with a single technique would be the optimal choice for our predictive system. While the number of properties that such data can adequately characterise will be limited compared to a tandem approach, narrowing the scope of the QSPR to ensure it is making robust predictions on a smaller subset of characteristics will show if QSPR models can be used to model such classification problems for supramolecular materials with sufficiently high-quality data. Conversely, if we focus on gathering more diverse data from fewer data points, it becomes increasingly difficult to build a robust model and to determine whether the poor models that may result arise from too few data points or from the data's poor amenability to classification modelling. As previously discussed, while microscopy techniques are advantageous for their simple, interpretable outputs, their reliance on drying and vitrification methods can lead to significant artefacts in preparation, especially in soft supramolecular materials. The *in situ* nature of scattering measurements allows the characterisation of representative systems, focusing on the global average structure and providing more representative statistics of the observed assemblies. Given the key advantages of scattering and the drawbacks of microscopy, it was clear that scattering would be the technique of choice for characterising our supramolecular systems to build a QSPR dataset. We note that with the appropriate reporting of our data and protocols, future iterations of our QSPR system could be expanded in future investigations by applying the same protocol to additional compounds, or the potential properties capable of modelling can be broadened by using our scattering data in tandem with microscopy techniques to yield QSPR models with further functionality.

4.1.6 Small Angle X-ray Scattering

The characterisation of supramolecular materials by scattering techniques requires access to highly specialised equipment. A synchrotron SAXS instrument offers several advantages critical to the development of QSPR datasets compared with laboratory SAXS and synchrotron SANS instruments. The required exposure time for both alternative techniques would be at least an order of magnitude greater. As we have no laboratory SAXS

instrumentation, obtaining prolonged access to such equipment would necessitate reliance on data from multiple sources, leading to greater data inhomogeneity and hindering the development of robust QSPR models.[43] Furthermore, compared to synchrotron sources, a laboratory SAXS instrument suffers from a more limited q range. As q-space and real-space measurements are inversely proportional, a reduction in q range will limit the size of structures that can be accurately quantified. One of the major limitations of scattering techniques is the difficulty in accessing synchrotron facilities. Several synchrotron facilities offer SAXS services, including those at ANL (USA), DESY (Germany), Diamond Light Source (UK), ESRF (France), Spring-8 (Japan), and SSRL (USA).[44] Many beamlines accept proposals from researchers worldwide, granting beamtime on the scientific merit of proposals as determined by a peer-reviewed panel of researchers. The available time at such facilities is limited and in high demand, making access difficult, even for established academic researchers.[45] However, the demand for tools that can facilitate the rational design of supramolecular materials gave us synchrotron access to Diamond Light Source's I22 beamline (experiment SM40119). In Section 4.1 we also highlighted the extra complexity multicomponent systems can afford to make more highly-tuned materials. However, this also gives rise to greater complexity in material characterisation. A significant limitation of SAXS in the context of multicomponent systems is the reliance on electron density to provide structural contrast, which for organic molecules such as naphthalenes and pyrenes varies very little. This makes the masking of each component impossible, and the characterisation of the structures afforded by each component individually unachievable. While the high-throughput SAXS data can afford some insight into the aggregates formed, more information may be obtained from complementary techniques such as small angle neutron scattering, which may afford more insight through the use of selective deuteration and contrast matching to investigate individual components for future datasets.[29]

4.1.7 SAXS Model Complexity

Despite the high-quality data afforded by a synchrotron SAXS scattering setup, the fitting procedure (even when aided with ample *a priori* knowledge) is a non-trivial task, requiring reasonable expertise and investigation to ensure that the appropriate models are used, potentially requiring days of trial and error to find and optimise the appropriate models for only a handful of samples. For example, if we consider the naphthalenes and perylenes, there is a wealth of literature establishing cylindrical models as the appropriate topology to investigate, along with spherical models and power laws, which indicate the presence of small micellar aggregates and larger-order structures, respectively. If we consider the typical

cylindrical models applied to these systems, there are five that often provide informative fits (cylinder, elliptical cylinder, flexible cylinder, flexible elliptical cylinder, and hollow cylinder). If we consider that, for many, a single model may not be sufficient and either requires some application of polydispersity or a combination with a sphere or power-law model, this may mean that each sample may have 20 different models that warrant investigation and optimisation. Such efforts are of interest for future models, as fitting an extensive library of self-assembling solutions could afford predictive models capable of identifying candidates with ideal topologies. For example, as discussed previously, the longer fibre lengths afforded by PBI-A, compared to other amino acid-functionalised perylene bisimides, are thought to be a significant factor in the emergence of the unique mechanoresponsiveness of their resulting films.[3] By developing models that could predict targets with a propensity to form longer structures, new target materials could be designed. While tools such as SasView have been developed and continuously improved to simplify fitting topological models to experimental scattering patterns, there remains a need for sufficient expertise to ensure that appropriate conclusions are drawn from the fitting procedure.[46] Such limitations are among the major drawbacks of small angle scattering techniques, leading to much of the data from high-throughput experimentation going unused.[47] The manual intervention required to develop an adequate dataset was deemed too great. As an initial step to realising such systems, it was decided that it was necessary first to develop a simpler system. We highlight, however, the work by Roberts *et al*., which offers a solution to this significant limitation. Using their own machine learning framework, they have developed a system that processes scattering curves to identify appropriate nanoparticle morphologies and corresponding parameters, with accuracy comparable to that of human experts.[47] Even in such a case where the optimal model may not be found, structural parameters can be further optimised using conventional fitting algorithms, which converge quickly when initialised from good estimates predicted by ML. Using this methodology, thousands of scattering curves can be feasibly processed.[47] Such systems will be vital to the future development of our QSPR system. Of particular note is this system's ability to classify cylindrical and spherical systems, making it an ideal tool for handling the large volume of scattering data obtained in this project.

Therefore, given the complexity and time-consuming nature of competently fitting scattering data, a targeted analysis of the more informative samples is required. Furthermore, given the inaccessibility of scattering techniques, being able to predict those combinations worth measuring would allow for more useful dataset for the future development of increasingly

complex predictive systems. While we can now predict the solubility of amino acid functionalised rylene derivatives across organic solvents, this provides limited information on aggregation that occurs in solution. Subsequently, by characterising the scattering of solutions of the soluble molecules identified in Chapter 3, a QSPR dataset can be developed to probe and predict the aggregation in solution (Figure 4.5). Using these QSPR systems in tandem could be an effective approach to narrowing the scope of investigation of useful material candidates for future material and QSPR dataset development.

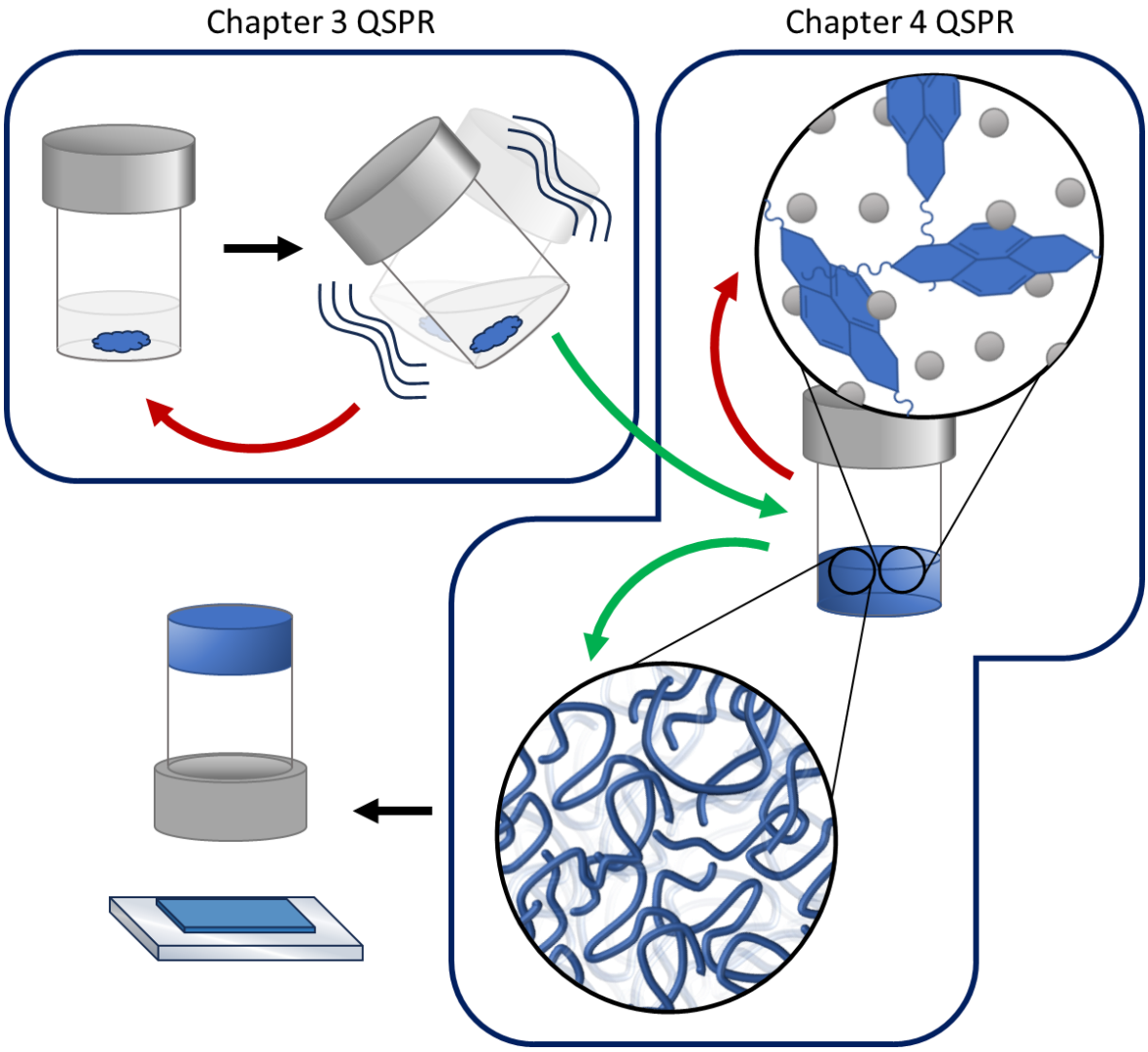


Figure 4.5. Cartoon summary of Chapter 3 and Chapter 4 QSPRs. While the earlier QSPR can identify candidates capable of solubilising in select solvents, an additional predictive system is required to identify which candidate combinations are molecularly dissolved or form micellar aggregates. By identifying those that form aggregates, the scope of possible candidates can be narrowed to those suitable for designing functional materials.

4.2 Results and Discussion

4.2.1 Data Collection

Before collecting SAXS data to be used for the models, we first needed to determine how the data would be collected, this would be ideally carried out using a well characterised and studied system as an example 'model' dataset. The extensive investigation of 2Nap-FF (**1**) previously undertaken in the group makes it an ideal candidate for assessing the quality of the resulting SAXS data.[48-50] 2Nap-FF self-assembles at high pH into long, anisotropic structures that entangle to form viscous solutions.[48] These micelles aggregate into well-defined one-dimensional cylinders which have previously been characterised using SAXS and cryo-EM to elucidate the nanostructures arising from self-assembly (Figure 4.6 and Table 4.2).[48-50] By quantifying and modelling the topological parameters obtained from the scattering setup described in Section 4.4.3, we find that the resulting data afforded models comparable with prior analyses, indicating that the scattering data (when fit appropriately) is a high-quality representation of the nanostructural characteristics. We could therefore ensure the quality of future data collection to this system to ensure the high-quality data collection throughout the study.

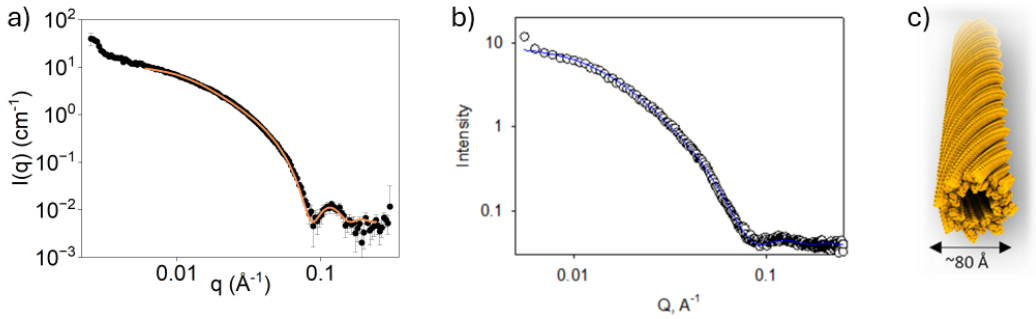


Figure 4.6. SAXS data and fits. a) shows the data (filled circles) and fit (line) for a solution of 2Nap-FF (**1**) at a concentration of 10 mg/mL measured on Diamond's I22 beamline. b) shows the data (hollow circles) and fit (line) for a solution of 2Nap-FF (**1**) at a concentration of 10 mg/mL previously measured using a SAXSLAB Ganesha 300XL instrument.[48] c) shows the cryo-TEM density map of 2Nap-FF (**1**) and the dimensions from the resulting computational model.[50]

Table 4.2. The model fit parameters generated for 2Nap-FF (**1**) by fitting models to the SAXS data in SasView.

	This Project	Previous Analysis [48]
Model	Flexible Cylinder	Flexible Cylinder
Background (cm-1)	0.0055 ± 0.00040	0.0398 ± 0.00013
Kuhn Length (Å)	95.5 ± 0.0	63.0 ± 2.4
Length (Å)	955.1 ± 10.2	949.6 ± 5.6
Radius (Å)	43.4 ± 0.1	40.5 ± 0.2
Scale	$0.00090 \pm 9.0\times10^{-6}$	$0.0059 \pm 1.1\times10^{-5}$
$\chi2$	7.4906	4.1614

It is important to note that small changes to the measured system, which may go unnoticed when reviewing literature, can significantly impact self-assembly. For example, while the scattering data obtained by Draper *et al*. for 2Nap-FF (**1**) have also been previously fit to simpler cylinder models, these analogous samples were instead prepared in D_2O.[49] This use of deuterated solvents for characterisation is widely accepted as not affecting topology and is used throughout the literature. Whilst intermolecular electrostatic forces are known to be the predominant force governing aggregation, many changes to the system can be made to afford different assembly topologies.[51] The solvent isotope effect has received little investigation, which, if significant, makes the use of characterisations obtained from deuterated solvents to explain observations in their hydrogenated analogues problematic.[52, 53] While it was hoped that the fitted model parameters obtained from prior scattering data, predominantly obtained from deuterated analogues, could be used in future models to augment the data obtained in this project for future QSPR model development, the impact of solvent deuteration demonstrated by differing topological fits of 2Nap-FF across solvent isotopes identifies the need to account for such changes when building an appropriate model. Therefore, it was deemed necessary to treat solvent isotopes as independent solvents, requiring descriptors that could appropriately identify the impact of the solvent deuteration on the prediction classification, if necessary.

When building topological datasets, it is important to assess that the reported metrics have been derived from appropriate characterisations. We highlight the hierarchical nature of supramolecular self-assembly across significantly different length scales and the need for a tandem characterisation approach to obtain a complete image.[30] The robustness of characterisation afforded by the scattering technique is governed by the range in reciprocal space that is quantified accurately. Our characterisations afforded accurate data, typically down to $q = 0.003$ Å^{-1}, corresponding to real-space distances of around 2000 Å^{-1}. Self-assembling micelles can afford multi-micron systems which are beyond the limit of quantification of the characterisations achieved here. Therefore, to ensure that the QSPR data is built on the most accurate data, it is necessary to dismiss characterisations beyond these limits. This allows experimental endpoints to be modelled from purely reliable descriptors and highlights the danger of using literature data without proper review, as such limits of quantification may not be addressed in reporting of scattering models. While lacking such information can limit the insights obtainable about each system, learning the fundamental nature of assembly in each solvent-solute combination and the resulting micelle cross-sectional parameters can still be highly informative. For example, if it is possible to

identify the presence or lack of aggregates, one can narrow down the candidates worth further investigation and topological fitting. A more focused investigation of those that scatter could better identify those which have ideal topologies. For example, identifying those with cylindrical topologies with lengths beyond the limit of characterisation could further narrow the scope of viable candidates for making functional materials that rely on the entanglement of such long one-dimensional fibres. Not only would it aid in identifying useful, applicable candidates, but it would also allow the development of QSPR datasets that can probe and predict more complex aspects of the self-assembly process.

With a robust characterisation regime and accurate, useful endpoints that can be characterised, a dataset suited to QSPR modelling could be obtained. As discussed, for self-assembly to occur, they must dissolve sufficiently to allow molecular mobility. Therefore, a library of solute-solvent combinations, focused on those expected to be sufficiently soluble to enable assembly, was prepared using the materials described in Appendix 4.1. Given that self-assembly is governed by the balance of interactions between solute molecules and their affinity for molecular dissolution in each solvent, a range of solvents varying in polarity was selected to assess and predict the formation of self-assembled aggregates as solvent interactions are varied. To that end, acetone, dimethyl formamide (DMF), dimethyl sulfoxide (DMSO), ethanol, tetrahydrofuran (THF), and water were investigated, chosen to cover a range of polarities and to utilise the insights afforded by the solubility characterisation data obtained in Chapter 3 to better narrow down the investigation of useful candidate systems. The complexity and functionality of materials resulting from the self-assembly afforded by these materials can be modulated by combining them to afford multicomponent systems as described in Section 4.1.7. While each organic solution was prepared at 5 $mgmL^{-1}$, a concentration commonly investigated for these materials, aqueous samples were characterised at both 5 $mgmL^{-1}$ and 10 $mgmL^{-1}$.[54] The 10 $mgmL^{-1}$ aqueous solutions were subsequently used to prepare 1:1 multicomponent systems with 5 $mgmL^{-1}$ concentrations of each component (Section 4.4.2).[13] Building a predictive system that probes self-assembly using these different modes of tunability could yield a system with maximised utility and insight. Using the high-quality, high-throughput SAXS afforded by the I22 beamline, over 1,000 measurements were analysed. After removing any problematic samples and those repeated due to an error in the measurement, a collection of 742 samples was reliably obtained across solvents, concentrations, and combinations (Section 4.4.3).

4.2.2 Scattering Quality

While predicting the more complex topological parameters from the molecular descriptors is the ideal target, the data-processing limitations necessitated a simpler one. A useful predictive target for these systems is to classify whether the molecules scatter well enough to produce high-quality profiles from which such topological parameters can be determined. This endpoint has significant uses as it allows one to narrow down rational candidates for future scattering experiments. As previously discussed, access to synchrotron radiation facilities is limited, making the on-demand characterisation of materials unachievable and driving interest in target predictive systems. This also means that, when SAXS is obtained, it is ideally undertaken on candidates that will provide useful insights. Furthermore, because the small angle scattering technique requires the presence of aggregated structures to yield significant detectable scattering, the endpoint of classifying scattering quality can serve as a useful proxy for identifying which molecule-solvent combinations afford self-assembled structures, allowing candidate combinations to be predicted and identified. Such tools could narrow the scope of experimental investigation to those candidates most likely to afford interesting functional materials.

As discussed previously, for an effective QSPR to be established, a well-defined endpoint is required. It was necessary, therefore, to find a useful classification regime to separate the samples with poor scattering profiles from those with substantial scattering in the QSPR dataset. The sample preparation and characterisation procedures are described in Section. 4.4.2 and 4.4.3. With such significant volumes of data, it was clear that manual classification would be ineffective, overly reliant on poorly articulated intuition, and prone to human error. It was therefore necessary to find a well-defined metric to classify data quality.

To fit SAXS data to topological models, the two-dimensional (2D) scattering intensity (I(q)) pattern is azimuthally integrated to yield a 1D scattering profile that measures I(q) as a function of the scattering vector q (Section 4.1.3). As each 1D datapoint in q is defined by a series of 2D I(q) measurements that will vary in intensity, the azimuthal integration provides both the reduced 1D I(q) and the resulting error of I(q). As the magnitude and quality of the scattering decrease, the error of I(q) increases. This error, therefore, offers a useful insight into the quality of the scattering profile and an approach to quantify it. Where a significant proportion of the datapoints have measurement errors in excess of the scattering intensity, the scattering is deemed to be poor, arising from the absence of well-defined aggregates from which scattering can occur. Subsequently the difference between I(q) and its error was

calculated for each datapoint, and the proportion of positive and negative values calculated for each profile. As q in increases, I(q) naturally decreases, therefore it is expected that the scattering intensity and error are more likely to converge at high-q. Therefore, the proportion positive datapoints was determined across three thresholds to afford a set of useful rules for classifying the scattering quality of our data. Where 90% of points below q = 0.1 and q = 0.2, and 95% of points below q = 0.3 were positive, the sample was deemed to afford good scattering. Where all criteria failed, the samples were deemed to afford bad scattering. The few samples which only passed one or two criteria were flagged for manual assessment. To ensure a consistent and articulable protocol was applied, it was deemed appropriate to assign these intermediate samples as poor scatterers (Figure 4.7). Each SAXS plot and classification analysis is provided in Appendix 4.2, with summaries of the classifications provided in Appendix 4.3. By adequately automating the large scale dataset and utilising a symlog plot to visually represent the quality of each in an easy to assess format, a manual validation of the QSPR dataset could be more reliably undertaken, ensuring that a robust dataset is prepared.

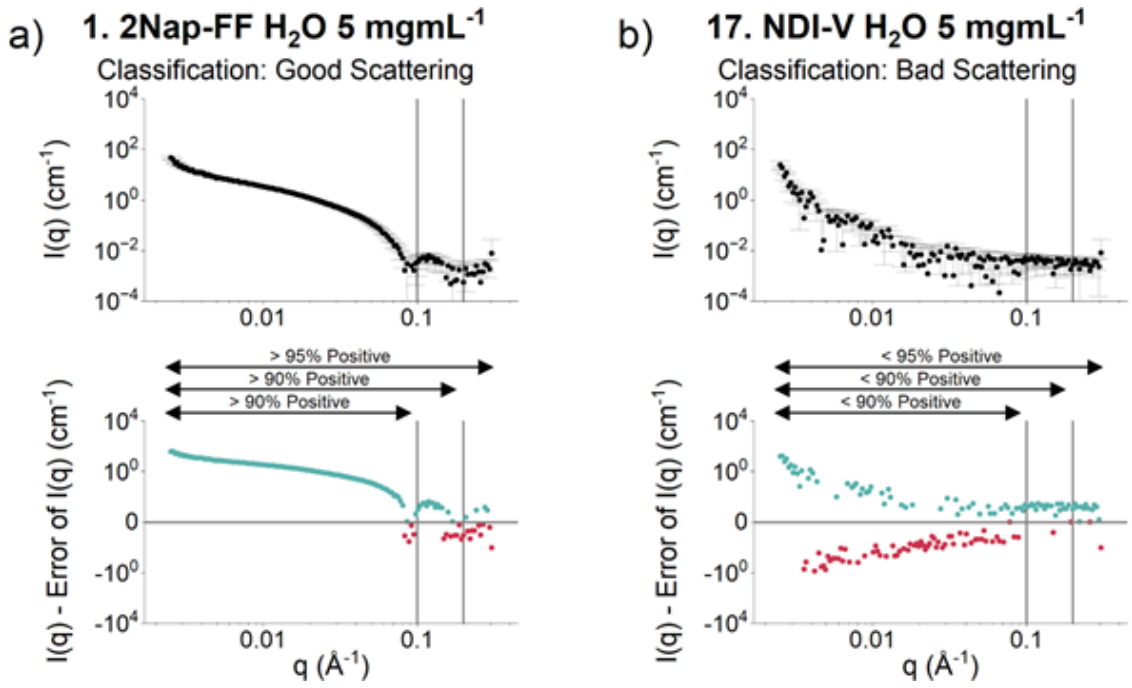


Figure 4.7. Example scattering quality classification regime for a) 2Nap-FF (**1**) and b) NDI-V (**17**) in H_2O at 5 $mgmL^{-1}$. The top plot shows a 1D SAXS plot of q vs I(q) with 1.0 background subtraction in log-log scales. The lower plot shows q vs the difference between I(q) and its error after background subtraction, on log-symlog scales.

Previous analysis of naphthalenes and rylenes using SAXS has shown that the difference in scattering length density (SLD), that is a measure of the material's electron density that determines X-ray scattering power, is sufficient to afford a reliable contrast between the

solvent and the self-assembled structure that forms in solution.[48] Subsequently, a loss in scattering quality must arise from the molecule's preference to molecularly dissolve rather than aggregate with other solute molecules. Therefore, while the model is built on the well-defined endpoint of identifying solute-solvent systems as good or bad scatterers, it is a useful and simple proxy for predicting the more complex question of whether those molecules are self-assembling. By enabling the prediction of combinations that are soluble enough such they can enter solution, but not so soluble that they prefer to molecularly dissolve completely, the scope of investigation of useful candidates of material preparation application can be significantly narrowed.

While the endpoint used in this study was sufficient to provide a simple assessment of SAXS data quality, future models may achieve more robust predictions if the chosen endpoint better captures the characteristics of the scattering profile. As scattering techniques become increasingly widespread and are employed in progressively more complex analytical workflows, there remains no consensus on how experimental uncertainty should be quantified. This is particularly important given the numerous sources of error that can arise during both SAXS measurement and data processing. Such sources include sample preparation artefacts, radiation damage during measurement, scattering contributions from experimental components, detector noise, counting statistics, and inaccuracies introduced during buffer subtraction.[55] Future models may benefit from estimating the scattering quality using metrics such as the signal-to-noise ratio (SNR) which can similarly be described by I(q) and its error (σ(q)) (Equation 4.4).[56]

$$SNR = \frac{I(q)}{\sigma(q)} \qquad \text{(Equation 4.4)}$$

Signal-to-noise ratio has recently been employed to investigate the effect of data quality on ML systems developed to fit topological models of nanoparticle scattering data.[56] As a relative measure of uncertainty, SNR may provide a more appropriate assessment of data quality across the full q-range than absolute error metrics. Because SAXS intensities typically span several orders of magnitude, absolute measures are often dominated by the large intensities observed at low q, potentially obscuring trends elsewhere in the profile. Models trained using SAXS data with low SNR exhibited poor predictive performance, whereas substantial improvements were observed as SNR increased. Based on these findings, it has been proposed that SNR thresholds could be used to identify datasets of sufficient quality for downstream analysis and modelling.[56] Subsequently assessing

scattering quality on the basis of SNR could make a useful endpoint metric for future QSPR modelling.

It should be noted, however, that both the endpoint metric used in this work and the SNR metric implicitly assume that the uncertainties reported in the one-dimensional SAXS profile accurately describe the noise present in the data and are primarily governed by Poisson counting statistics. In practice, experimentally derived uncertainties often deviate from these assumptions owing to the substantial contribution of both systematic and statistical error sources.[55] As a result, although both metrics provide useful indicators of SAXS data quality, neither is capable of fully describing the reliability of a dataset in isolation. Therefore, while the model developed here can help narrow the search space towards more promising solute-solvent combinations, further advances in the quantification of SAXS uncertainty may be required to support the development of more predictive and broadly applicable modelling approaches.

Applying our classification regime to samples measured yielded a classified dataset of 742 samples. The samples represent solutions prepared in 6 solvents (acetone, DMF, DMSO, ethanol, THF, and water), with some examples prepared in deuterated solvent analogues as well as their non-deuterated counterparts. While all organic samples were analysed at 5 $mgmL^{-1}$, some aqueous samples were analysed at both 5 and 10 $mgmL^{-1}$ and included in the QSPR dataset to assess its ability to predict based on concentration, as well as solvent and solute structure. The 10 $mgmL^{-1}$ aqueous solutions were also combined 1:1 to yield multicomponent solutions with 5 $mgmL^{-1}$ each of two components. The samples were separated into different datasets to build QSPRs from varying data and assess how the information content afforded by different solution types in the QSPR dataset impacts the predictive power of the resulting models. A dataset comprised of all solutions quantified by SAXS was first prepared (519 good, 223 bad scatterers). From this, all multicomponent samples were removed to afford a dataset comprising all single-component solutions in organic and aqueous solvents (100 good, 147 bad scatterers). A further dataset was prepared with only single-component aqueous samples (72 good, 33 bad scatterers), from which all deuterated samples were removed to afford a final single-component H_2O only dataset (48 good, 33 bad scatterers) (Figure 4.8). With datasets prepared and classified, they could now be processed to train QSPR models.

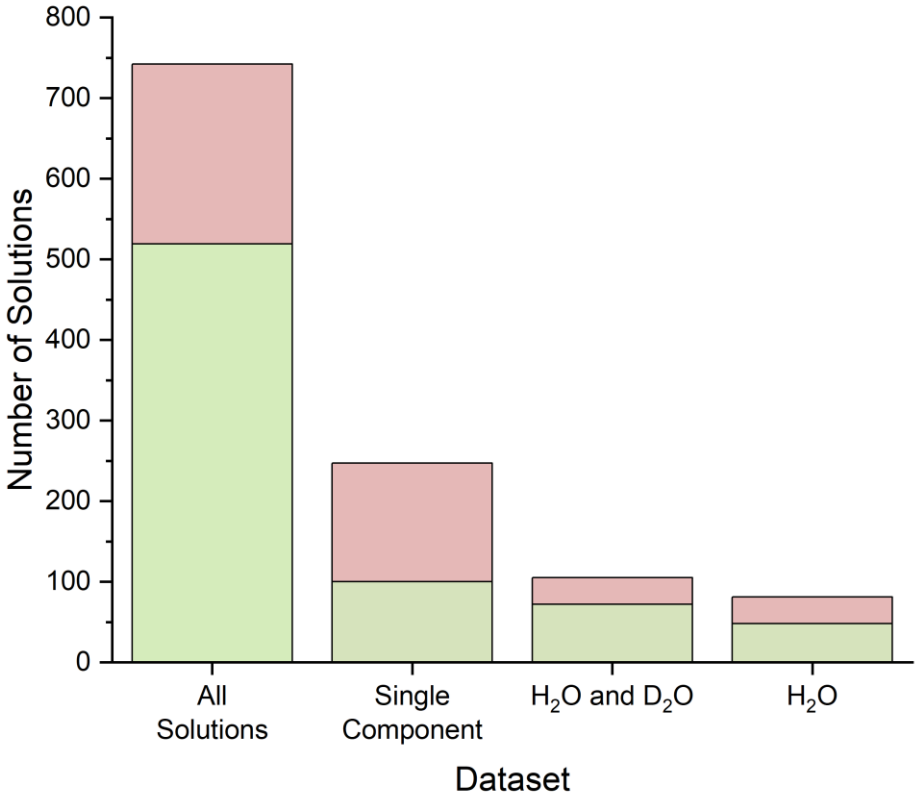


Figure 4.8. Class distribution of solutions which comprise the training, test, and validation sets. The green segment represents the portion of compounds deemed to scatter well by the classification procedure described in Section 3.4.7, while red represents those deemed to scatter poorly.

4.2.3 Data Fitting

As discussed in Section 4.1, SAXS data can be fitted to topological models to gain further insight into the aggregated structures that have formed. However, due to the complexity of scattering patterns, and the need for domain expertise and *a priori* knowledge to appropriately fit and optimise topological models, properly characterising such structures can be time-consuming and difficulty. Subsequently, having a quality metric that allows profiles, likely to lack regular structure, to be accelerated by enabling targeted focus samples to have well-defined structures. This would better enable the identification of useful candidate materials and the search for structures likely to yield the desired properties. For example, we find that for 6MeO2Nap-FI (**6**), screening out the bad scatterers effectively removes those with no well-defined topological parameters to fit (Figure 4.9).

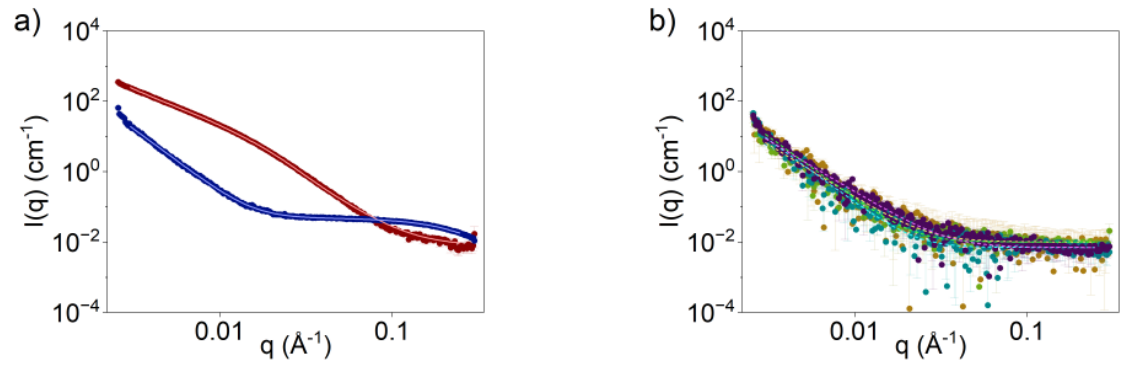


Figure 4.9. SAXS data (filled circles) and fits (line) for a solution of 6MeO2Nap-FI (**6**) at a concentration of 5 mgmL^{-1}. a) represents the good scatterers: acetone (red), and THF (blue). b) represents the bad scatterers: DMF (yellow), DMSO (green), ethanol (teal), and water (purple).

Each of the profiles shown in Figure 4.9 was afforded by solutions prepared at 5 mgmL^{-1}, well below the saturation solubility measured in Chapter 3. Subsequently, one might assume that each solution is molecularly dissolved and monodisperse. Indeed, for the DMF, DMSO, ethanol, and water solutions, all the data overfit even to simple, structure-independent power laws, highlighting the frivolity of fitting such data, and identifying them as combinations that are dissolved as expected. In contrast, acetone and THF solutions were both fitted to cylindrical models. While 6MeO2Nap-FI (**6**) formed long rigid rods in THF, switching to acetone allows for the formation of thicker, flexible fibres (Appendix 4.4.1). The significant differences between the good and bad scattering profiles highlight the information that can be obtained from a simple model; however, much information is lost. For example, with PBI-DOPA (**22**), screening the bad scatters was not sufficient to guarantee that large anisotropic structures useful for the design of supramolecular materials are found (Figure 4.10).

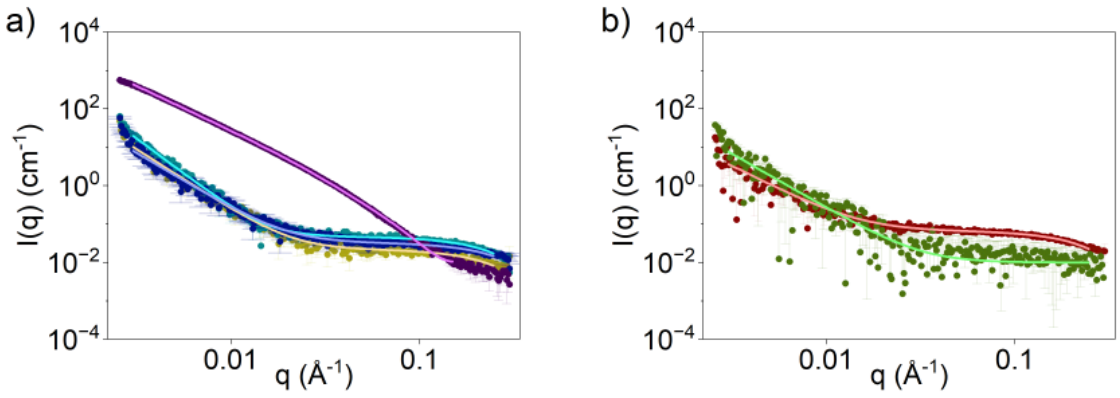


Figure 4.10. SAXS data (filled circles) and fits (line) for a solution of 6MeO2Nap-FI (**6**) at a concentration of 5 mgmL^{-1}. a) represents the good scatterers: DMF (yellow), ethanol (teal), THF (blue), and water (purple). b) represents the bad scatterers: acetone (red) and DMSO (green).

While the scattering profile of PBI-DOPA in water indicates the presence of longer flexible fibres, the scattering of the DMF, ethanol, and THF solution which were also considered good scatterers, only contained small spherical aggregates (Appendix 4.4.1). Therefore, while predicting solubility can screen out molecules incapable of molecular mobility in a given solvent, and characterising scattering quality can screen out those that are too soluble to form micellar aggregates, future models will require more complex topological descriptor datasets to provide topological predictions.

Fitting of select compounds also confirmed the necessity of treating solvents and their deuterated analogues as independent solvents. Comparing 13 different amino acid functionalised perylene bisimides in aqueous solutions, it was found that the effect of changing the solvent isotope was significant in roughly half of the measured samples, with no observable pattern distinguishing which structures were susceptible to such effects. While a similar investigation of 3 perylenes in ethanol and their deuterated analogues yielded comparable profiles, the investigation of 9 perylenes and a naphthalene also showed that, in a few cases, organic solutions of these compounds may exhibit a significant difference in the scattering profile upon deuterating the solvent (Appendix 4.2 – 4.8). For example, PBI-E (**23**) shows minor change when prepared in D_2O compared to the H_2O counterpart, both fitting to flexible elliptical cylinder models (Figure 4.11a, Appendix 4.4.2). The main differences arise in dimensions, where in the aggregates in D_2O appear to be slightly shorter, less flexible, and with more variable radii. In contrast, PBI-M (**29**) forms well-defined one-dimensional aggregates in the nanoscale region in H_2O, but such aggregates are not observed in D_2O. Instead, PBI-M (**29**) prefers the formation of smaller spherical aggregates (Figure 4.11b, Appendix 4.4.2). Given the importance of forming long one-dimensional fibres for the development of many functional supramolecular systems, materials such as PBI-M may be discounted as viable candidates when characterised in deuterated environments and not treated as independent of the non-deuterated analogue.

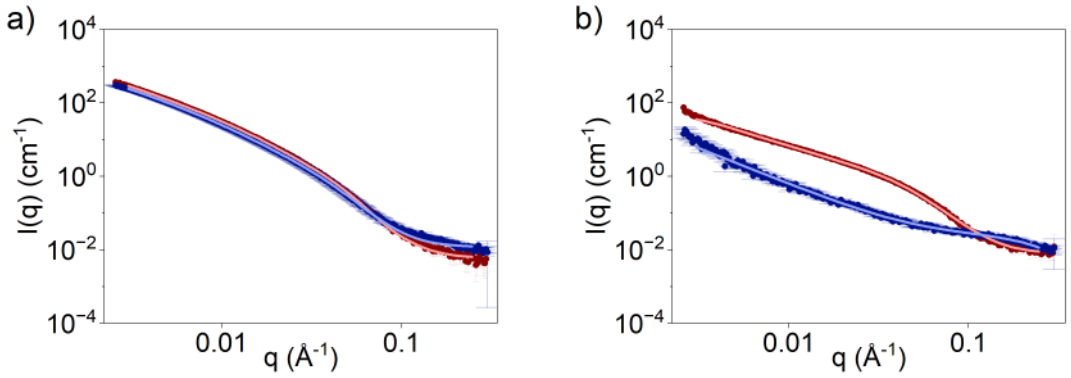


Figure 4.11. SAXS data (filled circles) and fits (line) for a solution of a) PBI-E (**23**) and b) PBI-M (**29**) at a concentration of 5 mgmL^{-1}. Red data represents the solution in H_2O blue for solution in D_2O.

It was also noted that the isotope effect was capable of promoting and inhibiting the self-assembly of different molecules in DMSO and its deuterated analogue. While 2Nap-IF (**4**) formed rigid cylindrical fibres in DMSO, aggregation did not occur when dissolved in d_6-DMSO (Figure 4.12a, Appendix 4.4.3). Conversely, while dissolving PBI-L in DMSO did not afford aggregation, rigid cylindrical aggregates were formed in d_6-DMSO.

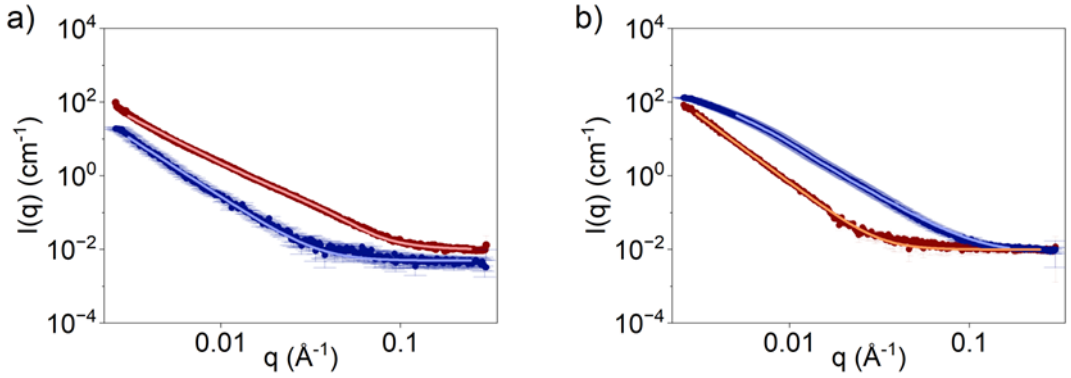


Figure 4.12. SAXS data (filled circles) and fits (line) for a solution of a) 2Nap-IF (**4**) and b) PBI-L (**28**) at a concentration of 5 mgmL^{-1}. Red data represents the solution in H_2O blue for solution in D_2O.

Conversely, no significant isotope effect was identified across the ethanol and d_6-ethanol samples. For example, PBI-DOPA (**22**) only formed small, well-defined spherical aggregates in both solvents (Figure 4.13, Appendix 4.4.4). Therefore, the solvent isotope effect can provide access to different aggregates, for which there is currently no way to predict how changing the solvent isotope may affect the resulting topology.

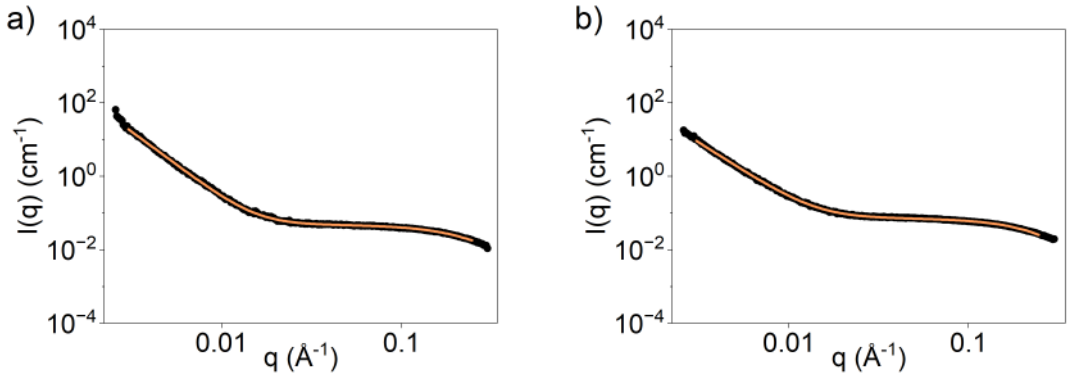


Figure 4.13. SAXS data (filled circles) and fits (line) for a solution of PBI-DOPA (**22**) at a concentration of 5 mgmL^{-1}. a) represents the sample prepared in ethanol. b) represents the sample prepared in d_6-ethanol.

Subsequently, conclusions drawn from techniques and characterisations that rely on solvent deuteration should not be assumed to apply to the typical environment. Instead, it must be shown, through supplemental characterisation, that there are no significant differences between the systems. If the assumption holds, such analyses may be used interchangeably in developing our QSPR dataset. However, our results indicate that this assumption, if not accounted for, would be a significant source of data inhomogeneity in the dataset. To maximise the data available to train the model, both the deuterated and non-deuterated data points were used as systems in the QSPR dataset. However, to distinguish between samples, it would be necessary to introduce descriptors that uniquely describe solvents and their deuterated analogues.

4.2.4 QSPR Data Preprocessing

To build a QSPR model capable of predicting a property from the chemical structure, the model must first interpret the compound's structural features. Therefore, each molecule was generated *in silico* using ChemDraw and converted to SMILES format.[21, 22] From the SMILES representation, molecular descriptors were calculated in KNIME (Chapter 2: Section 2.4.2).[23] In addition to representing the molecular structure of the solute, it is necessary to provide some description of the other relevant variables of the solution preparation. While the concentration is easy to represent numerically, we opted to include the gravimetric concentration (5 mgmL^{-1} or 10 mgmL^{-1}); there are many more ways to describe the solvent. As a molecule, the solvent can be represented by a separate fingerprint. However, to aid interpretability, simpler metrics to differentiate the solvents were considered. Given the vital importance of solubility and the nonbonding interactions afforded by the solvent during self-assembly, a simple metric capturing solvent polarity was deemed a rational choice. While no single metric completely describes the polarity of a

solvent, the dielectric constant is a common metric used to describe a solvent's polarity. As discussed in section 4.2.1, comparisons of scattering profiles between deuterated and non-deuterated solution analogues revealed significant differences in the scattering profile in some cases. Subsequently, to adequately differentiate the solutions when both data sets are included in the models, an additional descriptor was required to identify the solvent isotopes. Although the isotope effect on the dielectric constant is very small, many other physicochemical parameters change more significantly when the isotope distribution is altered (Table 4.3).[57] While in the case of water, viscosity is one of the most significantly affected parameters upon deuteration, the appropriate values for other organic solvents are not well described in the literature. Density, while exhibiting a much smaller relative difference compared to viscosity, is affected much more than the dielectric constant. Density information for solvent isotopes is also much more accessible. While there is no centralised literature source for these values; they are readily available from chemical distributors in reports on their physicochemical properties. This made density an ideal additional descriptor to discriminate between solutions and their deuterated analogues.

Table 4.3. Physicochemical properties of H_2O and D_2O and their relative difference compared to H_2O.[57]

	H_2O	D_2O	Relative Difference (%)
Dielectric Constant	78.304	77.937	-0.47
Boiling Point (°C)	100.00	101.42	1.42
Density (g cm^{-3})	0.997	1.104	10.77
Viscosity (N s m^{-2})	8.9×10^{-4}	1.10×10^{-3}	22.98

While the dielectric constant is often predicted to serve as a quantitative measure of solvent polarity, it assumes solvents are an isotropic continuum rather than composed of individual molecules with their own solvent-solvent interactions. Furthermore, the dielectric constant fails to fully describe the charge distribution responsible for the correlation with polarity and does not account for the influence of higher multipole moments.[58] Future models may be improved by retraining them using descriptors that better capture the characteristics of the solvent. There are hundreds of quantitative measures of solvent polarity, each derived from a different physical basis.[59] Solvent polarity is much better described by molecular–microscopic empirical solvent parameters derived from suitable solvent-dependent reference processes.[60] Dimroth and Reichardt's $E_T(30)$ parameter, an empirical solvent polarity parameter derived from the measurement of intermolecular charge transfer absorption bands in solvatochromic dye solutions, has one of the most extensive libraries of polarity

quantifications.[61] $E_T(30)$ has previously been established as an appropriate metric for a uniform polarity scale of organic solvents and ionic liquids.[62] Therefore, future models may extract more useful relationships when using $E_T(30)$ as a descriptor of solvent polarity.

As discussed in Section 3.2.5, to establish such a relationship, descriptors capable of reliably segregating the property classifications are necessary. Ideally, simple relationships between descriptors and properties of interest can be identified to determine Lipinski rule-of-five-like relationships.[31] Similar to the relationships discussed in Chapter 3, no simple relationship was found with the simplest physicochemical descriptors. However, while there was still significant overlap between the classes across the property distribution, there was a clear trend for larger solutes to afford better scattering. This is certainly expected for the dataset, made up in part of multicomponent systems, where if the solution contains at least one component that scatters well individually, the multicomponent system will also scatter well (assuming no disruption of self-assembly by a secondary component). However, these relationships also persist in the single-component and aqueous solution model data, indicating that while unable to segregate the property classes alone perfectly, larger aromatic solutes will tend to afford better scattering (Figure 4.14, Appendix 4.4). As discussed, it was necessary to include descriptors of the solution composition, which can also be investigated for simple relationships (Figure 4.15, Appendix 4.5). While increasing concentration is shown to improve the likelihood that the solution scatters well, it is still not sufficient to predict this alone. The distributions for the H_2O and D_2O models highlight a limitation of the dataset: all D_2O samples analysed were analogues of H_2O samples that scatter well. The distribution highlights that all D_2O scatter well, with no negative examples. The inclusion of these samples in the dataset may bias model predictions by introducing duplicate compounds, whose D_2O equivalents provide no new predictive information. To investigate the possibility, a dataset using only the non-deuterated compounds was used to build an additional QSPR relationship. We also attempted to find a simple relationship using more information-rich Morgan fingerprint descriptors, a simple representation of the presence of different chemical substructures, but no relationship was found (Figure 4.16, Appendix 4.6).[33] As with the solubility models, the lack of well-segregating descriptors suggests that developing Lipinski-like rules to predict solubility would be challenging for these solutions. Again, this is expected, as the numerous interactions that affect a molecule's ability to self-assemble and afford well-defined scattering patterns make it difficult for a single descriptor to capture the nuances of solubility classification. However, the presence of descriptors that

improve class discrimination relative to those discussed in Chapter 3 suggests that QSPR classification models with even better predictive performance may be achieved.

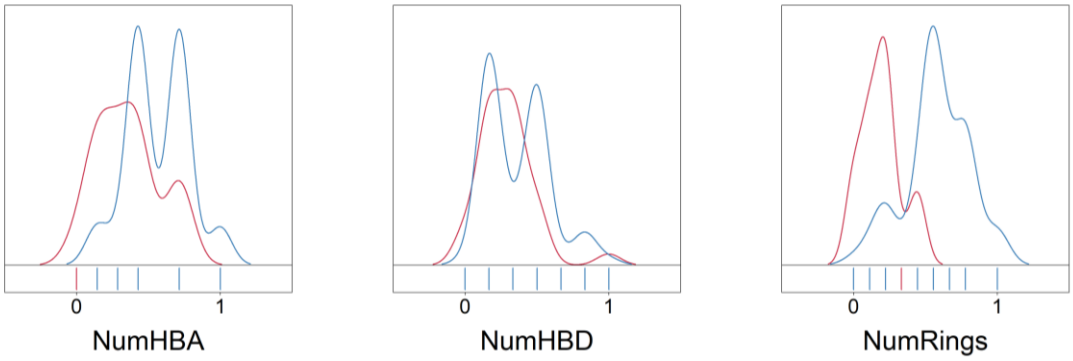


Figure 4.14. Select physicochemical descriptors calculated and their distribution, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the H_2O and D_2O model. The value of each molecule's descriptor is depicted by | marker. Values are normalised from 0 to 1.

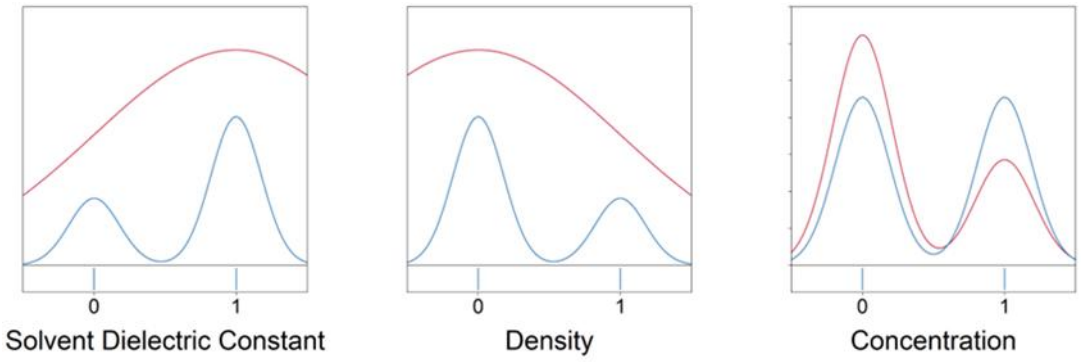


Figure 4.15. Solvent descriptor distributions, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the H_2O and D_2O QSPR models. The value of each molecule's descriptor is depicted by | marker. Values are normalised from 0 to 1.

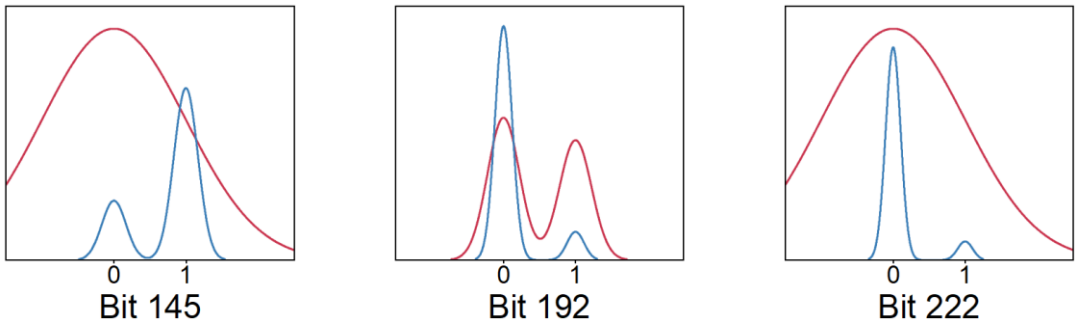


Figure 4.16. Select Morgan fingerprint descriptors and their distributions, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the H_2O and D_2O model.

The final step before modelling can begin is to prepare the datasets. An important step is partitioning the dataset to enable effective assessment of the model's performance. As discussed in Chapter 2, it is necessary to segregate a portion of the data to test the performance of the machine learning models. While Chapter 3 used solubility measurements

as the basis for the QSPR dataset, which can be obtained in a standard chemistry laboratory, one major limitation of our scattering dataset is the inaccessibility of scattering techniques. This makes the procurement of new data non-trivial without access to laboratory SAXS equipment. Subsequently, even if the synthetic limitations discussed in Chapter 3 were overcome, the analysis of new solutions is not guaranteed. Subsequently, rather than synthesising new validation molecules for final external validation, it was deemed necessary to prepare validation sets from the overall QSPR datasets. Before any assessment, approximately 10% of the classified data was partitioned from the dataset, stratified so that the validation set maintained a class distribution similar to that of the overall dataset. While removing data limits the predictive information available to the model during training, the larger datasets obtained by high-throughput experimentation should minimise the impact on the final model. From the remaining data, a 20% portion of the QSPR dataset was removed for independent testing after modelling to assess predictive performance on unseen data and to assess the appropriate applicability domain.

To minimise the impact of partitioning the QSPR dataset, the modelability index (MODI) was used to screen potential dataset partitions (Chapter 2, Section 2.4.3). While the models from Chapter 3 showed that maximising the dataset MODI did not guarantee maximised performance, it was clear that the lowest MODI partitions did not yield effective models. Therefore, to minimise the investigation of partitions less likely to afford effective models, stratified partitions were taken at random, and only carried forward if the MODI of the training set exceeded the typical MODI threshold of 0.65 (Table 4.4).[63] In addition, to maximise the applicability domain of the resulting models, the proportion of models found to be within the maxdist and dsz applicability domains as defined by Sahigara *et al.* was calculated (Chapter 2, Section 2.4.4).[64] If the proportion of test-set molecules deemed to lie outside either domain exceeded 20%, the partition procedure was repeated. By focusing on models with a larger domain of applicability, the final model will have greater utility, better describing a wider range of possible solute-solvent combinations.

Table 4.4. MODI of training set partitions across 10 random stratified partitions, and the specific MODI values of the best-performing model.

Model	Hyperparameters	Training Set MODI Range	Optimal Training Set MODI
All Solutions	Original	0.72 – 0.76	0.76
Single Component	Original	0.66 – 0.72	0.72
H_2O and D_2O	Original	0.72 – 0.80	0.72
Single Component	Updated	0.66 – 0.72	0.67
H_2O	Updated	0.66 – 0.82	0.69

As with the datasets discussed in Chapter 3, it is important to screen and remove descriptors with low information content and subsequently balance the datasets. Here, we applied a low variance threshold of 0.1 and a high-correlation threshold of $r^2 = 0.9$ using Spearman's rho rank correlation to reduce the number of physicochemical, solvent, and fingerprint descriptors from 1036 to between 59 and 75 descriptors of higher informational content (Chapter 2: Section 2.4.5).[65] The synthetic minority oversampling technique (SMOTE) was employed instead of simpler undersampling techniques due to avoid removing crucial predictive information from the training set (Chapter 2: Section 2.4.6).[34] As previously discussed in Chapter 3, SMOTE was configured to use the 5 nearest neighbours to generate synthetic datapoints as it is the typical choice, future models may benefit from an investigation of the optimal SMOTE configuration on these datasets.

4.2.5 Machine Learning Model Training

The QSPR models were generated using a variety of machine learning methods: support vector machines (SVM),[35, 36] random forests (RF),[37] k nearest neighbours (kNN),[38] XGBoost (XG),[39] logistic regression (LR),[40] naive Bayesian (NB),[40] and C4.5 ensemble decision trees.[37] Each modelling method employed fingerprint, physicochemical, and solvent descriptors to capture molecular properties. As discussed in Chapter 2 (Section 2.5), each machine learning technique has its own strengths and weaknesses. By identifying multiple predictive models, our capacity to rationally design new materials with tailored properties would be significantly improved, as the average of multiple good prediction models should be closer to the actual measurement than any given individual model.

For each machine learning algorithm to successfully predict the properties of interest, its setup must be tuned to establish effective relationships. Hyperparameters are the variables of a machine learning model that can be modified to control its training process. One of the greatest barriers to efficient model development is selecting the appropriate hyperparameter values to investigate. To find parameters which optimised the QSPR performance of

scattering quality prediction, a brute-force grid search of the same hyperparameter space discussed for previous models was undertaken (Chapter 3: Section 3.2.6, Table 3.15). However, we discussed the limitations of the original hyperparameter space and potential improvements suggested from the sparse literature. Therefore, to attempt to improve the quality of the updated models, a set of smaller hyperparameters was applied to later iterations of our models (Table 4.5). By reducing the learning rate and the tree depth, the chance of model overfitting should be reduced, giving rise to more meaningful models. By reducing the number of trees and cross-validation steps, the computational cost of the optimisation procedure is reduced, which, if the models maintain or improve in performance, improves the efficiency of the training procedure.[66-69]

Table 4.5. Updated hyperparameter combinations optimised by a 5-fold cross-validation procedure.

Model	Parameter	Start	Stop	Step
RF	Maximum Depth	5	20	5
	Number of Trees	100	1000	100
C4.5	Minimum Node Size	2	10	2
	Number of Trees	100	1000	100
	M_{Try}	0.2	0.8	0.2
	Sample Fraction	0.2	0.8	0.2
	Maximum Depth	5	20	5
kNN	Number of Neighbours	3	20	1
SVM	σ	-5	5	0.05
XG [a]	Depth	5	20	5
	Number of Trees	100	1000	100
	Learning Rate	0.01	0.1	0.01
LR	Learning Rate	0.01	0.1	0.01
	Variance	0.1	0.8	0.1

[a] For simplicity, the log10 of the SVM σ hyperparameter input has been reported. Therefore, while all other parameters use a linear interval, the σ interval represents the logarithmic interval of each iteration.

A repeated 5-fold cross-validation approach was used to select the optimal QSPR model for each method based on the largest Matthew's correlation coefficient (MCC) (Chapter 2: 2.5.7). An ideal model has an MCC value of 1, a random model has a value of 0, and an acceptable value is > 0.5. Using a cross-validated approach provides a good estimate of the models' predictive power. The models generated by each machine learning method, along with their associated statistics, are shown in Appendix 4.7. While the classes are significantly more balanced across the models, compared to the solubility models discussed in Chapter 3, there is still a reasonable imbalance with an approximate good: bad scattering class ratio of approximately 0.7–2.3:1 across the datasets (Figure 4.8). Therefore, the balanced accuracy, Matthew's correlation coefficient, and Cohen's kappa (κ), which provide metrics independent of skewed class sizes, were used (Chapter 2: Section 2.6.1).

In any data set, model performance will vary with the dataset composition. The negative effects of training and test set portioning can be significant, especially when modelling with small datasets. Therefore, it was expected that the changing partition of the training data would heavily impact model performance. Therefore, for each dataset, 10 randomised partitions were investigated. The split was stratified to ensure that solubility classifications in the training set were representative of the entire dataset. The model with the best metrics was carried forward for further investigation. Appendix 4.8 shows a dataset dependence on prediction performance across the models. While the smaller datasets vary significantly, the larger "all solutions" QSPR dataset is much more robust to dataset partitioning. The training set performance of the optimal dataset split is summarised in Table 4.6.

Table 4.6. Overall quality of predictive performance of optimised QSPR models on the training set. Model is deemed to be good (G) when the following are satisfied: balanced accuracy > 0.7, MCC > 0.5, $\kappa > 0.4$. Models are otherwise deemed to be bad (B).

Model	Hyperparameters	Method						
		RF	C4.5	kNN	SVM	XG	NB	LR
All Solutions	Original	G	G	G	G	G	B	G
Single Component	Original	G	G	B	B	G	B	G
H_2O and D_2O	Original	G	G	G	B	G	G	G
Single Component	Updated	G	G	B	G	G	B	G
H_2O	Updated	G	G	G	G	G	G	G

4.2.6 Applicability Domain

As discussed in Chapter 3, assessing the model's quality by applying it only to the training set is insufficient to gauge its true predictive performance on unseen data. To assess if a candidate molecule has sufficient chemical similarity, distance-based metrics described by Sahigara *et al*. were used to establish well-defined applicability domains.[64] For each training partition, the Morgan fingerprint descriptors as well as the normalised physicochemical and solvent descriptors were used to characterise well-defined distance-based applicability domains (Chapter 2: Section 2.4.4). It is expected that the resulting QSPRs will better model molecules which are more similar to the training set. Therefore, it is important to choose the appropriate applicability domain to ensure the reliability of predictions is well-defined. As previously discussed, the maxdist was equivalent to the 2d, 3d, and p95 in performance metrics; therefore, as one of the more generous boundaries, the metrics obtained after

applying the maxdist domain were compared with the much more pessimistic dsz boundary. By considering the models' performance metrics after applying the maxdist and dsz applicability domains, the impact of these domains can be identified.

Contrary to previous models, applying the more restrictive dsz applicability domain yielded poorer predictive performance than the less restrictive maxdist domain. This could arise from a slightly increased spread of descriptor values, making the data centroid less representative of each data point. We find that descriptor values vary more than the descriptors described in Chapter 3. For example, the average descriptor distribution for the Acetone solubility classification model after filtering was 0.317 ± 0.432 (136%), while the H_2O and D_2O solution model had average descriptor values of 0.229 ± 0.355 (155%) after descriptor filtering. It could also suggest that a more generalised model has been achieved, one that better describes a wider applicability domain rather than one overfit to the narrower space of the training data. The performance of the models on the in-domain test sets is reported in Appendix 4.7 and summarised in Table 4.7.

Table 4.7. Overall quality of predictive performance of optimised QSPR models on the in-domain maxdist test set. Model is deemed to be good (G) when the following are satisfied: ROC A, balanced accuracy > 0.7, MCC > 0.5, and K > 0.4. Models are otherwise deemed to be bad (B).

Model	Hyperparameters	Method						
		RF	C4.5	kNN	SVM	XG	NB	LR
All Solutions	Original	G	G	G	B	G	B	G
Single Component	Original	G	G	B	B	G	B	G
H_2O and D_2O	Original	G	G	G	B	G	G	G
Single Component	Updated	G	G	G	B	G	G [a]	G
H_2O	Updated	G	G	G	B	G	G	G

[a] An in-domain test set yielded good metrics despite poor training metrics. The model is expected to be less effective and not carried forward.

The predictive performance of each model is further assessed using the receiver operating characteristic (ROC) curve (Appendix 4.9). ROC plots the true positive prediction rate against the false positive rate. While the high predictive power afforded by these models on the test set means there is very little change in performance with changes in the choice of applicability domain, the maxdist offered marginal improvements over the dsz, further rationalising it as the appropriate choice for these models (Figure 4.17).

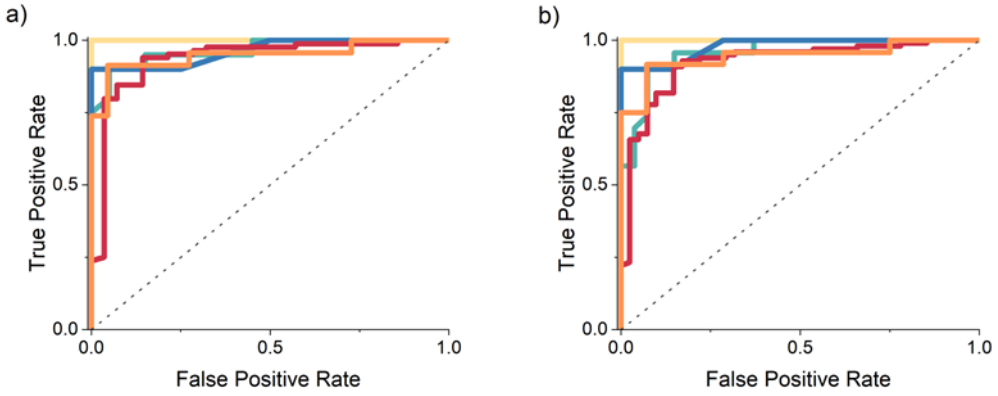


Figure 4.17. ROC curves of the random forest models for each model. a) shows the ROC performance of the maxdist domain. b) shows the ROC performance of the dsz domain. Curves are coloured based on the dataset and hyperparameters as follows: Original hyperparameter models: all solutions (red), single component (orange), H_2O and D_2O (yellow); updated hyperparameter models: single component (green) and H_2O (blue).

4.2.7 Model Robustness

To further investigate the robustness of the optimised models, randomisation tests were conducted in which hyperparameter optimisation was conducted on a randomised training set. The performance metrics should be much poorer than those derived from the true data. These metrics identify the statistical significance of the QSPR model.[43] Indeed, the MCC metrics resulting from the randomisation tests were significantly poorer than those of the models found with true data (Table 4.8).

Table 4.8. Comparison of MCC scores of the in-domain test set on models built using the true data and models built on randomised data.

			Model and Hyperparameters				
			All Solutions Original	Single Component Original	H_2O and D_2O Original	Single Component Updated	H_2O Updated
Method	RF	True	0.699	0.675	1.000	0.763	0.887
		Randomised	0.00 ± 0.14	0.00 ± 0.19	0.00 ± 0.27	0.00 ± 0.10	0.08 ± 0.29
	C4.5	True	0.718	0.736	0.894	0.793	0.887
		Randomised	0.01 ± 0.19	0.10 ± 0.16	0.22 ± 0.27	0.06 ± 0.12	0.07 ± 0.25
	kNN	True	0.498	0.775	0.553	0.484	0.509
		Randomised	0.07 ± 0.10	0.04 ± 0.11	0.12 ± 0.36	-0.03 ± 0.11	-0.02 ± 0.17
	XG	True	0.615	0.741	1.000	0.950	0.789
		Randomised	0.00 ± 0.09	-0.12 ± 0.09	-0.13 ± 0.13	-0.03 ± 0.07	0.03 ± 0.25
	LR	True	0.670	0.664	1.000	0.585	0.887
		Randomised	-0.03 ± 0.1	0.05 ± 0.14	-0.08 ± 0.24	0.06 ± 0.06	0.02 ± 0.33

This decrease in performance on random data increases our confidence in the strong predictive performance of random forest, C4.5, kNN, XGBoost, and logistic regression models. When applying the maxdist applicability domain to the randomised models, all models showed a decrease in performance compared to the true models; however, three of the four kNN models did not show a significant decrease (the single component and the H_2O and D_2O models using the original hyperparameters, and the H_2O model using the updated hyperparameters). Again, this may indicate poorer generalisability for these models (Table 4.9). Given the consistently poor robustness of the simple kNN models, they were deemed ineffective and excluded from further investigation. Of particular note was the improvement in robustness across all models when using the original and updated hyperparameters to build models on the single-component data. This demonstrates that reducing tree depth and learning rates across models has likely reduced the potential for overfitting, making the discovery of chance correlation in both the true and randomised datasets much less likely. Comparing the H_2O and D_2O models with the original hyperparameters, and the H_2O model with the updated hyperparameters, we do not observe the same trend. This likely arises from the removal of the D_2O data, which could be adequately separated and predicted using the solvent descriptors alone, due to the homogeneity of their scattering classifications. This suggests that while building a model capable of predicting differences between deuterated and non-deuterated analogues will be crucial for models predicting topological parameters, the solvent isotope effect did not afford a sufficient change in the simpler metric modelled here to afford sufficient predictive information. Although less robust, the H_2O model with the updated hyperparameters was expected to provide more insight into the solute-structure contribution to the prediction of scattering quality; it was therefore carried forward as the appropriate model.

Table 4.9. Z-scores for "good" models on the test and in-domain test sets applied to models built with randomised training data. Scores above the 0.05 confidence threshold are highlighted in green, and scores below are highlighted in red.

Model	Hyperparameters	Method				
		RF	C4.5	kNN	XG	LR
All solutions	Original	4.983	3.753	4.394	7.305	6.078
Single component	Original	3.639	4.125	1.468	9.62	4.348
H_2O and D_2O	Original	3.712	2.492	1.203	8.448	4.524
Single component	Updated	7.554	5.869	4.787	13.623	5.303
H_2O	Updated	2.755	3.241	1.562	3.049	2.668

To further show that the prediction performance of the optimised models outperformed randomised models, a one-tailed one-sample t-test was performed comparing the MCC values for the optimised models with those obtained from the cross-validation hyperparameter modelling on randomised training sets (Table 4.10). The t-test shows that the performance improvements for each model were statistically significant, with the highest p-values corresponding to the poorest Z-scoring models.

Table 4.10. One-tailed one-sample t-test p-values for "good" models on the test, and in-domain test sets applied to models built with randomised training data. Scores below the 0.05 p-value threshold are highlighted in green.

Model	Hyperparameters	Method			
		RF	C4.5	XG	LR
All solutions	Original	3.67×10^{-08}	5.5×10^{-04}	1.41×10^{-09}	2.37×10^{-09}
Single component	Updated	9.42×10^{-10}	8.79×10^{-09}	4.90×10^{-12}	3.54×10^{-10}
H_2O	Updated	5.58×10^{-06}	6.93×10^{-05}	2.42×10^{-06}	7.25×10^{-06}

4.2.8 Model Validation

Given that the 4 models demonstrated strong performance in predicting scattering quality classification on test sets across the 3 datasets, a final validation on unseen data was required. As discussed in Section 4.2.5, the inaccessibility of synchrotron beamlines makes subsequent analysis impractical or even impossible. While quantifying the scattering quality of novel molecules derived from predictions in an *in silico* library of compounds is the ideal approach, we settled on using an independent validation set, partitioned from the original data. To avoid biasing the modelling procedure, the data were not investigated in any capacity until the final model validation. While the chemical space these compounds represent is much more limited than that achievable by synthesising new targets, they will still inform the general predictive power of the resulting models.

To externally validate the model, predictions of 75, 31, and 10 solutions were calculated for the "all solutions", "single component", and "H_2O" models (Appendix 4.10). The machine learning model predictions are typically in agreement with one another; fewer than 10% of solutions yield mixed predictions in each case. Limiting predictions to those with complete consensus across all 4 machine learning models, we find that the models perform well on unseen data, with balanced accuracy ranging from 71% on the smallest dataset to 86% on the largest dataset. However, the MCC and Cohen's kappa scores for the H_2O model fall below the threshold for a good model. While these results highlight that significant further

improvements can be made, they also show that reasonable predictive ability has been achieved for the QSPR models built from the All solutions and single component dataset (Figure 4.19). While H_2O can predict with similar accuracy, its performance on the small validation dataset is not sufficiently better than chance to conclude that a powerful predictive model has been achieved, suggesting that the model requires a larger dataset to achieve better performance and a validation set sufficiently large to ensure robust predictions. The higher predictive performance of the "all solutions" and "single component" models is a clear indication of the ability of the QSPR system to establish effective relationships using both molecular structure descriptors describing both single component and multicomponent in tandem with descriptors representing changes to the material processing, including solvent and concentration. This provides a good indication that using the topological parameters obtained from fitting solutions predicted to be good scatterers, in combination with molecular descriptors, will be an effective next step toward developing more complex QSPR relationships.

Table 4.11. Summary of validation predictions for each model. Model is deemed to be good (G) when the following are satisfied: ROC A, balanced accuracy > 0.7, MCC > 0.5, and K > 0.4. Models are otherwise deemed to be bad (B).

Model	Hyper-parameters	All models agree					Models differ	Out of domain
		Correct	Incorrect	Balanced accuracy	MCC	Cohen's kappa		
All solutions	Original	61	9	0.86	0.73	0.73	5	0
Single component	Updated	21	7	0.74	0.50	0.49	3	0
H_2O	Updated	5	4	0.71	0.38	0.25	0	1

Almost all predictions were deemed within the domain of applicability, yet several were incorrectly predicted. To assess if the applicability domain selection procedure during the model evaluation could be improved, the ability of the dsz to exclude the incorrect predictions in the validation was assessed (Table 4.12). While narrowing the scope of applicability improves the proportion of correct predictions within the domain for the H_2O model, the other models exclude a significant proportion of their correct predictions, causing the proportion of incorrect predictions within the domain to increase. Therefore, the distance-based metrics may not be the optimal applicability domain metric for these models.

Table 4.12. Proportion of solutions correctly and incorrectly predicted by each model deemed to be within the dsz domain.

Model	Hyperparameters	Correct	Incorrect

		Prediction		Prediction	
		In dsz	Out dsz	In dsz	Out dsz
All solutions	Original	41	20	6	3
Single component	Updated	12	9	6	1
H_2O	Updated	5	0	2	2

The inability of the simple distance-based applicability domains to adequately describe the model may arise from the clusters of chemical space represented in the dataset. Given that the distance-based metrics defined by Sahigara *et al*. are defined by the distance from the average descriptor value, the resulting applicability domain may be defined from a point in the descriptor space between these clusters, requiring it to either overextend the space to adequately capture the points, including a wide range of chemical space unrelated to the training set. Future models may be improved by investigating more nuanced applicability domains. Approaches such as tree maps have been used previously to offer useful visual assessments of high-dimensional datasets such as those prepared here.[70] Tree MAP (TMAP) visualisation reduces the dimensionality of the data, representing clusters of similar data as branches (Figure 4.18). However, such representations have been shown to struggle to consistently identify poorly described predictions.[71] Sahigara *et al*. assessed several other domain types finding each to have significant limitations. Just as we have employed an ensemble of machine learning models to confirm predictions across methods and improve prediction confidence, it is suggested that evaluating model applicability using all available strategies could increase confidence in the results of the applicability assessment.[64]



Figure 4.18. TMAP representation of compounds in training (blue circles) and test (orange crosses) sets. Out-of-domain compounds are shown as black crosses. The distance between points represents similarity between the descriptor sets. Reproduced from Simpson *et al*.[71]

In contrast to the validation results discussed in Chapter 3, the prediction confidence of the incorrectly classified samples was much larger. At the same time, previous models found that incorrect predictions had the lowest confidence; a majority of incorrect predictions had

confidence over 50%, with several exceeding 90%. Subsequently, prediction confidence is a much less reliable indicator compared to the previous models. While this limits the confidence we can have in each prediction, incorrect predictions with the highest confidence indicate the chemical spaces that are overfit to the training data. Subsequently, those solutions represent the best areas for expanding the dataset and further improving the model's predictive performance.

4.2.9 Feature Importance

As discussed in Chapter 2 (Section 2.7.1), a key feature of a high-quality model is the inclusion of a mechanistic interpretation, if possible. Having identified models with predictive ability, chemical substructure importance can be calculated to extract useful insights for future investigation and rational design. We previously used simple metrics to identify important features derived from shallow decision trees. While they are simple to calculate and can provide some basic insight into the features that underpin the resulting model, a metric indicating the direction in which the presence or absence of a descriptor shifts the prediction would better aid the selection of appropriate features when rationally designing candidates. Shapley values describe the contributions of each descriptor across each possible coalition of descriptors, providing an estimate of each descriptor's importance.

To investigate the feature importance, the Shapley values were determined using KNIME's native Shapley values loop feature.[72] As it was the most robust model, the XGBoost was chosen for investigation. The high robustness metrics indicate that the relationship modelled by the XGBoost method may be more insightful than other, less robust models. It is important to note that because each model learns from the data differently, other models may identify features that the XGBoost model cannot. Using the test data, the training data can be perturbed to assess how changes in the dataset affect predictions. This allows each prediction to be explained and quantifies each feature's contribution. While identifying how each prediction is made, examining trends in prediction importance across descriptors yields much more useful information for identifying the characteristics that govern the property of interest. As discussed previously, Shapley feature importance can be easily visualised as beeswarm plots, where the Shapley values of important features will clearly separate descriptor values. i.e., where the presence of a descriptor is correlated with a positive Shapley value, this indicates that an increase in the descriptor (that being the inclusion of the structure represented by a given fingerprint bit, or an increase in the value of a solvent or physicochemical descriptor) contributes to a "good scattering" classification, and *vice versa*.

While an investigation of the Shapley values for the solubility models in Chapter 3 was inconclusive, the models described here yielded plots with differing levels of variability. While the Shapley values across many descriptors were near zero, indicating very little contribution to the model, several models afforded well-defined importance relationships. Subsequently, Shapley values were used to identify the most important features of each XGBoost model (Figure 4.19, Appendix 4.11).

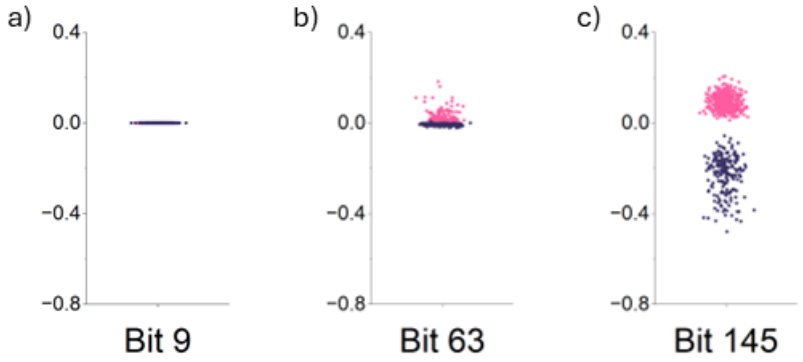


Figure 4.19. Beeswarm plot of Shapley values for "all solutions" XGBoost model. Shapley values identify the importance of descriptors. Examples of descriptors with a) no importance, b) reasonable importance, and c) significant importance to the final prediction.

We note, however, that as the dataset size decreases, these relationships become significantly weaker. As such, to ensure that appropriate relationships between structure and property are found, further work must be undertaken to expand QSPR datasets for supramolecular applications, if these materials are to be better designed (Figure 4.20).

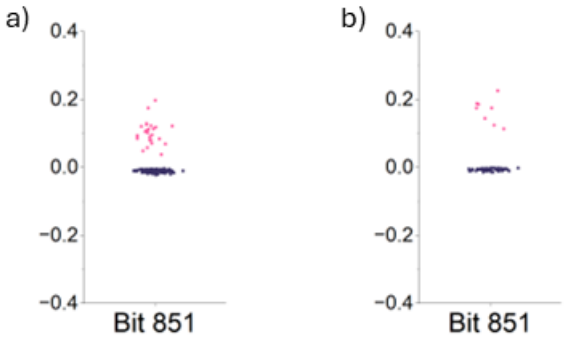


Figure 4.20. Beeswarm plot of Shapley values for Bit 851 for the a) "single component" and b) "H_2O" XGBoost models. As the dataset size decreases, the density of Shapley plots also decreases, making it less reliable for identifying important features.

It is known that changes in the preparation of solutions of self-assembling molecules can have a large impact on the resulting assembly.[73] These include changes to both the solvent system and the solute concentration. It is therefore unsurprising that the model consistently identified such descriptors throughout the training process.[74]

All three models highlight the importance of aromatic systems, particularly perylenes, for the "all solution" and "single component" models, with Bit 145 and 1001 being among the most important fingerprint descriptors in those models, respectively. Bits 145 and 1001 both represent core substructures unique to the perylene bisimide molecules. While the H_2O solutions model does not contain an analogous fingerprint bit of high importance, the number of rings is shown to be an important physicochemical parameter in its prediction.

As observed at the start of the QSPR procedure, while the physicochemical parameters alone are insufficient to derive a complete predictive relationship, some split the classes well, resulting in their inclusion as some of the most important descriptors, suggesting that the physicochemical parameters are effective at capturing the general trends, with the fingerprints better able to highlight specific structural features that are outliers to these general trends. In general, the Shapley values of the physicochemical parameters indicate that solutions of larger solutes with more polar structural features exhibit better scattering. This is a rational outcome, as the molecules in this dataset are those with increased aromatic character or those better able to utilise polar interactions for aggregation. And with increased aggregation, the scattering quality is expected to increase.

As discussed previously, one of the significant advantages of Shapley values is that it not only highlights the important features but also allows features which specifically promote or inhibit the property of interest. This allows features that simpler models would highlight as important to be better understood. While chemical intuition may suggest that including such features will promote the property of interest, without understanding how a descriptor shifts a prediction, incorrect relationships may be established. For example, the physicochemical parameters indicate that increased polarity and aromaticity afford solutions with better scatter. As well as this, Bits 289 and 310, corresponding to substructures of the tyrosine and levodopa amino acid side chains, are shown to be important structural features. Without further understanding, one may conclude that the polar hydroxyl groups and aromatic ring, therefore, promote scattering. However, the Shapley values for both systems are negative, indicating that the presence of these features actually diminishes the scattering quality. Indeed, an investigation of the data shows that several multicomponent solutions with PBI-Y (**33**) and NDI-YI (**49**) do not afford as effective aggregates as other rylenes (Figure 4.21).

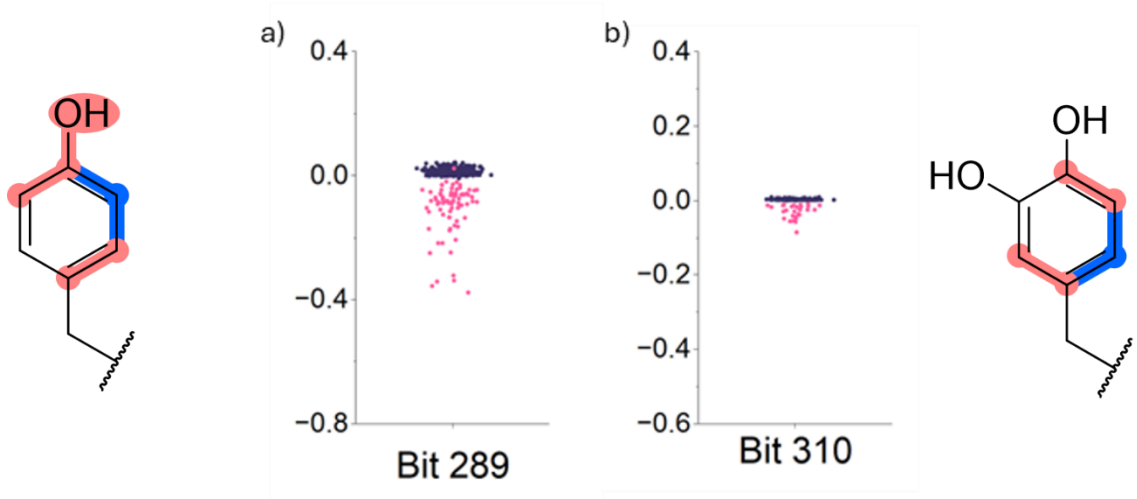


Figure 4.21. Negative Shapley value relationships indicate the presence of a substructure disrupting the scattering. a) Bit 289 in the "all solutions" model representing the tyrosine substructure and b) Bit 310 in the "single component" model representing the levodopa substructure. The core atom represented by the fingerprint bit of interest is highlighted in blue, with the extended connectivity captured by the structural descriptor highlighted in red.

Conversely, Bit 119 indicates that the presence of long alkyl chains in the asymmetric hexyl NDIs (**43**–**45**) enhances scattering, suggesting that these structures may promote molecular aggregation. It is important to consider, however, that such systems lack true chemical understanding. Therefore, it is important to consider why such features have been selected. While the long-chain alkyl structure may directly afford aggregation, such structures are only present on these asymmetrically functionalised molecules. Subsequently, this fingerprint may serve as a proxy for the asymmetric nature of these molecules. Both conclusions offer valuable insight into potential chemical structures that may afford good scattering classifications, and therefore, a propensity to self-assemble.

4.3 Conclusions

In conclusion, we have demonstrated that SAXS scattering quality, as a proxy for the presence of self-assembled aggregates in solution, can be modelled using QSPR models. We have shown that investigating physicochemical descriptions alone is insufficient to elucidate the complexity governing self-assembly, and that more robust machine-learning approaches are required to model how the nuances of chemical structure affect bulk properties. The open-source framework we have developed provides reliable scattering-quality predictions of scattering quality. We have demonstrated the adaptability of our QSPR framework to tackle prediction problems across the hierarchical scales of self-assembly. While the robust results show that overfitting has been reduced, the high confidence in incorrect validation predictions indicates that further work to reduce model overfitting and to implement additional applicability-domain metrics could yield an even more predictive system. In particular we show that XGBoost ML models yield high performance metrics and exceptional robustness metrics, suggesting that a truly effective structure-property relationship has been established with this model. Using Shapley values, we can better probe the models' dependence on one another to draw useful interpretations that could aid the design of future materials.

We demonstrate how the use of different machine learning models could be used in tandem to narrow the scope of experimental investigation, which would hopefully allow for the targeted investigation of candidates more likely to be effective in the application and design of functional materials. Firstly, given the need for molecules to sufficiently dissolve to enable their kinetic self-assembly, the large chemical space of possible candidates can be narrowed by screening those predicted to be insoluble in a given solvent. We can then take a reduced pool of candidates and screen out those that are predicted to not form self-assembled aggregates, leaving only a dataset of candidate solute-solvent combinations that affords self-assembled materials. While current limitations in predicting aggregate morphology mean that aggregate shape cannot be predicted, the systems described here would enable targeted investigation and characterisation, allowing future QSPRs to identify the expected morphology for a given solute-solvent combination. If such predictive models are accessed, further insight could be gained, enabling the development of new and effective design rules for functional materials.

4.4 Experimental

4.4.1 Materials

All chemicals and solvents were purchased from Alfa Aesar, Fischer Scientific, Honeywell, Sigma-Aldrich, Thermo Scientific, and VWR Chemicals and used as received. Deionised water was used throughout. A comprehensive list of compounds used in this chapter is described in the Appendix. 3.1 and 4.1.

4.4.2 Solution Preparation

Single-component solutions were prepared at concentrations of 5 $mgmL^{-1}$ or 10 $mgmL^{-1}$ by dispersing the appropriate napthalenyloxy acetyl (Nap), naphthalene diimide (NDI) or perylene bisimide (PBI) derivatives in either acetone, dimethyl formamide (DMF), dimethyl sulfoxide (DMSO), ethanol, tetrahydrofuran (THF) or deionised water. Solutions prepared in deionised water were basified by the addition of one equivalent (Nap, PBI) or two equivalents (NDI) of sodium hydroxide (0.1 M, aqueous). The solutions were mixed overnight on using a magnetic stirring plate (SCILOGEX, MS-H280-Pro; IKA, RO 10) until all the solute had visibly dissolved.

Select samples were prepared in deuterated solvents D_2O, d_6-DMSO, and d_6-ethanol. Solution prepared in D_2O were basified by the addition of one equivalent of sodium deuteroxide (0.1 M, D_2O). The solutions were mixed overnight on using a magnetic stirring plate (SCILOGEX, MS-H280-Pro; IKA, RO 10) until all the solute had visibly dissolved.

Multicomponent solutions were prepared by combining the basified 10 $mgmL^{-1}$ aqueous solutions in a 1:1 ratio. Multicomponent solutions were shaken by inversion followed by agitation for 1 hour using an orbital shaker (Cleaver Scientific, CS-NOR).

4.4.3 Small Angle X-ray Scattering

Small angle X-ray scattering (SAXS) measurements were performed using the I22 instrument (Diamond Light Source, Didcot, UK) under experiment SM40119.

Measurements on the I22 instrument were performed using a monochromatic beam of 12.4 eV corresponding to 1 Å fixed wavelength. A Pilatus P3-2M sample detector placed at a distance of 5.72 m allowed a q-range of approximately 0.003 – 0.3 $Å^{-1}$ to be characterised. Solutions were measured in 1.56 mm diameter borosilicate glass capillaries (Vitrex Medical A/S, 112213). Measurements were run at 25°C. Solutions were prepared as described in Section 4.4.2.

The 2D scattering data were reduced to 1D via azimuthal integration using the fast azimuthal integration Python library[75] (PyFAI) and data analysis workbench[76] (DAWN) software. The scattering from acetone, DMF, DMSO, ethanol, THF, and water was also measured and subtracted from the data using a custom Python script. The script would calculate the difference between the samples I(q) and the appropriate background sample (multiplied by a given scalar) to obtain the background-subtracted I(q). Furthermore, the background-corrected error in I(q) was determined by the Euclidean distance between the sample and background errors (Equation 4.5).

$$Error_{Corrected} = \sqrt{Error_{Sample}^2 + Error_{Background}^2} \qquad \text{(Equation 4.5)}$$

The resulting data were instrument-independent 1D scattering curves for each solution of intensity vs q. The scattering length density (SLD) of each material was calculated using the National Institute of Standards and Technology's neutron activation and scattering calculator (Table 4.13).[32]

Table 4.13. X-ray scattering length density of materials and solvents characterised by SAXS.

Sample	**Scattering Length Density (× 10^{-6} Å^{-2})**
Acetone	7.41
DMF	8.85
d_6-DMF	8.82
DMSO	10.14
d_6-DMSO	10.18
Ethanol	7.59
d_6-Ethanol	7.58
THF	7.27
H2O	9.44
D2O	9.46
Nap / NDI / PBI materials	13.54 – 14.137

4.5 References

1. N. Singh, N. Singh and A. P. Singh, *International Journal of Pharmaceutical Chemistry and Analysis*, 2021, **7**, 166–171.
2. R. Lundy, E. R. Draper and J. J. Walsh, *New Journal of Chemistry*, 2018, **42**, 19020–19025.
3. V. Adams, J. Cameron, M. Wallace and E. R. Draper, *Chemistry – A European Journal*, 2020, **26**, 9879–9882.
4. R. E. Ginesi, N. R. Murray, R. M. Dalgliesh, J. Doutch and E. R. Draper, *Chemistry – A European Journal*, 2023, **29**, e202301042.
5. E. R. Draper and D. J. Adams, *Chem*, 2017, **3**, 390–410.
6. C. Naranjo, A. Doncel-Giménez, R. Gómez, J. Aragó, E. Ortí and L. Sánchez, *Chemical Science*, 2023, **14**, 9900–9909.
7. M. M. Safont-Sempere, G. Fernández and F. Würthner, *Chemical Reviews*, 2011, **111**, 5784–5814.
8. E. R. Draper, J. R. Lee, M. Wallace, F. Jäckel, A. J. Cowan and D. J. Adams, *Chemical Science*, 2016, **7**, 6499–6505.
9. K. Nakamura, R. Kubota, T. Aoyama, K. Urayama and I. Hamachi, *Nat Commun*, 2023, **14**, 1696.
10. S. Datta, H. Itabashi, T. Saito and S. Yagai, *Nature Chemistry*, 2025, **17**, 477–492.
11. M. C. Jones, L. P. Cavalcanti, G. N. Smith, R. Dalgliesh, S. R. Euston, F. Vilela, V. Arrighi and G. O. Lloyd, *Communications Chemistry*, 2025, **8**, 265.
12. E. R. Draper and D. J. Adams, *Chemical Society Reviews*, 2018, **47**, 3395–3405.
13. R. I. Randle, R. E. Ginesi, O. Matsarskaia, R. Schweins and E. R. Draper, *Macromolecular Rapid Communications*, 2023, **44**, 2200709.
14. L. Li, R. Sun and R. Zheng, *Materials & Design*, 2021, **197**, 109209.
15. R. G. Weiss and P. Terech, *Molecular gels: materials with self-assembled fibrillar networks*, Springer, Dordrecht, 2006.
16. G. Yu, X. Yan, C. Han and F. Huang, *Chemical Society Reviews*, 2013, **42**, 6697–6722.
17. V. S. Raghuwanshi and G. Garnier, *Advances in Colloid and Interface Science*, 2019, **274**, 102044.
18. C. A. Schneider, W. S. Rasband and K. W. Eliceiri, *Nature Methods*, 2012, **9**, 671–675.
19. R. Aston, K. Sewell, T. Klein, G. Lawrie and L. Grøndahl, *European Polymer Journal*, 2016, **82**, 1–15.
20. R. Jacobs, M. Shen, Y. Liu, W. Hao, X. Li, R. He, J. R. C. Greaves, D. Wang, Z. Xie, Z. Huang, C. Wang, K. G. Field and D. Morgan, *Cell Reports Physical Science*, 2022, **3**, 100876.
21. C. R. M. MacDonald and E. R. Draper, *Beilstein Journal of Organic Chemistry*, 2024, **20**, 2608–2634.
22. B. J. Inkson, in *Materials Characterization Using Nondestructive Evaluation (NDE) Methods*, eds. G. Hübschen, I. Altpeter, R. Tschuncky and H.-G. Herrmann, Woodhead Publishing, 2016, DOI: https://doi.org/10.1016/B978-0-08-100040-3.00002-X, pp. 17–43.
23. H. Jin, C. Ma, W. Wang, Y. Cai, J. Qi, T. Wu, P. Liao, H. Li, Q. Zeng, M. Xie, J. Huang and Y. Yan, *ACS Materials Letters*, 2022, **4**, 145–152.
24. L. L. E. Mears, E. R. Draper, A. M. Castilla, H. Su, Zhuola, B. Dietrich, M. C. Nolan, G. N. Smith, J. Doutch, S. Rogers, R. Akhtar, H. Cui and D. J. Adams, *Biomacromolecules*, 2017, **18**, 3531–3540.
25. D. J. Adams, *Journal of the American Chemical Society*, 2022, **144**, 11047–11053.

26. D. Katrantzi, S. Micklethwaite, N. Hondow, A. Brown and L. Dougan, *Faraday Discussions*, 2025, **260**, 55–81.
27. P. Terech, *Croatica chemica acta*, 1992, **65**.
28. *Soft Matter Characterization*, Springer Netherlands, 1 edn., 2008.
29. D. McDowall, D. J. Adams and A. M. Seddon, *Soft Matter*, 2022, **18**, 1577–1590.
30. C. R. M. MacDonald and E. R. Draper, *Beilstein J Org Chem*, 2024, **20**, 2608–2634.
31. D. McDowall, B. J. Greeves, R. Clowes, K. McAulay, A. M. Fuentes-Caparrós, L. Thomson, N. Khunti, N. Cowieson, M. C. Nolan, M. Wallace, A. I. Cooper, E. R. Draper, A. J. Cowan and D. J. Adams, *Advanced Energy Materials*, 2020, **10**, 2002469.
32. National Institute of Standards and Technology: Neutron Activation and Scattering Calculator, https://www.ncnr.nist.gov/resources/activation/, (accessed 15th October 2025).
33. C. J. Gommes, S. Jaksch and H. Frielinghaus, *Journal of Applied Crystallography*, 2021, **54**, 1832–1843.
34. A. Tropsha, *Molecular Informatics*, 2010, **29**, 476–488.
35. B. Khalid, H. Ghorab and A. Benkhemissa, *Journal of the Indian Chemical Society*, 2022, **99**, 100672.
36. D. Fourches, E. Muratov and A. Tropsha, *J Chem Inf Model*, 2010, **50**, 1189–1204.
37. D. Young, T. Martin, R. Venkatapathy and P. Harten, *QSAR & Combinatorial Science*, 2008, **27**, 1337–1345.
38. P. Agarwal and D. B. Searls, *Briefings in Bioinformatics*, 2008, **9**, 479–492.
39. M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, *Chemical Society Reviews*, 2025, **54**, 1125–1150.
40. F. Song, S. Parekh, L. Hooper, Y. K. Loke, J. Ryder, A. J. Sutton, C. Hing, C. S. Kwok, C. Pang and I. Harvey, 2010, **14**, 8.
41. B. Duyx, M. J. E. Urlings, G. M. H. Swaen, L. M. Bouter and M. P. Zeegers, *Journal of Clinical Epidemiology*, 2017, **88**, 92–101.
42. C. Colquhoun, E. R. Draper, E. G. B. Eden, B. N. Cattoz, K. L. Morris, L. Chen, T. O. McDonald, A. E. Terry, P. C. Griffiths, L. C. Serpell and D. J. Adams, *Nanoscale*, 2014, **6**, 13719–13725.
43. J. C. Dearden, M. T. D. Cronin and K. L. E. Kaiser, *SAR and QSAR in Environmental Research*, 2009, **20**, 241–266.
44. A. G. Kikhney and D. I. Svergun, *FEBS Letters*, 2015, **589**, 2570–2577.
45. A. R. Hirst, I. A. Coates, T. R. Boucheteau, J. F. Miravet, B. Escuder, V. Castelletto, I. W. Hamley and D. K. Smith, *Journal of the American Chemical Society*, 2008, **130**, 9113–9121.
46. e. a. J. Krzywon, *Journal*, DOI: 10.5281/zenodo.18352296.
47. G. Roberts, M.-P. Nieh, A. W. K. Ma and Q. Yang, *Digital Discovery*, 2025, **4**, 1467–1477.
48. E. R. Draper, H. Su, C. Brasnett, R. J. Poole, S. Rogers, H. Cui, A. Seddon and D. J. Adams, *Angewandte Chemie*, 2017, **129**, 10603–10606.
49. E. R. Draper, B. Dietrich, K. McAulay, C. Brasnett, H. Abdizadeh, I. Patmanidis, S. J. Marrink, H. Su, H. Cui, R. Schweins, A. Seddon and D. J. Adams, *Matter*, 2020, **2**, 764–778.
50. R. R. Sonani, S. Bianco, B. Dietrich, J. Doutch, E. R. Draper, D. J. Adams and E. H. Egelman, *Cell Reports Physical Science*, 2024, **5**, 101812.
51. R. E. Ginesi and E. R. Draper, *Soft Matter*, 2024, **20**, 3887–3896.
52. T. R. Canrinus, F. J. R. Cerpentier, B. L. Feringa and W. R. Browne, *Chemical Communications*, 2017, **53**, 1719–1722.
53. K. McAulay, H. Wang, A. M. Fuentes-Caparrós, L. Thomson, N. Khunti, N. Cowieson, H. Cui, A. Seddon and D. J. Adams, *Langmuir*, 2020, **36**, 8626–8631.

54. E. R. Draper, B. J. Greeves, M. Barrow, R. Schweins, M. A. Zwijnenburg and D. J. Adams, *Chem*, 2017, **2**, 716–731.
55. S. M. Sedlak, L. K. Bruetzel and J. Lipfert, *J Appl Crystallogr*, 2017, **50**, 621–630.
56. N. Monge, M.-R. Amini and A. Deschamps, *Acta Crystallographica Section A*, 2024, **80**, 405–413.
57. G. Jancsó, in *Handbook of Nuclear Chemistry*, eds. A. Vértes, S. Nagy, Z. Klencsár, R. G. Lovas and F. Rösch, Springer US, Boston, MA, 2011, DOI: 10.1007/978-1-4419-0720-2_15, pp. 699–725.
58. C. Reichardt and T. Welton, *Solvents and solvent effects in organic chemistry*, John Wiley & Sons, 2011.
59. A. R. Katritzky, D. C. Fara, H. Yang, K. Tämm, T. Tamm and M. Karelson, *Chemical Reviews*, 2004, **104**, 175–198.
60. F. Shakerizadeh-Shirazi, B. Hemmateenejad and A. M. Mehranpour, *Analytical Methods*, 2013, **5**, 891–896.
61. C. Reichardt, *Chemical Society Reviews*, 1992, **21**, 147–153.
62. S. Spange, N. Weiß, C. H. Schmidt and K. Schreiter, *Chemistry–Methods*, 2021, **1**, 42–60.
63. A. Golbraikh, E. Muratov, D. Fourches and A. Tropsha, *Journal of Chemical Information and Modeling*, 2014, **54**, 1–4.
64. F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni and R. Todeschini, *Molecules*, 2012, **17**, 4791–4810.
65. C. Spearman, *International Journal of Epidemiology*, 2010, **39**, 1137–1150.
66. J. D. Hedengren, Random Forest, https://apmonitor.com/pds/index.php/Main/RandomForest, (accessed 25th March 2026).
67. RBF SVM parameters, https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html, (accessed 26th March 2026).
68. Random Forest Hyperparameters & Tuning Strategies, https://baekholab.com/2025/03/22/part-6-random-forest-hyperparameters-tuning-strategies/, (accessed 25th March 2026).
69. P. Banerjee, A Guide on XGBoost hyperparameters tuning, https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning, (accessed 25th March 2026).
70. M. Langevin, C. Grebner, S. Güssregen, S. Sauer, Y. Li, H. Matter and M. Bianciotto, *ACS Omega*, 2023, **8**, 23148–23167.
71. J. D. Simpson, L. Thomson, C. M. Woodley, C. M. Wallace, B. Dietrich, A. S. Loch, D. J. Adams and N. G. Berry, *Advanced Materials*, 2025, **37**, 2415031.
72. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel and B. Wiswedel, Berlin, Heidelberg, 2008.
73. X. Yang, X. Xu and H.-F. Ji, *The Journal of Physical Chemistry B*, 2008, **112**, 7196–7202.
74. L. Thomson, D. McDowall, L. Marshall, O. Marshall, H. Ng, W. J. A. Homer, D. Ghosh, W. Liu, A. M. Squires, E. Theodosiou, P. D. Topham, L. C. Serpell, R. J. Poole, A. Seddon and D. J. Adams, *ACS Nano*, 2022, **16**, 20497–20509.
75. J. Kieffer, J. Orlans, N. Coquelle, S. Debionne, S. Basu, A. Homs, G. Santoni and D. De Sanctis, *Journal of Applied Crystallography*, 2025, **58**, 138–153.
76. M. Basham, J. Filik, M. T. Wharmby, P. C. Y. Chang, B. El Kassaby, M. Gerring, J. Aishima, K. Levik, B. C. A. Pulford, I. Sikharulidze, D. Sneddon, M. Webber, S. S. Dhesi, F. Maccherozzi, O. Svensson, S. Brockhauser, G. Naray and A. W. Ashton, *Journal of Synchrotron Radiation*, 2015, **22**, 853–858.

Chapter 5: Conclusions and Future Work

5.1 Conclusions

Across these projects, we have discussed the development and application of QSPRs to address the large gaps in understanding and prediction in the preparation of smart materials using small-molecule supramolecular systems. We have taken care to design and develop these systems with a 'traditional' bench chemist in mind, to provide accessible, functional systems. We highlight throughout that the use of KNIME provides significant advantages in accessibility for end-user applications and data interrogation. It is important to highlight that accessibility was also key to the significant progress made in developing these systems. With no background in programming, the author has successfully designed and implemented systems that would likely have been unachievable with traditional programming methods. It is worth noting that the collaborative nature of cheminformatics scientists, particularly those who the open-source, accessible KNIME platform has fostered, was instrumental in the author's development and implementation of machine learning system design.

At the outset, we ambitiously (and perhaps naively) planned to build systems capable of probing each length scale of the hierarchical assembly involved in supramolecular assembly, from molecular-scale interactions to bulk properties. It is clear to us now that each modellable endpoint has a wealth of considerations that must be tackled to ensure a meaningful relationship is established. We were able to tie molecular structure to molecular-scale solvation interactions and then later to insights into nano- and microscale assemblies, further relationships that describe interesting bulk properties remain to be realised. What is clear is that the development of small, targeted, predictive systems holds great promise, especially when used in tandem to build a more complete picture of supramolecular systems.

We have shown how each of these systems addresses QSPR development in accordance with the OECD guidelines, ensuring that robust and meaningful models are developed and assessed. Each of the final systems afforded reasonable predictive ability, but it is clear that limitations in the available data play a significant role in the performance achieved. We show that, even with limited data, effective local investigations and predictive interpolations are achievable: predictions across organic solvents can now be made with useful confidence metrics, as well as predictions of the presence of well-defined self-assembled aggregates. The latter models provide useful chemical insights through Shapley feature importance metrics, identifying specific substructures that may disrupt self-assembly. It is necessary to highlight the need for more data to develop systems applicable to a broader range of chemical systems. However, a significant advantage of enabling experienced chemists to develop their

own QSPR systems is that it forces them to appreciate firsthand the limitations and difficulties of developing both the computational system and the chemical datasets necessary to train it. By unifying experience in both fields, data collection workflows can be better optimised from a QSPR standpoint, by enabling the user to consider the most effective endpoints for the computational model, but also from a chemical standpoint, allowing the user to guide the investigation to more realistic chemical spaces and providing more informed chemical understanding throughout the development process.

While we have applied several metrics to assess and ensure that the most effective models were carried forward, we highlight throughout that there is no one-size-fits-all approach to many parts of the QSPR process, ranging from the machine learning models utilised, the approach to optimising those models, to the metrics used to assess their performance. For each approach, several others could have been taken, with unknown effects on the resulting models. However, several changes can be made to ensure that future iterations of the QSPR workflow produce superior models, as highlighted throughout the discussions. In particular, we do not address the application of other partitioning approaches that could provide better metrics for extrapolative performance, the impact of tuning SMOTE or exploring more advanced alternatives, the effectiveness of Bayesian hyperparameter optimisation on our larger datasets, or the application of more complex and informative applicability-domain approaches. However, as we learned by tackling these problems, developing effective systems is a continuous process. Not only do the chemical insights afforded by each model inform the predictive limitations that need to be tackled in the next iteration, but so too does each computation problem that arises lead to a more efficient, robust computational system.

One of the most pivotal points made throughout this work is the need for a significant improvement in the standardisation of characterisation techniques and their subsequent reporting. Although this is necessary for QSPR of any problem, the point is particularly salient for supramolecular chemists, given the current lack of standards. We hope that more complex and insightful QSPRs can be developed to tackle the lack of design principles in supramolecular chemistry; however, it is clear that the current state of the literature is not amenable to QSPR applications, which is why a data mining approach was not used. We show that, even with robust knowledge and domain expertise in organic synthetic chemistry, developing a sufficient library is no small task; significant experimental effort yielded only a small expansion of the available molecules in this project, highlighting the limited amount of data any one group can generate. This suggests that it may take a unified supramolecular community to (self-)assemble the most effective standardised measurements that will afford

the greatest insight into supramolecular design. However, instituting such standards would require significant collaboration and collective effort, though the potential benefits they could afford are, in our view, certainly worth the effort.

5.2 Future Work

We have demonstrated that simpler relationships can be found that tie molecular and process-dependent descriptors to the presence of well-defined self-assembled aggregates, but we plan to begin developing a larger dataset of topological parameters that could be used to establish more complex and insightful relationships that will finally bridge the predictive gap between molecular structure, self-assembled morphology, and bulk property.

The dataset collected in Chapter 4 provides a perfect starting point for more informative model training about the nano- and microscale morphology. By fitting topological models to the scattering profiles, the resulting parameters provide a useful endpoint for modelling, hopefully allowing QSPRs to link molecular structure and solution preparation to these topologies. Ultimately, we hope to address the lack of prediction tools for bulk properties of supramolecular materials. By measuring and cataloguing the useful properties of each preparation in the resulting materials, we foresee that a QSPR could be built to establish relationships with the bulk properties of interest, utilising descriptors of solution preparation and the self-assembled structures it yields.

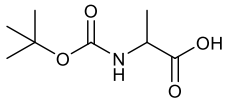
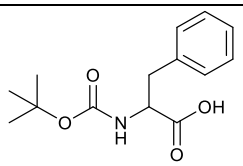
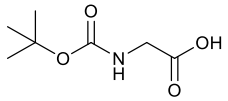
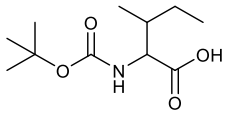
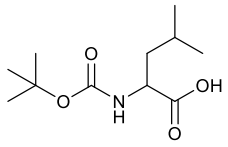
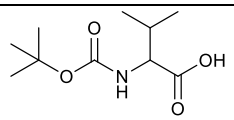
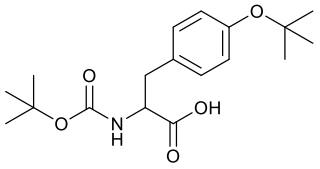
It is clear from our concluding remarks that any future work must be undertaken with reproducibility and data accessibility in mind, and we hope that we can continue contributing to the discourse of standardised supramolecular characterisation and material design.

Appendices

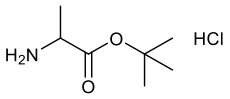
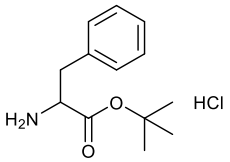
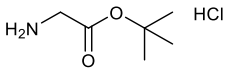
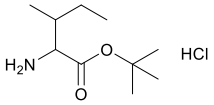
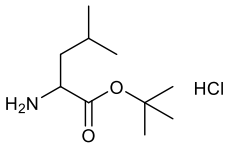
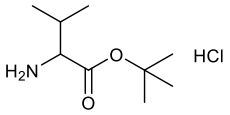
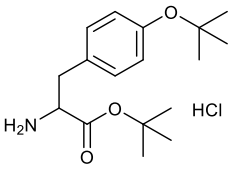
Chapter 3 Appendices

A 3.1 Synthetic Characterisations

A 3.1.1 Dipeptide Synthesis Boc Reagents

Compound Number	Name	Structure
A1	Boc-A	
A2	Boc-F	
A3	Boc-G	
A4	Boc-I	
A5	Boc-L	
A6	Boc-V	
A7	Boc-Y-^{t}Bu	

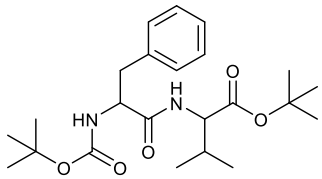
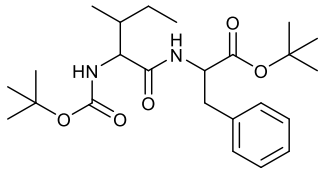
A 3.1.2 Dipeptide Synthesis O^{t}Bu Reagents

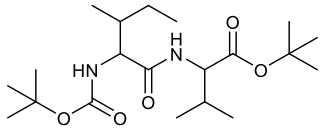
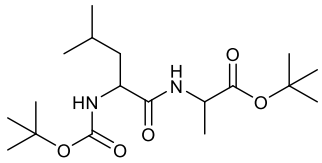
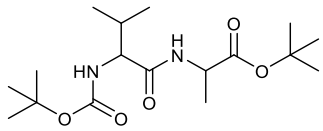
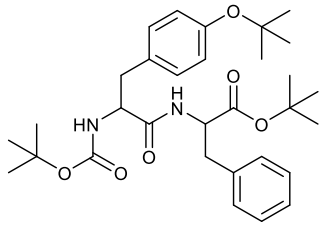
Compound Number	Name	Structure
B1	A-O^{t}Bu·HCl	
B2	F-O^{t}Bu·HCl	
B3	G-O^{t}Bu·HCl	
B4	I-O^{t}Bu·HCl	
B5	L-O^{t}Bu·HCl	
B6	V-O^{t}Bu·HCl	
B7	Y-O^{t}Bu·HCl	

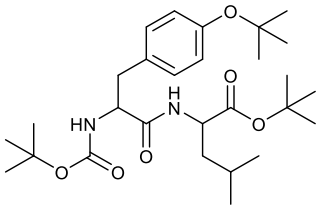
A 3.1.3 Protected Dipeptide Products

^{1}H and ^{13}C NMR spectra were recorded on a Bruker Avance III 400 MHz spectrometer. Chemical shifts (δ) are reported in parts per million (ppm) to the residual solvent peak (DMSO-d_6: 2.50 ppm for ^{1}H and 39.52 ppm for ^{13}C). Coupling constants (*J*) are given to the nearest 0.1 Hertz. Peak multiplicities are reported as singlet (s), doublet (d), triplet (t), multiplet (m), and/or broad (b). Peak assignments are reported as: H/C_{Ar} = aromatic H or C. Amino acids are listed as A: alanine; F: phenylalanine; G: glycine; I: isoleucine; L: leucine; V: valine; Y: tyrosine. High resolution accurate mass measurements (HRMS) were performed on an Agilent 6546 LC/QTOF (quadrupole-time of flight) instrument with an Agilent ESI or APCI source.

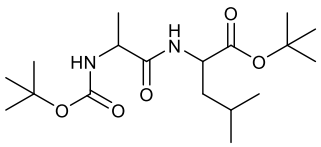
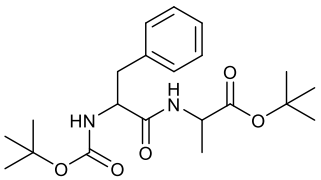
Synthesised and Characterised by Connor MacDonald

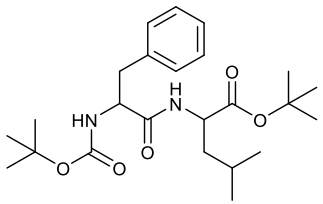
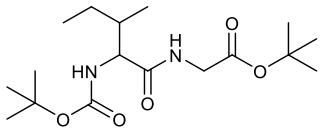
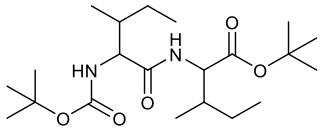
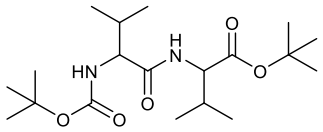
Compound Number	Name	Structure	Characterisation
C1	Boc-FV-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 7.94 (1H, d, *J* 8.3, N<u>H</u>), 7.43 – 7.11 (5H, m, <u>H</u>$_{ArF}$), 6.94 (1H, d, *J* 8.7, N<u>H</u>), 4.27 (1H, td, *J* 10.3, 3.8, NH-C<u>H</u>$_F$), 4.10 (1H, dd, *J* 8.3, 5.8, NH-C<u>H</u>$_V$), 2.97 (1H, dd, *J* 13.8, 3.8, C<u>H</u>$_{2F}$), 2.74 (1H, dd, *J* 13.7, 10.6, C<u>H</u>$_{2F}$), 2.05 (1H, dt, *J* 13.3, 6.6, C<u>H</u>$_V$), 1.42 (9H, s, C(CH$_3$)$_3$), 1.30 (9H, s, C(CH$_3$)$_3$), 0.91 (6H, t, *J* 6.5, C<u>H</u>$_{3V}$). δ_C (101 MHz, DMSO-d_6) 171.9, 170.4, and 155.2 (<u>C</u>=O), 138.1, 129.2, 127.9, and 126.1 (<u>C</u>$_{ArF}$), 80.7 and 78.0 (<u>C</u>(CH$_3$)$_3$), 57.7 (NH-<u>C</u>H$_F$), 55.4 (NH-<u>C</u>H$_V$), 37.3 (<u>C</u>H$_{2F}$), 30.2 (<u>C</u>H$_V$), 28.1 and 27.6 (C(CH$_3$)$_3$), 18.8, 17.9 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 443.2516, found 443.2524.
C2	Boc-IF-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.15 (1H, d, *J* 7.6, N<u>H</u>), 7.30 – 7.15 (5H, m, <u>H</u>$_{ArF}$), 6.59 (1H, d, *J* 9.3, N<u>H</u>), 4.46 – 4.02 (1H, m, *J* 7.6, NH-C<u>H</u>$_F$), 3.84 (1H, t, *J* 8.4, NH-C<u>H</u>$_I$), 3.01 – 2.84 (2H, m, *J* 13.9, 7.4, C<u>H</u>$_{2F}$), 1.68 – 1.53 (1H, m, C<u>H</u>$_I$), 1.41 – 1.27 (19H, m, C(CH$_3$)$_3$ overlapping C<u>H</u>$_{2I}$), 1.12 – 0.96 (1H, m, C<u>H</u>$_{2I}$), 0.88 – 0.72 (6H, m, C<u>H</u>$_{3I}$). δ_C (101 MHz, DMSO-d_6) 171.7, 170.9, and 155.6 (<u>C</u>=O), 137.6, 129.6, 128.5, and 126.9 (C$_{ArF}$), 81.0 and 78.4 (<u>C</u>(CH$_3$)$_3$), 59.0 (NH-<u>C</u>H$_F$), 54.4 (NH-<u>C</u>H$_I$), 37.5 (<u>C</u>H$_{2F}$), 37.2 (<u>C</u>H$_I$), 28.6 and 27.9 (C(CH$_3$)$_3$), 24.7 (<u>C</u>H$_{2I}$), 15.7 and 11.4 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 457.2673, found 457.2673.

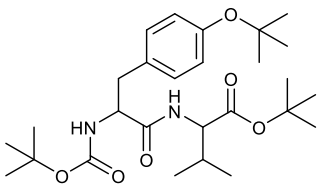
C3	Boc-IV-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 7.84 (1H, d, *J* 8.3, N<u>H</u>), 6.75 (1H, d, *J* 9.2, N<u>H</u>), 4.06 (1H, dd, *J* 8.3, 6.3, NH-C<u>H</u>$_V$), 3.88 (1H, t, *J* 8.2, NH-C<u>H</u>$_L$), 2.10 – 1.91 (1H, m, C<u>H</u>$_L$), 1.66 (1H, s, C<u>H</u>$_V$), 1.38 (19H, d, *J* 9.6, C(C<u>H</u>$_3$)$_3$ overlapping C<u>H</u>$_{2I}$), 1.10 (1H, s, C<u>H</u>$_{2I}$), 0.94 – 0.74 (12H, m, C<u>H</u>$_{3I}$ and C<u>H</u>$_{3L}$). δ_C (101 MHz, DMSO-d_6) 171.6, 170.4, and 155.3 (<u>C</u>=O), 80.5 and 77.9 (<u>C</u>(CH$_3$)$_3$), 58.6 (NH-<u>C</u>H$_I$), 57.7 (NH-<u>C</u>H$_V$), 36.3 (<u>C</u>H$_V$), 30.1 (CH$_L$), 28.1 and 27.6 (C(CH$_3$)$_3$), 24.4 (<u>C</u>H$_{2I}$), 18.8 and 18.0 (<u>C</u>H$_{3V}$), 15.3 and 10.8 (<u>C</u>H$_{3L}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 409.2673, found409.2677.
C4	Boc-LA-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.07 (1H, d, *J* 7.0, N<u>H</u>), 6.80 (1H, d, *J* 8.6, N<u>H</u>), 4.09 (1H, dq, *J* 7.2, 7.2, NH-C<u>H</u>$_A$), 3.98 (1H, td, *J* 8.9, 6.0, NH-C<u>H</u>$_L$), 1.62 (1H, dt, *J* 13.5, 6.8, C<u>H</u>$_L$), 1.43 – 1.39 (2H, m, C<u>H</u>$_{2L}$), 1.37 (18H, d, *J* 4.5, C(CH$_3$)$_3$), 1.24 (3H, d, *J* 7.3, C<u>H</u>$_{3A}$), 0.86 (6H, dd, *J* 6.6, 5.1, C<u>H</u>$_{3V}$). δ_C (101 MHz, DMSO-d_6) 172.4, 171.7, and 155.3 (C=O), 80.2 and 77.8 (<u>C</u>(CH$_3$)$_3$); 52.3 (NH-<u>C</u>H$_A$), 48.2 (NH-<u>C</u>H$_A$); 40.9 (<u>C</u>H$_{2L}$), 28.2 and 27.6 (C(CH$_3$)$_3$), 24.2 (<u>C</u>H$_V$), 23.1 and 21.4 (<u>C</u>H$_{3V}$), 16.9 (<u>C</u>H$_{3A}$). HRMS (ESI) m/z: [M+H]+ calculated 359.2540, found 359.2550.
C5	Boc-VA-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.15 (1H, d, *J* 6.8, N<u>H</u>), 6.55 (1H, d, *J* 9.1, N<u>H</u>), 4.22 – 4.07 (1H, m, NH-C<u>H</u>$_A$), 3.88 – 3.80 (1H, m, NH-C<u>H</u>$_V$), 2.02 – 1.86 (1H, m, C<u>H</u>$_V$), 1.39 (18 H, d, *J* 2.6, C(CH$_3$)$_3$), 1.24 (3H, d, *J* 7.3, C<u>H</u>$_{3A}$), 0.88 (3H, d, *J* 6.7, C<u>H</u>$_{3V}$), 0.83 (3H, d, *J* 6.7, C<u>H</u>$_{3V}$). δ_C (101 MHz, DMSO-d_6) 171.6, 171.0, and 155.3 (<u>C</u>=O), 80.3 and 77.9 (<u>C</u>(CH$_3$)$_3$), 59.0 (NH-<u>C</u>H$_V$), 48.2 (NH-<u>C</u>H$_A$), 30.6 (<u>C</u>H$_V$), 28.1 and 27.6 (C(CH$_3$)$_3$), 19.2 (<u>C</u>H$_{3A}$), 18.0 and 17.0 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 367.2203, found 367.2230.
C6	Boc-YF-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.20 (1H, d, *J* 7.5, N<u>H</u>), 7.34 – 7.18 (5H, m, <u>H</u>$_{ArF}$), 7.14 (2H, d, *J* 8.4, <u>H</u>$_{ArY}$), 6.85 (3H, dd, *J* 8.5, 4.6, <u>H</u>$_{ArY}$ overlapping N<u>H</u>), 4.45 – 4.28 (1H, m, NH-C<u>H</u>$_Y$), 4.17 (1H, td, *J* 10.7, 3.7, NH-C<u>H</u>$_F$), 3.06 – 2.91 (2H, m, C<u>H</u>$_{2Y}$), 2.87 (1H, dd, *J* 13.8, 3.6, C<u>H</u>$_{2F}$), 2.63 (1H, dd, *J* 13.7, 10.9, C<u>H</u>$_{2F}$), 1.35 – 1.13 (27H, m, C(CH$_3$)$_3$). δ_C (101 MHz, DMSO-d_6) 171.7, 170.4, and 155.1 (<u>C</u>=O), 153.3, 137.1, 132.7, 129.6, 129.2, 128.1, 126.5, and 123.2 (<u>C</u>$_{ArF}$ and <u>C</u>$_{ArY}$), 80.7, 77.9, and 77.5 (<u>C</u>(CH$_3$)$_3$), 55.5 (NH-<u>C</u>H$_F$), 54.0 (NH-<u>C</u>H$_Y$), 36.9 (<u>C</u>H$_{2F}$ and <u>C</u>H$_{2Y}$ overlapped), 28.5, 28.1, and 27.5 (C(CH$_3$)$_3$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 563.3092, found 5563.3085.

C7	Boc-YL-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.11 (1H, d, *J* 7.6, N<u>H</u>), 7.17 (2H, d, *J* 8.4, <u>H</u>$_{ArY}$), 6.85 (3H, d, *J* 8.2, <u>H</u>$_{ArY}$ overlapping N<u>H</u>), 4.27 – 4.02 (2H, m, C<u>H</u>$_{2Y}$ and C<u>H</u>$_{2L}$), 3.00 – 2.82 (1H, m, NH-C<u>H</u>$_Y$), 2.73 – 2.57 (1H, m, NH-C<u>H</u>$_V$), 1.65 (1H, d, *J* 6.8, C<u>H</u>$_V$), 1.56 – 1.46 (2H, m, *J* 8.5, C<u>H</u>$_{2Y}$ and C<u>H</u>$_{2L}$), 1.39 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.27 (18H, d, *J* 11.0, C(C<u>H</u>$_3$)$_3$), 0.91 (3H, d, *J* 6.5, C<u>H</u>$_{3V}$), 0.85 (3H, d, *J* 6.5, C<u>H</u>$_{3V}$). δ_C (101 MHz, DMSO-d_6) 171.8, 171.5, and 155.1 (<u>C</u>=O), 153.3, 132.7, 129.6, and 123.2 (<u>C</u>$_{ArY}$), 80.4, 77.9, and 77.5 (<u>C</u>(CH$_3$)$_3$), 55.4 (<u>C</u>H$_{2Y}$), 51.0 (<u>C</u>H$_{2L}$), 36.8 (<u>C</u>H$_{2Y}$ and <u>C</u>H$_{2L}$ overlapped), 28.5, 28.1, and 27.6 (C(CH$_3$)$_3$), 24.2 (<u>C</u>H$_V$), 22.7, 21.4 (<u>C</u>H$_3$). HRMS (ESI) m/z: $[M+H]^+$ calculated 507.3429, found 507.3441.

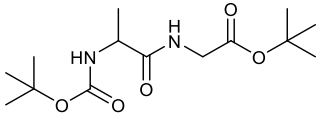
<u>Synthesised by Dr Rebecca Ginesi and Characterised by Connor MacDonald</u>

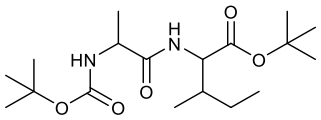
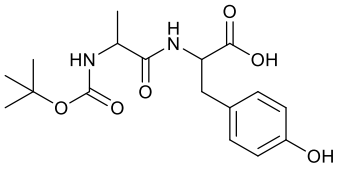
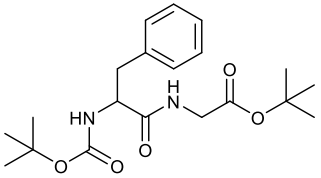
Compound Number	Name	Structure	Characterisation
C8	Boc-AL-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 7.74 (1 H, d, *J* 8.1, N<u>H</u>), 6.91 (1 H, d, *J* 7.8, N<u>H</u>), 4.09 (1 H, dd, *J* 8.2, 6.0, NH-C<u>H</u>$_L$), 4.06 – 3.98 (1 H, m, NH-C<u>H</u>$_A$), 1.74 (1 H, s, C<u>H</u>$_L$), 1.38 (19 H, m, C(C<u>H</u>$_3$)$_3$ overlapping C<u>H</u>$_{2L}$), 1.22 (1 H, dd, *J* 13.6, 8.3, C<u>H</u>$_{2L}$), 1.16 (3 H, d, *J* 7.1, C<u>H</u>$_{3A}$), 0.85 (6 H, t, *J* 6.8, C<u>H</u>$_{3V}$). δC (101 MHz, DMSO-d_6) 172.7, 170.3, and 155.0 (<u>C</u>=O), 80.5 and 78.0 (<u>C</u>(CH$_3$)$_3$), 56.6 (NH-<u>C</u>H$_L$), 49.4 (NH-<u>C</u>H$_A$), 36.7 (<u>C</u>H$_L$), 28.1 and 27.6 (C(CH$_3$)$_3$), 24.8 (<u>C</u>H$_2$), 18.0 (<u>C</u>H$_{3A}$), 15.3 and 11.2 (<u>C</u>H$_V$). HRMS (ESI) m/z: $[M+H]^+$ calculated 359.2540, found 359.2545.
C9	Boc-FA-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.27 (1H, d, *J* 7.0, N<u>H</u>), 7.32 – 7.12 (5H, m, <u>H</u>$_{ArF}$), 6.85 (1H, d, *J* 8.8, N<u>H</u>), 4.23 – 4.07 (2H, m, NH-C<u>H</u>$_F$ and NH-C<u>H</u>$_A$), 2.99 (1H, dd, *J* 13.8, 3.4, C<u>H</u>$_{2F}$), 2.71 (1 H, dd, *J* 13.7, 11.0, C<u>H</u>$_{2F}$), 1.39 (9 H, s, C(C<u>H</u>$_3$)$_3$), 1.31 – 1.16 (12 H, m, C(C<u>H</u>$_3$)$_3$ overlapping C<u>H</u>$_{3A}$). δ_C (101 MHz, DMSO-d_6) 171.8, 171.7, and 155.3 (<u>C</u>=O), 138.3, 129.2, 128.0, and 126.2 (<u>C</u>$_{ArF}$), 80.4 and 77.9 (<u>C</u>(CH$_3$)$_3$), 55.4 (NH-<u>C</u>H$_F$), 48.4 (NH-<u>C</u>H$_A$), 37.5 (<u>C</u>H$_{2F}$), 28.2 and 27.6 (C(CH$_3$)$_3$), 17.0 (<u>C</u>H$_{3A}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 415.2203, found 415.2207.

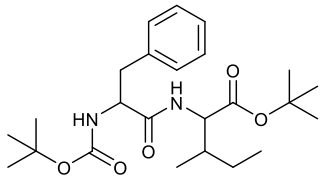
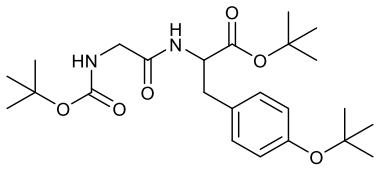
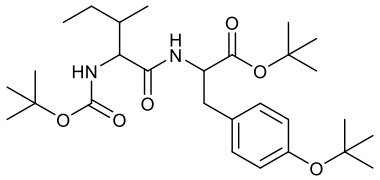
C10	Boc-FL-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.14 (1H, d, *J* 7.8, N<u>H</u>), 7.38 – 7.14 (5H, m, <u>H</u>$_{ArF}$), 6.85 (1H, d, *J* 8.8, N<u>H</u>), 4.18 (2H, tt, *J* 13.5, 6.5, NH-C<u>H</u>$_F$ and NH-C<u>H</u>$_L$), 2.96 (1H, dd, *J* 13.8, 3.8, C<u>H</u>$_{2F}$), 2.72 (1H, dd, *J* 13.7, 10.6, C<u>H</u>$_{2F}$), 1.65 (1H, dq, *J* 12.9, 6.5, C<u>H</u>$_L$), 1.51 (2H, tt, *J* 16.5, 8.0, C<u>H</u>$_{2L}$), 1.39 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.29 (9H, s, C(C<u>H</u>$_3$)$_3$), 0.91 (3H, d, *J* 6.6, C<u>H</u>$_{3L}$), 0.85 (3H, d, *J* 6.5, C<u>H</u>$_{3L}$). δ_C (101 MHz, DMSO-d_6) 171.8, 171.6, and 155.2 (<u>C</u>=O), 138.2, 129.2, 128.0, and 126.2 (<u>C</u>$_{ArF}$), 80.5 and 78.0 (<u>C</u>(CH$_3$)$_3$), 55.5 (NH-<u>C</u>H$_F$), 51.0 (NH-<u>C</u>H$_L$), 37.4 (<u>C</u>H$_{2F}$), 28.1, and 27.6 (C(CH$_3$)$_3$), 24.2 (<u>C</u>H$_L$), 22.8, 21.4 (<u>C</u>H$_{3L}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 457.2673, found 457.2665.
C11	Boc-IG-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.17 (1H, t, *J* 5.7, N<u>H</u>), 6.65 (1H, d, *J* 9.1, N<u>H</u>), 3.83 (1H, t, *J* 8.3, NH-C<u>H</u>$_I$), 3.80 – 3.54 (2H, m, NH-C<u>H</u>$_{2G}$), 1.74 – 1.58 (1H, m, C<u>H</u>$_I$), 1.44 (1H, dd, *J* 8.8, 4.8, C<u>H</u>$_{2I}$), 1.38 (18H, d, *J* 8.5, C(C<u>H</u>$_3$)$_3$), 1.16 – 1.01 (1H, m, C<u>H</u>$_{2I}$), 0.87 – 0.74 (6H, m, C<u>H</u>$_{3I}$). δ_C (101 MHz, DMSO-d_6) 171.8, 168.9, and 155.3 (<u>C</u>=O), 80.6 and 78.0 (<u>C</u>(CH$_3$)$_3$), 58.6 (NH-<u>C</u>H$_I$), 41.4 (NH-<u>C</u>H$_{2G}$), 36.6 (<u>C</u>H$_I$), 28.2 and 27.7 (C(CH$_3$)$_3$), 24.3 (<u>C</u>H$_2$), 15.3 and 11.0 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 367.2203, found 367.2211.
C12	Boc-II-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 7.86 (1H, d, *J* 8.2, N<u>H</u>), 6.72 (1H, d, *J* 9.1, N<u>H</u>), 4.10 (1H, dd, *J* 8.2, 6.6, NH-C<u>H</u>$_I$), 3.87 (1H, t, *J* 8.4, NH-C<u>H</u>$_I$), 1.81 – 1.60 (2H, m, C<u>H</u>$_I$), 1.43 (2H, dd, *J* 8.8, 4.7, C<u>H</u>$_{2I}$), 1.38 (18H, d, *J* 7.8, C(C<u>H</u>$_3$)$_3$), 1.27 – 1.01 (2H, m, C<u>H</u>$_{2I}$), 0.91 – 0.76 (12H, m, C<u>H</u>$_{3I}$). δ_C (101 MHz, DMSO-d_6) δ 171.5, 170.4, and 155.3 (<u>C</u>=O), 80.5, 78.0 (<u>C</u>(CH$_3$)$_3$), 58.6 and 56.7 (NH-<u>C</u>H$_I$), 36.4 and 36.4 (<u>C</u>H$_I$), 28.2 and 27.6 (C(CH$_3$)$_3$), 24.8 and 24.4 (<u>C</u>H$_{2I}$), 15.3, 11.1, and 10.9 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 423.2829, found 423.2824.
C13	Boc-VV-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 7.83 (1H, d, *J* 8.3, N<u>H</u>), 6.69 (1H, d, *J* 9.1, N<u>H</u>), 4.06 (1H, dd, *J* 8.2, 6.1, NH-C<u>H</u>$_V$), 3.91 – 3.82 (1H, m, NH-C<u>H</u>$_V$), 2.07 – 1.86 (2H, m, C<u>H</u>$_V$), 1.38 (18H, d, *J* 7.4, C(CH$_3$)$_3$), 0.85 (12H, dt, *J* 18.5, 6.2, C<u>H</u>$_{3V}$). δ_C (101 MHz, DMSO-d_6) 171.6, 170.4, and 155.4 (<u>C</u>=O), 80.5, and 78.0 (<u>C</u>(CH$_3$)$_3$), 59.6 and 57.8 (NH-<u>C</u>H$_V$), 30.3 and 30.0 (<u>C</u>H$_V$), 28.1 and 27.6 (C(CH$_3$)$_3$), 19.2, 18.9, 18.2, and 18.0 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 395.2516, found 395.2523.

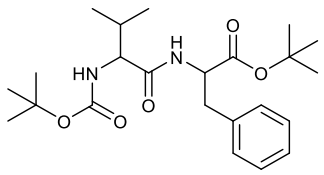
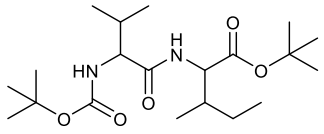
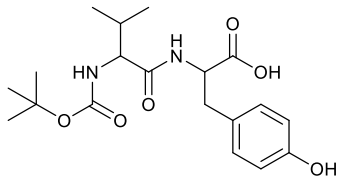
C14	Boc-YV-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 7.91 (1H, d, *J* 8.3, N<u>H</u>), 7.17 (2H, d, *J* 8.4, <u>H</u>$_{ArY}$), 6.91 (1H, d, *J* 8.8, N<u>H</u>), 6.85 (2H, d, *J* 8.4, <u>H</u>$_{ArY}$), 4.24 (1H, td, *J* 10.4, 3.8, NH-C<u>H</u>$_Y$), 4.08 (1H, dd, *J* 8.3, 5.8, NH-C<u>H</u>$_V$), 2.92 (1H, dd, *J* 13.9, 3.7, C<u>H</u>$_{2Y}$), 2.67 (1H, dd, *J* 13.7, 10.8, C<u>H</u>$_{2Y}$), 2.10 – 1.98 (1H, m, C<u>H</u>$_V$), 1.40 (9H, s, C(CH$_3$)$_3$), 1.27 (18H, d, *J* 13.1, C(CH$_3$)$_3$), 0.89 (6H, t, *J* 6.3, C<u>H</u>$_{3V}$). δ_C (101 MHz, DMSO-d_6) 172.1, 170.5, and 155.3 (<u>C</u>=O), 153.4, 132.8, 129.7, and 123.3 (<u>C</u>$_{ArY}$), 80.8, 78.1, and 77.6 (<u>C</u>(CH$_3$)$_3$), 57.8 (NH-<u>C</u>H$_V$), 55.6 (NH-<u>C</u>H$_Y$), 36.9 (<u>C</u>H$_{2Y}$), 30.2 (<u>C</u>H$_V$), 28.6, 28.2, and 27.7 (C(CH$_3$)$_3$), 18.9 and 18.0 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 515.3092, found 515.3104.

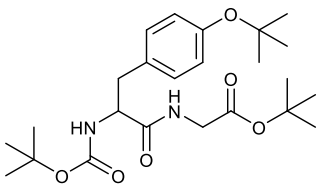
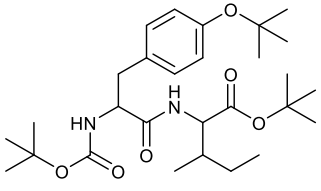
<u>Synthesised and Characterised by Dr. Bart Dietrich</u>

Compound Number	Name	Structure	Characterisation
C15	Boc-AG-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.08 (1H, t, *J* 5.78, N<u>H</u>), 6.90 (0.8H, d, *J* 7.76, N<u>H</u>), 6.50 (0.1H, br m, N<u>H</u>), 4.02-3.95 (0.85H, m, NH-C<u>H</u>$_A$), 3.93-3.82 (0.15H, br m, NH-C<u>H</u>$_A$), 3.75 (1H, dd, *J* 17.33, 6.09, C<u>H</u>$_{2G}$), 3.65 (1H, dd, *J* 17.37, 5.69, C<u>H</u>$_{2G}$), 1.39 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.37 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.18 (3H, d, *J* 7.17, C<u>H</u>$_3$). δ_C (100 MHz, DMSO-d_6) 173.10, 168.85, and 154.97 (<u>C</u>=O), 80.50 and 77.95 (<u>C</u>(CH$_3$)$_3$), 49.45 (NH-<u>C</u>H$_A$), 41.35 (<u>C</u>H$_{2G}$), 28.18 and 27.67 (C(<u>C</u>H$_3$)$_3$), 18.25 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 325.1734; found 325.1741.

C16	Boc-AI-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 7.85-7.78 (0.15H, br, N<u>H</u>), 7.74 (0.85H, d, *J* 8.15, N<u>H</u>), 6.94 (0.85H, d, *J* 7.76, N<u>H</u>), 6.53-6.42 (0.15H, br, N<u>H</u>), 4.13-3.91 (2H, m, NHC<u>H</u>*), 1.81-1.67 (1H, m, NHCH*C<u>H</u>*), 1.39 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.37 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.35-1.32 (0.7H, br, C<u>H</u>$_2$), 1.24-1.19 (1.3H, m, C<u>H</u>$_2$), 1.16 (3H, d, *J* 7.12, CH*CH*C<u>H</u>$_3$), 0.87-0.84 (6H, m, C<u>H</u>$_3$). δ_C (100 MHz, DMSO-d_6) 172.73, 170.32, and 155.01 (<u>C</u>=O), 80.53 and 77.98 (<u>C</u>(CH$_3$)$_3$), 56.64 and 49.42 (NH-<u>C</u>H), 36.67 (<u>C</u>H$_I$), 28.12 and 27.58 (<u>C</u>H$_{3I}$), 24.77 (<u>C</u>H$_{2I}$), 18.00 (<u>C</u>H$_{3I}$), 15.26 and 11.22 (<u>C</u>H$_{3A}$ and <u>C</u>H$_{3I}$). HRMS (ESI) m/z: [M+Na]$^+$ calculated 381.2360; found 381.2363.
C17	Boc-AY-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.00 (1H, d, *J* 7.50, N<u>H</u>), 7.12 (2H, d, *J* 8.40, <u>H</u>$_{ArY}$), 6.86 (2.8H, pseudo-d, *J* 8.36, <u>H</u>$_{ArF}$ and N<u>H</u>), 6.52-6.39 (0.15H, br, N<u>H</u>), 4.40-4.24 (1H, m, NH-C<u>H</u>), 4.05-3.90 (0.87H, m, NH-C<u>H</u>), 3.90-3.78 (0.13H, br, NH-C<u>H</u>), 2.89 (2H, d, *J* 7.24, C<u>H</u>$_{2Y}$), 1.36 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.28 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.26 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.13 (3H, d, *J* 7.14, C<u>H</u>$_{3A}$). δ_C (100 MHz, DMSO-d_6) 172.58, 170.42, and 154.90 (<u>C</u>=O), 153.56, 131.66, 129.74, and 123.45 (<u>C</u>$_{ArY}$), 80.53, 77.96, and 77.60 (<u>C</u>(CH$_3$)$_3$), 53.92 and 49.49 (NH-<u>C</u>H), 36.32 (<u>C</u>H$_{2Y}$), 28.49, 28.14, and 27.45 (C(<u>C</u>H$_3$)$_3$), 18.16 (<u>C</u>H$_{3A}$). HRMS (ESI) m/z: [M+Na]$^+$ calculated 487.2779; found 487.2785.
C18	Boc-FG-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.37-8.30 (0.15H, br, N<u>H</u>), 8.29 (0.85H, t, *J* 5.87, N<u>H</u>), 7.34-7.22 (4H, m, <u>H</u>$_{ArF}$), 7.22-7.14 (1H, m, <u>H</u>$_{ArF}$), 6.92 (0.85H, d, *J* 8.92, N<u>H</u>), 6.46 (0.15H, d, *J* 8.50, N<u>H</u>), 4.24-4.14 (0.85H, m, C<u>H</u>$_F$), 4.14-4.06 (0.15H, br, C<u>H</u>$_F$), 3.79 (1H, dd, *J* 17.37, 6.08, C<u>H</u>$_{2G}$), 3.70 (1H, dd, *J* 17.37, 5.70, C<u>H</u>$_{2G}$), 3.00 (1H, dd, *J* 13.89, 3.66, C<u>H</u>$_{2F}$), 2.72 (1H, dd, *J* 13.75, 10.93, C<u>H</u>$_{2F}$), 1.41 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.28 (7.7H, s, C(C<u>H</u>$_3$)$_3$), 1.19 (1.4H, s, C(C<u>H</u>$_3$)$_3$). δ_C (100 MHz, DMSO-d_6) 172.20, 168.89, and 155.24 (<u>C</u>=O), 138.30, 129.17, 127.97, and 126.12 (<u>C</u>$_{ArF}$), 80.59 and 77.91 (<u>C</u>(CH$_3$)$_3$), 55.54 (NH-<u>C</u>H$_F$), 41.45 (<u>C</u>H$_{2G}$), 37.52 (<u>C</u>H$_{2F}$), 28.13 and 27.70 (C(<u>C</u>H$_3$)$_3$). HRMS (ESI) m/z: [M+Na]$^+$ calculated 401.2047; found 401.2050.

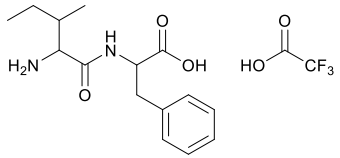
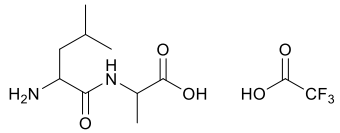
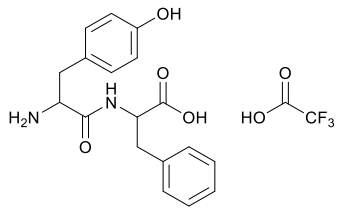
C19	Boc-FI-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.07 (0.15H, d, *J* 7.27, N<u>H</u>), 7.95 (0.85H, d, *J* 8.17, N<u>H</u>), 7.37-7.14 (5H, m, <u>H</u>$_{ArF}$), 6.93 (0.85H, d, *J* 8.70, N<u>H</u>), 6.45 (0.15H, d, *J* 8.10, N<u>H</u>), 4.30-4.20 (0.85H, m, NH-C<u>H</u>), 4.20-4.16 (0.15H, m, NH-C<u>H</u>), 4.12 (1H, dd, *J* 8.04, 6.12, NH-C<u>H</u>), 2.95 (0.85H, dd, *J* 13.82, 3.87, C<u>H</u>$_{2F}$), 2.92-2.88 (0.15H, br, C<u>H</u>$_{2F}$), 2.73 (0.85H, dd, *J* 13.67, 10.62, C<u>H</u>$_{2F}$), 2.68-2.61 (0.15H, br, C<u>H</u>$_{2F}$), 1.85 (1H, C<u>H</u>$_I$), 1.48-1.35 (10H, m, C(C<u>H</u>$_3$)$_3$ and C<u>H</u>$_{2I}$), 1.34-1.13 (10H, m, C(C<u>H</u>$_3$)$_3$ and C<u>H</u>$_{2I}$), 0.93-0.82 (6H, m, C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 171.75, 170.35, and 155.20 (<u>C</u>=O), 138.09, 129.16, 127.91, and 126.11 (<u>C</u>$_{ArF}$), 80.61 and 77.97 (<u>C</u>(CH$_3$)$_3$), 56.74 and 55.41 (NH-<u>C</u>H), 37.27 (<u>C</u>H$_{2F}$), 36.67 (<u>C</u>H$_I$), 28.06 and 27.60 (C(<u>C</u>H$_3$)$_3$), 24.80 (<u>C</u>H$_{2I}$), 15.29 and 11.21 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 457.2673; found 457.2677.
C20	Boc-GY-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.05 (1H, d, *J* 7.50, N<u>H</u>), 7.12-7.10 (2H, m, <u>H</u>$_{ArY}$), 6.93 (0.85H, t, *J* 5.98, N<u>H</u>), 6.89-6.85 (2H, m, <u>H</u>$_{Ar}$), 6.56-6.48 (0.15H, br, N<u>H</u>), 4.34 (1H, dd, *J* 14.68, 7.27, NH-C<u>H</u>), 3.54 (1.7H, d, *J* 6.11, C<u>H</u>$_{2G}$), 3.50-3.43 (0.3H, br, C<u>H</u>$_{2G}$), 2.92-2.83 (2H, m, C<u>H</u>$_{2F}$), 1.37 (7.7H, s, C(C<u>H</u>$_3$)$_3$), 1.34-1.30 (1.3H, br, C(C<u>H</u>$_3$)$_3$), 1.28 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.26 (9H, s, C(C<u>H</u>$_3$)$_3$). δ_C (100 MHz, DMSO-d_6) 170.54, 169.24 and 155.73 (<u>C</u>=O), 153.63, 131.57, 129.76, and 123.51 (<u>C</u>$_{ArY}$), 80.66, 77.98, and 77.66 (<u>C</u>(CH$_3$)$_3$), 54.01 (NH-<u>C</u>H$_Y$), 42.94 (<u>C</u>H$_{2G}$), 36.62 (<u>C</u>H$_{2F}$), 28.51, 28.19, and 27.47 (C(<u>C</u>H$_3$)$_3$). HRMS (ESI) m/z: $[M+Na]^+$ calculated 473.2622; found 473.2630.
C21	Boc-IY-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.14 (1H, d, *J* 7.47, N<u>H</u>), 7.12 (2H, d, *J* 8.44, <u>H</u>$_{ArY}$), 6.85 (2H, d, *J* 8.38, <u>H</u>$_{ArY}$), 6.62 (0.9H, d, *J* 9.20, N<u>H</u>), 6.29-6.18 (0.1H, br, N<u>H</u>), 4.39-4.30 (1H, m, NH-C<u>H</u>), 3.87-3.79 (0.9H, m, NH-C<u>H</u>), 3.77-3.69 (0.1H, m, NH-C<u>H</u>), 2.94-2.82 (2H, m, C<u>H</u>$_{2F}$), 1.68-1.54 (1H, m, C<u>H</u>$_I$), 1.42-1.31 (overlapped by *t*-Bu, 1H, C<u>H</u>$_{2I}$), 1.37 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.28 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.25 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.13-0.97 (1H, m, C<u>H</u>$_{2I}$), 0.78 (3H, t, *J* 7.40, C<u>H</u>$_{3I}$), 0.74 (3H, d, *J* 6.77, C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 171.17, 170.47, and 155.14 (<u>C</u>=O), 153.53, 131.70, 129.64, and 123.45 (<u>C</u>$_{ArY}$), 80.43, 77.91, and 77.58 (<u>C</u>(CH$_3$)$_3$), 58.48 (NH-<u>C</u>H), 53.93 (NH-<u>C</u>H), 36.70 (<u>C</u>H$_I$), 36.34 (C<u>H</u>$_{2F}$), 28.47, 28.13, and 27.43 (C(<u>C</u>H$_3$)$_3$), 24.30 (<u>C</u>H$_{2I}$), 15.16 and 10.87 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 507.342864; found 507.343753.

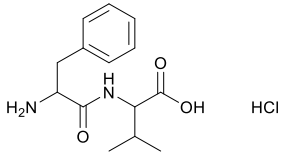
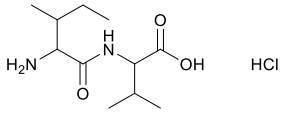
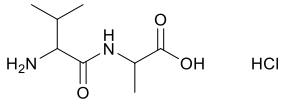
C22	Boc-VF-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.15 (1H, d, *J* 7.52, N<u>H</u>), 7.31-7.15 (5H, m, <u>H</u>$_{ArF}$), 6.56 (0.9H, d, *J* 9.25, N<u>H</u>), 6.25-6.12 (0.1H, br, N<u>H</u>), 4.46-4.30 (1H, m, NH-C<u>H</u>), 3.86-3.76 (0.9H, m, NH-C<u>H</u>), 3.73-3.63 (0.1H, br, NH-C<u>H</u>), 2.96 (1H, dd, *J* 14.05, 6.68, C<u>H</u>$_{2F}$), 2.90 (1H, dd, *J* 13.91, 8.29, C<u>H</u>$_{2F}$), 1.93-1.80 (1H, m, C<u>H</u>$_V$), 1.37 (8.1H, s, C(C<u>H</u>$_3$)$_3$), 1.30 (9.9H, s, C(C<u>H</u>$_3$)$_3$), 0.79 (3.3H, d, *J* 2.05, C<u>H</u>$_{3V}$), 0.71 (2.7H, d, *J* 2.07, C<u>H</u>$_{3V}$). δ_C (100 MHz, DMSO-d_6) 171.16, 170.39, and 155.25 (<u>C</u>=O), 137.10, 129.13, 128.07, and 126.40 (<u>C</u>$_{ArF}$), 80.55 and 77.92 (<u>C</u>(CH$_3$)$_3$), 59.38 and 53.91 (NH-<u>C</u>H), 36.95 (<u>C</u>H$_{2F}$), 30.52 (<u>C</u>H$_V$), 28.13 and 27.45 (C(<u>C</u>H$_3$)$_3$), 19.10 and 18.03 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 507.342864; found 507.343753.
C23	Boc-VI-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 7.85 (1H, d, *J* 8.08, N<u>H</u>), 6.69 (0.85H, d, *J* 9.10, N<u>H</u>), 6.36-6.21 (0.15H, br, N<u>H</u>), 4.10 (1H, dd, *J* 8.00, 6.53, NH-C<u>H</u>), 3.86-3.81 (0.85H, m, NH-C<u>H</u>*), 3.80-3.72 (0.15H, br, NH-C<u>H</u>), 1.98-1.84 (1H, m, C<u>H</u>$_V$), 1.80-1.68 (1H, m, C<u>H</u>$_I$), 1.47-1.29 (overlapped by *t*-Bu, 1H, C<u>H</u>$_{2I}$), 1.39 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.37 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.24-1.12 (1H, m, C<u>H</u>$_{2I}$), 0.86-0.81 (12H, m, C<u>H</u>$_{3V}$ and C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 171.38, 170.33, and 155.34 (<u>C</u>=O), 80.47 and 77.90 (<u>C</u>(CH$_3$)$_3$), 59.46 and 56.69 (N<u>C</u>*), 36.44 (<u>C</u>H*CH$_3$), 30.32 (<u>C</u>H(CH$_3$)$_2$), 28.10 and 27.56 (C(<u>C</u>H$_3$)$_3$), 24.75 (<u>C</u>H$_2$CH$_3$), 19.14 and 18.14 (CH(<u>C</u>H$_3$)$_2$), 15.26 and 11.08 (CH*<u>C</u>H$_3$ and CH$_2$<u>C</u>H$_3$). HRMS (ESI) m/z: $[M+H]^+$ calculated 409.267293; found 409.268155.
C24	Boc-VY-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.14 (1H, d, *J* 7.35, N<u>H</u>), 7.12 (2H, d, *J* 8.42, <u>H</u>$_{ArY}$), 6.85 (2H, d, *J* 8.35, <u>H</u>$_{ArY}$), 6.58 (0.85H, d, *J* 9.27, N<u>H</u>), 6.25-6.15 (0.15H, br, N<u>H</u>), 4.37-4.32 (1H, m, NH-C<u>H</u>), 3.82-3.78 (0.85H, NH-C<u>H</u>), 3.73-3.63 (0.15H, br, NH-C<u>H</u>), 2.93-2.83 (2H, m, C<u>H</u>$_{2Y}$), 1.90-1.82 (1H, m, C<u>H</u>$_V$), 1.38 (7.7H, s, C(C<u>H</u>$_3$)$_3$), 1.33 (1.3H, s, C(C<u>H</u>$_3$)$_3$), 1.28 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.25 (9H, s, C(C<u>H</u>$_3$)$_3$), 0.79 (3H, d, *J* 3.61, C<u>H</u>$_{3V}$), 0.77 (3H, d, *J* 3.69, C<u>H</u>$_{3V}$). δ_C (100 MHz, DMSO-d_6) 171.14, 170.50, and 155.26 (<u>C</u>=O), 153.55, 131.69, 129.65, and 123.48 (<u>C</u>$_{ArY}$), 80.46, 77.93, and 77.58 (<u>C</u>(CH$_3$)$_3$), 59.41 and 53.95 (NH-<u>C</u>H), 36.38 (<u>C</u>H$_{2Y}$), 30.56 (<u>C</u>H$_V$), 28.47, 28.13, and 27.44 (C(<u>C</u>H$_3$)$_3$), 19.09 and 18.06 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M+H]^+$ calcd for $C_{27}H_{45}N_2O_6$ 493.327214; found 493.328249.

C25	Boc-YG-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.35-8.29 (0.15H, br, N<u>H</u>), 8.26 (0.85H, t, *J* 5.74, N<u>H</u>), 7.16 (2H, d, *J* 8.36, <u>H</u>$_{ArY}$), 6.91 (0.85H, d, *J* 9.00, N<u>H</u>), 6.47 (0.15H, d, *J* 6.98, N<u>H</u>), 4.22-4.14 (0.85H, m, NH-C<u>H</u>$_Y$), 4.14-4.04 (0.15H, br, NH-C<u>H</u>$_Y$), 3.78 (1H, dd, *J* 17.34, 6.02, C<u>H</u>$_{2G}$), 3.69 (1H, dd, *J* 17.31, 5.67, C<u>H</u>$_{2G}$), 2.95 (1H, dd, *J* 13.86, 3.44, C<u>H</u>$_{2F}$), 2.66 (1H, dd, *J* 13.70, 11.29, C<u>H</u>$_{2F}$), 1.41 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.27 (7.7H, s, C(C<u>H</u>$_3$)$_3$), 1.25 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.20 (1.3H, s, C(C<u>H</u>$_3$)$_3$). δ_C (100 MHz, DMSO-d_6) 172.16, 168.84, and 155.18 (<u>C</u>=O), 153.31, 132.89, 129.63, and 123.27 (<u>C</u>$_{ArY}$), 80.56, 77.83, and 77.50 (<u>C</u>(CH$_3$)$_3$), 55.52 (NH-<u>C</u>H$_Y$), 41.42 (<u>C</u>H$_{2G}$), 36.98 (<u>C</u>H$_{2Y}$), 28.51, 28.11, and 27.68 (C(<u>C</u>H$_3$)$_3$). HRMS (ESI) m/z: $[M+H]^+$ calculated 451.2803; found 451.2813.
C26	Boc-YI-O^{t}Bu		δ_H (400 MHz, DMSO-d_6) 8.04 (0.15H, d, *J* 8.15, N<u>H</u>), 7.93 (0.85H, d, *J* 8.16, N<u>H</u>), 7.22 (0.3H, d, *J* 7.32, <u>H</u>$_{ArY}$), 7.17 (1.7H, d, *J* 8.34, <u>H</u>$_{ArY}$), 6.93 (0.85H, d, *J* 8.79, N<u>H</u>), 6.85 (2H, d, *J* 8.35, <u>H</u>$_{ArY}$), 6.46 (0.15H, d, *J* 7.68, N<u>H</u>), 4.28-4.19 (0.85H, m, NH-C<u>H</u>), 4.19-4.16 (0.15H, m, NH-C<u>H</u>), 4.16-4.07 (1H, m, NH-C<u>H</u>), 2.91 (0.85H, dd, *J* 13.91, 3.66, C<u>H</u>$_{2Y}$), 2.87-2.82 (0.15H, br, C<u>H</u>$_{2Y}$), 2.67 (0.85H, dd, *J* 13.70, 10.71, C<u>H</u>$_{2Y}$), 2.62-2.57 (0.15H, br, C<u>H</u>$_{2Y}$), 1.84-1.71 (1H, m, C<u>H</u>$_I$), 1.46-1.38 (1H, C<u>H</u>$_{2I}$, overlapped by C(CH$_3$)$_3$), 1.40 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.31-1.17 (1H, C<u>H</u>$_{2I}$, overlapped by C(CH$_3$)$_3$), 1.29 (9H, s, C(C<u>H</u>$_3$)$_3$), 1.26 (9H, s, C(C<u>H</u>$_3$)$_3$), 0.87 (3H, t, *J* 7.24, C<u>H</u>$_{3I}$), 0.86 (3H, d, *J* 6.69, C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 171.78, 170.33, and 155.16 (<u>C</u>=O), 153.33, 132.70, 129.64, and 123.19 (<u>C</u>$_{ArY}$), 80.59, 77.90, and 77.48 (<u>C</u>(CH$_3$)$_3$), 56.73 and 55.42 (NH-<u>C</u>H), 36.76 (<u>C</u>H$_{2F}$), 36.65 (<u>C</u>H$_I$), 28.50, 28.06, and 27.60 (C(<u>C</u>H$_3$)$_3$), 24.81 (<u>C</u>H$_{2I}$), 15.30 and 11.22 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 507.342864; found 507.343592.

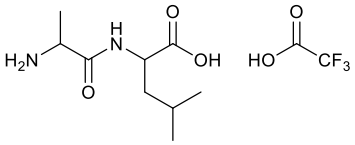
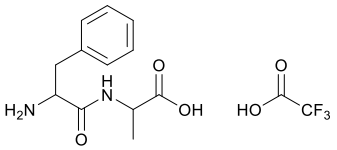
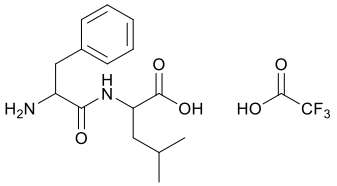
A 3.1.4 Deprotected Dipeptide Products

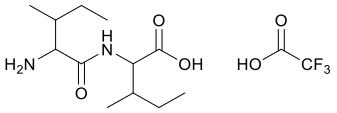
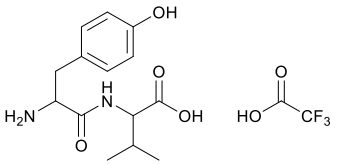
Synthesised and Characterised by Connor MacDonald

Compound Number	Name	Structure	Characterisation
D1	IF· TFA		δ_H (400 MHz, DMSO-d_6) 8.69 (1H, d, *J* 7.6, NH), 8.50 – 7.50 (2H, bs, NH) 7.34 – 7.18 (5H, m, H_{ArF}), 4.55 – 4.45 (1H, m, NH-CH_F), 3.64 (1H, d, *J* 5.1, NH-CH_I), 3.09 (1H, dd, *J* 14.1, 5.2, CH_{2F}), 2.95 (1H, dd, *J* 14.1, 8.8, CH_{2F}), 1.92 – 1.77 (1H, m, CH_I), 1.46 (1H, ddd, *J* 13.4, 7.5, 3.4, CH_{2I}), 1.08 (1H, ddd, *J* 16.5, 14.4, 7.2, CH_{2I}), 0.93 (3H, d, *J* 6.9, CH_{3I}), 0.84 (3H, t, *J* 7.4, CH_{3I}). δ_C (101 MHz, DMSO-d_6) 172.3 and 168.2 (C=O), 158.6, 158.3, 158.0, and 157.7 (F_3C-C=O), 137.3, 129.1, 128.3, and 126.6 (C_{ArF}), 121.7, 118.8, 115.8, 112.8 (CF_3), 56.5 (NH-CH_F), 53.9 (NH-CH_I), 36.5 (CH_{2F}), 36.4 (CH_I), 23.5 (CH_{2I}), 14.5 and 11.2 (CH_{3I}). HRMS (ESI) m/z: $[M+H]^+$ calculated 279.1703, found 279.1712.
D2	LA· TFA		Not fully characterised due to inseparable cyclisation byproduct. HRMS (ESI) m/z: $[M+H]^+$ calculated 203.1390, found 203.1392.
D3	YF· TFA		δ_H (400 MHz, DMSO-d_6) 9.35 (1H, s, OH), 8.80 (1H, d, *J* 7.9, NH), 8.50 – 7.50 (1H, bs, NH), 7.34 – 7.20 (4H, m, H_{ArF}), 7.05 (2H, dd, *J* 11.4, 8.5, H_{ArY}), 6.73 – 6.66 (2H, m, H_{ArY}), 4.56 – 4.46 (1H, m, NH-CH_F), 3.90 (1H, dd, *J* 8.1, 4.4, NH-CH_Y), 3.11 (1H, dd, *J* 14.0, 5.1, CH_{2F}), 3.01 (1H, dd, *J* 14.5, 4.5, CH_{2Y}), 2.95 (1H, dd, *J* 14.1, 8.6, CH_{2F}), 2.77 (1H, dd, *J* 14.4, 8.5, CH_{2Y}). δ_C (101 MHz, DMSO-d_6) 172.8 and 168.8 (C=O), 159.0, 158.7, 158.4, and 158.0 (F_3C-C=O), 157.1, 137.7, 131.1, 129.6, 128.8, 127.0, and 125.1 (C_{Ar}), 122.2, 119.2, and 116.3 (CF_3), 115.8 (C_{Ar}), 113.3 (CF_3), 54.3 (NH-CH_F), 53.9 (NH-CH_Y), 37.1 (CH_{2F}), 36.7 (CH_{2Y}). HRMS (ESI) m/z: $[M+H]^+$ calculated 329.1496, found 329.1506.

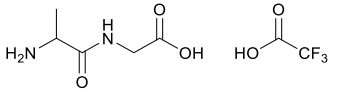
D4	FV·HCl		δ_H (400 MHz, DMSO-d_6) 8.71 (1H, d, *J* 8.4, N<u>H</u>), 7.37 – 7.21 (2H, bs, N<u>H</u>), 7.37 – 7.21 (5H, m, <u>H</u>$_{ArF}$), 4.18 (2H, ddd, *J* 8.5, 5.6, 3.2, NH-C<u>H</u>$_V$ and NH-C<u>H</u>$_F$), 3.15 (1H, dd, *J* 14.1, 5.6, C<u>H</u>$_{2F}$), 2.98 (1H, dd, *J* 14.1, 7.5, C<u>H</u>$_{2F}$), 2.16 – 2.03 (1H, m, C<u>H</u>$_V$), 0.92 (6H, dd, *J* 6.8, 1.6, C<u>H</u>$_{3V}$). δ_C (101 MHz, DMSO-d_6) 172.2 and 168.2 (<u>C</u>=O), 134.8, 129.6, 128.4 and 127.1 (<u>C</u>$_{ArF}$), 57.5 (NH-<u>C</u>H$_F$), 53.0 (NH-<u>C</u>H$_V$), 36.8 (<u>C</u>H$_{2F}$), 29.9 (<u>C</u>H$_V$), 19.1 and 18.0 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 265.1547, found 265.1553.
D5	IV·HCl		δ_H (400 MHz, DMSO-d_6) 8.75 – 7.83 (2H, m, N<u>H</u>), 8.50 (1H, d, *J* 8.0, N<u>H</u>), 4.16 (1H, dd, *J* 8.0, 5.6, NH-C<u>H</u>$_I$), 3.79 (1H, d, *J* 5.8, NH-<u>C</u>H$_V$), 2.16 – 2.04 (1H, m, C<u>H</u>$_I$), 1.85 (1H, ddt, *J* 12.8, 6.6, 3.4, C<u>H</u>$_V$), 1.55 (1H, ddd, *J* 13.6, 7.5, 3.4, C<u>H</u>$_{2I}$), 1.22 – 1.07 (1H, m, C<u>H</u>$_{2I}$), 0.94 (9H, d, *J* 6.8, C<u>H</u>$_{3V}$ and C<u>H</u>$_{3I}$), 0.86 (3H, t, *J* 7.4, C<u>H</u>$_{3I}$). δ_C (101 MHz, DMSO-d_6) 172.3 and 168.2 (<u>C</u>=O), 57.6 (NH-<u>C</u>H$_I$), 56.1 (NH-<u>C</u>H$_V$), 36.3 (<u>C</u>H$_V$), 29.7 (<u>C</u>H$_I$), 23.8 (<u>C</u>H$_{2I}$), 19.1 and 18.1 (<u>C</u>H$_{3V}$), 14.4 and 11.2 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 231.1703, found 231.1707.
D6	VA·HCl		δ_H (400 MHz, DMSO-d_6) 9.46 – 7.18 (2H, m, N<u>H</u>), 8.77 (1H, d, *J* 6.7, N<u>H</u>) 4.32 – 4.21 (1H, m, NH-C<u>H</u>$_A$), 3.60 (1H, d, *J* 5.6, NH-C<u>H</u>$_V$), 2.18 – 2.04 (1H, m, C<u>H</u>$_V$), 1.32 (3H, d, *J* 7.3, C<u>H</u>$_{3A}$), 0.96 (6H, d, *J* 6.9, C<u>H</u>$_{3V}$). δ_C (101 MHz, DMSO-d_6) 173.5 and 167.5 (<u>C</u>=O), 57.0 (NH-<u>C</u>H$_V$), 47.7 (NH-<u>C</u>H$_A$), 29.8 (<u>C</u>H$_V$), 18.2 (<u>C</u>H$_{3A}$), 17.9 and 17.0 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M-H]^-$ calculated 187.1088, found 187.1087.

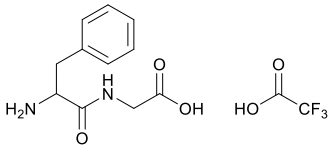
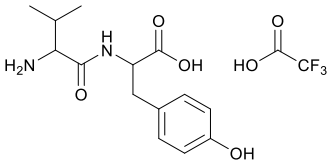
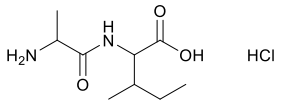
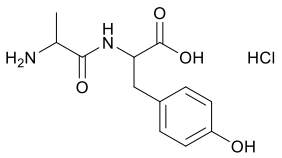
Synthesised by Dr Rebecca Ginesi and Characterised by Connor MacDonald

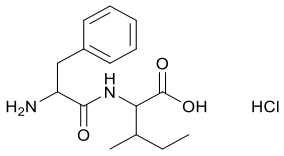
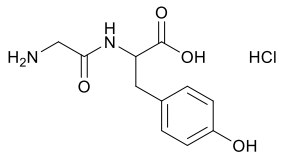
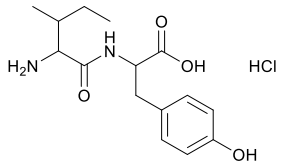
Compound Number	Name	Structure	Characterisation
D7	AL· TFA		δ_H (400 MHz, DMSO-d_6) 12.76 (1H, s, COO<u>H</u>), 8.56 (1H, d, *J* 8.2, N<u>H</u>), 8.26 (2H, s, N<u>H</u>), 4.20 (1H, dd, *J* 8.2, 5.6, NH-C<u>H</u>$_L$), 3.99 – 3.89 (1H, m, NH-C<u>H</u>$_A$), 1.89 – 1.74 (1H, m, C<u>H</u>$_L$), 1.48 – 1.37 (1H, m, C<u>H</u>$_{2L}$), 1.35 (3H, d, *J* 6.9, C<u>H</u>$_{3A}$), 1.30 – 1.14 (1H, m, C<u>H</u>$_{2L}$), 0.97 – 0.82 (6H, m, C<u>H</u>$_{3L}$). δ_C (101 MHz, DMSO-d_6) 172.5 and 169.8 (<u>C</u>=O), 56.7 (NH-<u>C</u>H$_L$), 47.9 (NH-<u>C</u>H$_A$), 36.3 (<u>C</u>H$_L$), 24.7 (<u>C</u>H$_{2L}$), 17.3 (<u>C</u>H$_{3A}$), 15.7 and 11.5 (). HRMS (ESI) m/z: $[M+H]^+$ calculated 203.1390, found 203.1398.
D8	FA· TFA		δ_H (400 MHz, DMSO-d_6) 8.82 (1H, d, *J* 7.3, N<u>H</u>), 7.38 – 7.20 (5H, m, <u>H</u>$_{ArF}$), 4.35 – 4.23 (1H, m, NH-C<u>H</u>$_A$), 4.03 (1H, dd, *J* 8.0, 5.3, NH-C<u>H</u>$_F$), 3.14 (1H, dd, *J* 14.1, 5.2, C<u>H</u>$_{2F}$), 2.93 (1 H, dd, *J* 14.2, 8.0, C<u>H</u>$_{2F}$), 1.33 (3 H, d, *J* 7.3, C<u>H</u>$_{3A}$). δ_C (101 MHz, DMSO-d_6) 173.88 and 168.36 (<u>C</u>=O), 135.34, 130.04, 128.99, and 127.61 (<u>C</u>$_{ArF}$), 53.79 (NH-<u>C</u>H$_F$), 48.18 (NH-<u>C</u>H$_A$), 37.42 (<u>C</u>H$_{2F}$), 17.72 (<u>C</u>H$_{3A}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 237.1234, found 237.1240.
D9	FL· TFA		δ_H (400 MHz, DMSO-d_6) 8.75 (1H, d, *J* 7.9, N<u>H</u>), 8.50 – 7.50 (2H, bs, N<u>H</u>), 7.38 – 7.24 (5H, m, <u>H</u>$_{ArF}$), 4.34 – 4.24 (1H, m, NH-C<u>H</u>$_L$), 4.05 (1H, dd, *J* 8.1, 5.0, NH-C<u>H</u>$_F$), 3.15 (1 H, dd, *J* 14.2, 4.9, C<u>H</u>$_{2F}$), 2.93 (1H, dd, *J* 14.2, 8.2, C<u>H</u>$_{2F}$), 1.74 – 1.61 (1 H, m, C<u>H</u>$_L$), 1.57 (2 H, t, *J* 7.3, C<u>H</u>$_{2L}$), 0.91 (6H, dd, *J* 17.6, 6.5, C<u>H</u>$_{3L}$). δ_C (101 MHz, DMSO-d_6) 173.4 and 168.1 (<u>C</u>=O), 134.7, 129.6, 128.5, and 127.2 (<u>C</u>$_{ArF}$), 53.2 (NH-<u>C</u>H$_F$), 50.5 (NH-<u>C</u>H$_L$), 40.0 (overlapped by DMSO-d_6, <u>C</u>H$_{2L}$), 36.9 (<u>C</u>H$_{2F}$), 24.2 (<u>C</u>H$_L$), 22.8 and 21.3 (CH$_{3L}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 279.1703, found 279.1708.

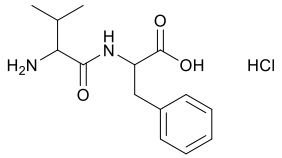
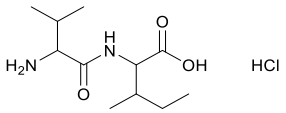
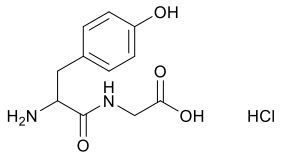
D10	II· TFA		δ_H (400 MHz, DMSO-d_6) 8.28 (1H, s, N<u>H</u>), 4.19 (1H, bs, NH-C<u>H</u>$_I$), 3.44 (1H, d, *J* 5.3, NH-C<u>H</u>$_I$), 1.86 – 1.72 (2H, m, <u>C</u>H$_V$), 1.45 (2H, dt, *J* 13.4, 6.1, <u>C</u>H$_{2V}$), 1.27 – 1.05 (2H, m, <u>C</u>H$_{2V}$), 0.98 – 0.71 (12 H, m, <u>C</u>H$_{3V}$). δ_C (101 MHz, DMSO-d_6) 173.19 and 169.43 (<u>C</u>=O), 57.5 and 57.3 (NH-<u>C</u>H$_I$), 37.1 (<u>C</u>H$_V$), 25.21 and 24.22 (<u>C</u>H$_{2I}$), 16.06, 15.2, 11.9, and 11.7 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 245.1860, found 245.1863..
D11	YV· TFA		δ_H (400 MHz, DMSO-d_6) 9.37 (1H, s, O<u>H</u>), 8.59 (1H, d, *J* 8.4, N<u>H</u>), 7.07 (2H, d, *J* 8.5, <u>H</u>$_{ArY}$), 6.71 (2H, d, *J* 8.5, <u>H</u>$_{ArY}$), 4.21 (1H, dd, *J* 8.4, 5.6, NH-C<u>H</u>$_A$), 4.04 (1H, dd, *J* 8.0, 5.0, NH-C<u>H</u>$_Y$), 3.01 (1H, dd, *J* 14.3, 4.9, C<u>H</u>$_{2Y}$), 2.81 (1H, dd, *J* 14.3, 8.1, C<u>H</u>$_{2Y}$), 2.17 – 2.04 (1H, m, <u>C</u>H$_V$), 0.93 (6H, d, *J* 6.8, <u>C</u>H$_{3V}$). δ_C (101 MHz, DMSO-d_6) 172.4 and 168.5 (<u>C</u>=O), 156.6, 130.5, 124.7, and 115.3 (<u>C</u>$_{ArY}$), 57.4 (NH-<u>C</u>H$_V$), 53.4 (NH-<u>C</u>H$_F$), 36.3 (<u>C</u>H$_{2Y}$), 30.0 (<u>C</u>H$_V$), 19.1 and 18.0 (<u>C</u>H$_{3V}$). HRMS (ESI) m/z: $[M+H]^+$ calculated 281.1496, found 281.1499.

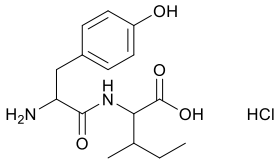
<u>Synthesised and Characterised by Dr Bart Dietrich</u>

Compound Number	Name	Structure	Characterisation
D12	AG· TFA		δ_H (400 MHz, DMSO-d_6) 8.71 (1H, t, *J* 5.85, N<u>H</u>), 8.12 (3H, br s, N<u>H</u>), 3.92-3.82 (3H, m, NH-C<u>H</u>$_A$ and C<u>H</u>$_{2G}$), 1.36 (3H, d, *J* 7.02, C<u>H</u>$_{3A}$). δ_C (100 MHz, DMSO-d_6) 170.76 and 170.00 (<u>C</u>=O), 158.51 (q, *J* 31.51, F_3C-<u>C</u>=O), 117.21 (q, *J* 299.09, <u>C</u>F$_3$), 48.12 and 40.79 (NH-<u>C</u>H and <u>C</u>H$_{2G}$), 17.22 (<u>C</u>H$_{3A}$). HRMS (ESI) m/z: [M+H]+ calculated 147.0764; found 147.0769.

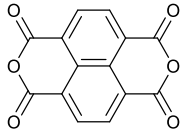
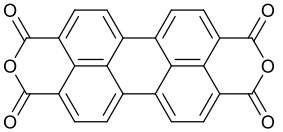
D13	FG·TFA		δ_H (400 MHz, DMSO-d_6) 9.22-7.56 (approx. 3H, br, N$\underline{H}_3^+$), 8.86 (1H, d, *J* 5.70, N$\underline{H}$), 7.36-7.24 (5H, m, $\underline{H}_{Ar}$), 4.08 (1H, dd, *J* 7.87, 5.45, C$\underline{H}^*$), 3.86 (2H, d, *J* 5.74, glycine C$\underline{H}_2$), 3.12 (1H, dd, *J* 14.06, 5.35, PhC$\underline{H}_2$), 2.95 (1H, dd, *J* 14.08, 7.94, PhC$\underline{H}_2$). δ_C (100 MHz, DMSO-d_6) 170.66 and 168.52 ($\underline{C}$=O), 158.44 (q, *J* 31.32, $CF_3\underline{C}$=O), 134.01, 129.66, 128.54, and 127.19 ($\underline{C}_{ArF}$), 117.24 (q, *J* 299.33, $\underline{C}F_3$), 53.36 (NH-$\underline{C}H_F$), 40.84 ($\underline{C}H_{2G}$), 37.04 ($\underline{C}H_{2F}$). HRMS (APCI) m/z: [M-TFA+H]+ calculated 2231077; found 223.1081.
D14	VY·TFA		δ_H (400 MHz, DMSO-d_6) 10.20-8.15 (br, N$\underline{H}$), 8.34 (1H, d, *J* 4.38, O$\underline{H}$), 7.04-6.98 (2H, m, $\underline{H}_{ArY}$), 6.68-6.60 (2H, m, $\underline{H}_{ArY}$), 4.43-4.26 (1H, br, NH-C$\underline{H}$), 3.42-3.39 (1H, m, NH-C$\underline{H}$), 2.94 (1H, dd, *J* 13.94, 5.09, C$\underline{H}_{2F}$), 2.81 (1H, dd, *J* 13.92, 8.13, C$\underline{H}_{2F}$), 2.11-1.96 (1H, m, C$\underline{H}_V$), 0.90 (3H, d, *J* 6.90, C$\underline{H}_{3V}$), 0.83 (3H, d, *J* 6.88, C$\underline{H}_{3V}$). δ_C (100 MHz, DMSO-d_6) 174.90 and 168.65 ($\underline{C}$=O), 160.78 (q, *J* 33.01, $CF_3\underline{C}$=O), 156.13, 131.27, and 129.28 ($\underline{C}_{ArY}$), 117.73 (q, *J* 296.03, $\underline{C}F_3$), 116.05 ($\underline{C}_{ArY}$), 58.50 and 56.47 (NH-$\underline{C}$H), 37.13 ($\underline{C}H_{2Y}$), 30.78 ($\underline{C}H_V$), 19.09 and 18.12 ($\underline{C}H_{3V}$). HRMS (APCI) m/z: [M-TFA-H]- calculated 279.1350; found 279.1349.
D15	AI·HCl		δ_H (400 MHz, DMSO-d_6) 12.75 (1H, br s, COO$\underline{H}$), 8.55 (1H, d, *J* 8.24, N$\underline{H}$), 8.25 (3H, br s, N$\underline{H}$), 4.20 (1H, dd, *J* 8.20, 5.66, NH-C$\underline{H}_V$), 3.94 (1H, q, *J* 6.90, NH-C$\underline{H}_A$), 1.89-1.74 (1H, m, C$\underline{H}_I$), 1.49-1.40 (1H, m, C$\underline{H}_{2I}$), 1.35 (3H, d, *J* 6.96, C$\underline{H}_{3A}$), 1.29-1.15 (1H, m, C$\underline{H}_{2I}$), 0.89 (3H, d, *J* 6.88, C$\underline{H}_{3I}$), 0.86 (3H, t, *J* 7.46, C$\underline{H}_{3I}$). δ_C (100 MHz, DMSO-d_6) 56.67 (NH-$\underline{C}H_A$), 47.87 ($NH_3\underline{C}H_I$), 36.27 ($\underline{C}H_I$), 24.71 ($\underline{C}H_{2I}$), 17.32 ($\underline{C}H_{3A}$), 15.68 ($\underline{C}H_{3A}$), 11.46 ($\underline{C}H_{3A}$). HRMS (ESI) m/z: [M-H]- calculated 201.1245; found 201.1246.
D16	AY·HCl		δ_H (400 MHz, DMSO-d_6) 12.82 (1H, br, COO$\underline{H}$), 9.32 (1H, s, O$\underline{H}$), 8.71 (1H, d, *J* 7.89, N$\underline{H}$), 8.22 (3H, br, N$\underline{H}$), 7.05 (2H, d, *J* 8.50, $\underline{H}_{ArY}$), 6.68 (2H, d, *J* 8.50, $\underline{H}_{ArY}$), 4.39-4.34 (1H, m, NH-C$\underline{H}$), 3.81 (1H, q, *J* 6.93, NH-C$\underline{H}$), 2.97 (1H, dd, *J* 13.99, 4.73, C$\underline{H}_{2F}$), 2.80 (1H, dd, *J* 14.00, 9.18, C$\underline{H}_{2F}$), 1.35 (3H, d, *J* 6.97, C$\underline{H}_{3A}$). δ_C (100 MHz, DMSO-d_6) 172.54 and 169.68 ($\underline{C}$=O), 156.17, 130.13, 127.40, and 115.18 ($\underline{C}_{ArY}$), 54.39 and 47.97 (NH-$\underline{C}$H), 35.66 ($\underline{C}H_{2F}$), 17.26 ($\underline{C}H_{3A}$). HRMS (ESI) m/z: [M-HCl+H]+ calculated 253.1183; found 253.1181.

D17	FI·HCl		δ_H (400 MHz, DMSO-d_6) 13.87-11.70 (0.6H, br, COO<u>H</u>), 8.68 (1H, d, *J* 8.31, N<u>H</u>), 8.47-7.80 (2.3H, br, N<u>H</u>), 7.35-7.23 (5H, m, <u>H</u>$_{ArF}$), 4.22 (1H, dd, *J* 8.26, 5.75, NH-C<u>H</u>), 4.15 (1H, dd, *J* 7.42, 5.65, NH-C<u>H</u>), 3.13 (1H, dd, *J* 14.06, 5.49, C<u>H</u>$_{2F}$), 2.96 (1H, dd, *J* 14.09, 7.59, C<u>H</u>$_{2F}$), 1.87-1.75 (1H, m, C<u>H</u>$_I$), 1.52-1.37 (1H, m, C<u>H</u>$_{2I}$), 1.29-1.15 (1H, m, C<u>H</u>$_{2I}$), 0.90 (3H, d, *J* 6.82, C<u>H</u>$_{3I}$), 0.87 (3H, t, *J* 7.31, C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 172.22 and 168.09 (<u>C</u>=O), 134.95, 129.72, 128.36, and 127.00 (<u>C</u>$_{ArF}$), 56.62 and 52.98 (NH-<u>C</u>H), 36.71 (<u>C</u>H$_{2F}$), 36.46 (<u>C</u>H$_I$), 24.70 (<u>C</u>H$_{2I}$), 15.57 and 11.41 (<u>C</u>H$_{3I}$). HRMS (APCI) m/z: [M-H]− calculated 277.1558; found 277.1561.
D18	GY·HCl		δ_H (400 MHz, DMSO-d_6) 14-11.60 (br, COO<u>H</u>), 9.31 (1H, br s, O<u>H</u>), 8.71 (1H, d, *J* 8.00, N<u>H</u>), 8.58-7.44 (br, N<u>H</u>), 7.02 (2H, d, *J* 8.48, <u>H</u>$_{ArY}$), 6.67 (2H, d, *J* 8.48, <u>H</u>$_{ArY}$), 4.45-4.39 (1H, m, NH-C<u>H</u>$_F$), 3.55 (1H, d, *J* 16.30, C<u>H</u>$_{2G}$), 3.47 (1H, d, *J* 16.29, C<u>H</u>$_{2G}$), 2.96 (1H, dd, *J* 13.90, 4.88, C<u>H</u>$_{2Y}$), 2.76 (1H, dd, *J* 13.91, 8.86, C<u>H</u>$_{2Y}$). δ_C (100 MHz, DMSO-d_6) 173.16 and 166.23 (<u>C</u>=O), 156.29, 130.77, 127.86, and 115.74 (<u>C</u>$_{ArY}$), 54.71 (NH-<u>C</u>H$_Y$), 40.36 (<u>C</u>H$_{2G}$), 36.51 (<u>C</u>H$_{2F}$). HRMS (APCI) m/z: [M-HCl-H]- calculated 237.0881; found 237.0880.
D19	IY·HCl		δ_H (400 MHz, DMSO-d_6) 9.31 (1H, br, O<u>H</u>), 8.70 (1H, d, *J* 7.57, N<u>H</u>), 7.06 (2H, d, *J* 8.51, <u>H</u>$_{ArY}$), 6.70-6.66 (2H, m, <u>H</u>$_{ArY}$), 4.44-4.35 (1H, m, NH-C<u>H</u>), 3.67 (1H, d, *J* 5.23, NH-C<u>H</u>), 2.95 (1H, dd, *J* 14.08, 5.31, C<u>H</u>$_{2F}$), 2.84 (1H, dd, *J* 14.06, 8.42, C<u>H</u>$_{2F}$), 1.93-1.80 (1H, m, C<u>H</u>$_I$), 1.55-1.42 (1H, m, C<u>H</u>$_{2I}$), 1.17-1.03 (1H, overlapped by diethyl ether, C<u>H</u>$_{2I}$), 0.92 (3H, d, *J* 6.92, C<u>H</u>$_{3I}$), 0.84 (3H, t, *J* 7.34, C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 172.48 and 168.07 (<u>C</u>=O), 156.18, 130.11, 127.25, and 115.16 (<u>C</u>$_{ArY}$), 56.34 and 54.33 (NH-<u>C</u>H), 36.38 (<u>C</u>H$_I$), 35.79 (CH$_2$F), 23.67 (<u>C</u>H$_2$), 14.57 and 11.31 (<u>C</u>H$_{3I}$). HRMS (APCI) m/z: [M-HCl-H]- calculated 293.1507; found 293.1512.

D20	VF·HCl		δ_H (400 MHz, DMSO-d_6) 13.40-12.20 (br, COO<u>H</u>), 8.81 (1H, d, *J* 7.50, N<u>H</u>), 8.62-7.60 (br, N<u>H</u>), 7.33-7.18 (5H, m, <u>H</u>$_{ArF}$), 4.51-4.46 (1H, m, NH-C<u>H</u>), 3.71-3.61 (1H, m, NH-C<u>H</u>), 3.08 (1H, dd, *J* 14.04, 5.25, C<u>H</u>$_{2F}$), 2.96 (1H, dd, *J* 14.05, 8.60, C<u>H</u>$_{2F}$), 2.22-2.08 (1H, m, C<u>H</u>$_V$), 0.95 (3H, d, *J* 6.92, C<u>H</u>$_{3V}$), 0.92 (3H, d, *J* 6.92, C<u>H</u>$_{3V}$). δ_C (100 MHz, DMSO-d_6) 172.29 and 168.03 (<u>C</u>=O), 137.35, 129.21, 128.26, and 126.53 (<u>C</u>$_{ArF}$), 56.88 and 53.93 (NH-<u>C</u>H), 36.43 (<u>C</u>H$_{2F}$), 29.88 (<u>C</u>H$_V$), 18.33 and 17.48 (<u>C</u>H$_{3V}$). HRMS (APCI) m/z: [M-HCl-H]- calculated 263.1401; found 263.1404.
D21	VI·HCl		δ_H (400 MHz, DMSO-d_6) 13.40-11.90 (br, COO<u>H</u>), 8.69-7.62 (br, N<u>H</u>), 8.51 (1H, d, *J* 7.78, N<u>H</u>), 4.19 (1H, dd, *J* 7.72, 5.70, NH-C<u>H</u>$_I$), 3.74 (1H, d, *J* 5.63, NH-C<u>H</u>$_V$), 2.16-2.03 (1H, m, C<u>H</u>$_V$), 1.88-1.74 (1H, m, C<u>H</u>$_I$), 1.52-1.38 (1H, m, C<u>H</u>$_{2I}$), 1.31-1.16 (1H, m, C<u>H</u>$_{2I}$), 0.96 (3H, d, *J* 2.03, C<u>H</u>$_{3V}$), 0.94 (3H, d, *J* 2.01, C<u>H</u>$_3$V), 0.91 (3H, d, *J* 6.86, C<u>H</u>$_{3I}$), 0.87 (3H, t, *J* 7.42, C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 172.30 and 168.08 (<u>C</u>=O), 56.78 and 56.72 (NH-<u>C</u>H), 36.17 (<u>C</u>H$_I$), 29.94 (<u>C</u>H$_V$), 24.72 (<u>C</u>H$_{2I}$), 18.22 and 17.86 (<u>C</u>H$_{3V}$), 15.60 and 11.44 (<u>C</u>H$_{3I}$). HRMS (ESI) m/z: [M-HCl+H]+ calculated 231.1703; found 231.1712.
D22	YG·HCl		δ_H (400 MHz, DMSO-d_6) 9.40 (1H, br s, O<u>H</u>), 8.92 (1H, t, *J* 5.73, N<u>H</u>), 8.83-7.46 (br, N<u>H</u>), 7.09 (2H, d, *J* 8.48, <u>H</u>$_{ArY}$), 6.70 (2H, d, *J* 8.48, <u>H</u>$_{ArY}$), 3.98 (1H, dd, *J* 7.36, 5.59, NH-C<u>H</u>$_Y$), 3.85 (1H, d, *J* 1.67, C<u>H</u>$_{2G}$), 3.83 (1H, d, *J* 1.82, C<u>H</u>$_{2G}$), 3.04 (1H, dd, *J* 14.15, 5.34, C<u>H</u>$_{2F}$), 2.87 (1H, dd, *J* 14.15, 7.58, C<u>H</u>$_{2F}$). δ_C (100 MHz, DMSO-d_6) 170.66 and 168.56 (<u>C</u>=O), 156.64, 130.68, 124.75, and 115.34 (<u>C</u>$_{ArY}$), 53.57 (NH-<u>C</u>H$_Y$), 40.78 (<u>C</u>H$_{2G}$), 36.08 (<u>C</u>H$_{2F}$). HRMS (APCI) m/z: [M-HCl-H]- calculated 237.0881; found 237.0882.

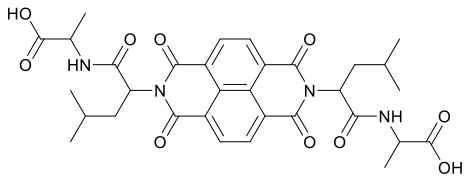
D23	YI·HCl		δ_H (400 MHz, DMSO-d_6) 13.9-11.8 (br, COO<u>H</u>), 9.40 (1H, br s, O<u>H</u>), 8.71 (1H, d, *J* 8.23, N<u>H</u>), 8.62-7.64 (br, N<u>H</u>), 7.08 (2H, d, *J* 8.52, <u>H</u>$_{ArY}$), 6.70 (2H, d, *J* 8.52, <u>H</u>$_{ArY}$), 4.22 (1H, dd, *J* 8.19, 5.77, NH-C<u>H</u>), 4.08 (1H, dd, *J* 7.27, 5.46, NH-C<u>H</u>), 3.04 (1H, dd, *J* 14.22, 5.20, C<u>H</u>$_{2F}$), 2.84 (1H, dd, *J* 14.22, 7.47, C<u>H</u>$_{2F}$), 1.89-1.75 (1H, m, C<u>H</u>$_I$), 1.52-1.38 (1H, m, C<u>H</u>$_{2I}$), 1.30-1.14 (1H, m, C<u>H</u>$_{2I}$), 0.92-0.83 (6H, m, C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 172.33 and 168.27 (<u>C</u>=O), 156.61, 130.65, 124.70, and 115.29 (C$_{ArY}$), 56.62 and 53.26 (NH-<u>C</u>H), 36.46 (<u>C</u>H$_I$), 36.00 (<u>C</u>H$_{2F}$), 24.74 (<u>C</u>H$_{2I}$), 15.60 and 11.42 (<u>C</u>H$_{3I}$). HRMS (APCI) m/z: [M-HCl+H]+ calculated 295.1652; found 295.1664.

A 3.1.5 Rylene Precursor Reagents

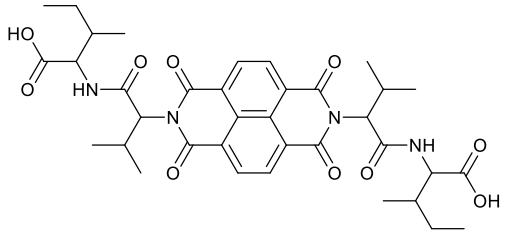
Compound Number	Name	Structure
E1	NTCDA	
E2	PTCDA	

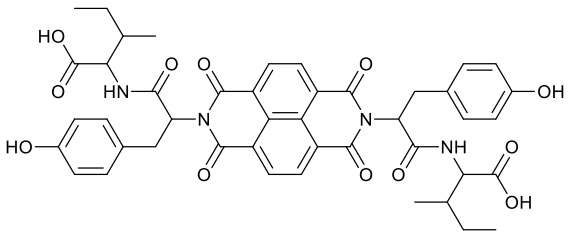
A 3.1.6 Rylene Products

Synthesised and Characterised by Connor MacDonald

Compound Number	Name	Structure	Characterisation
48	NDI-LA		δ_H (400 MHz, DMSO-d_6) 12.49 (2H, s, COO<u>H</u>), 8.71 (4H, d, *J* 2.5, Core <u>H</u>$_{Ar}$), 8.34 – 8.15 (2H, m, N<u>H</u>), 5.55 – 5.36 (2H, m, N-C<u>H</u>$_L$), 4.39 – 4.17 (2H, m, N-C<u>H</u>$_A$), 2.27 – 2.12 (2H, m, C<u>H</u>$_{2L}$), 2.10 – 1.81 (2H, m, C<u>H</u>$_{2L}$), 1.56 – 1.43 (2H, m, C<u>H</u>$_L$), 1.24 (1H, dd, *J* 7.3, 2.4, C<u>H</u>$_{3A}$), 1.08 (5H, d, *J* 7.3, C<u>H</u>$_{3A}$), 0.91 (6H, d, *J* 6.5, C<u>H</u>$_{3L}$), 0.84 (6H, dd, *J* 6.8, 1.8, C<u>H</u>$_{3L}$). δ_C (101 MHz, DMSO-d_6) 174.3, 168.3, and 162.8 (<u>C</u>=O), 130.5, 126.7, 126.5 (Core C$_{Ar}$), 52.5 (N-<u>C</u>H$_L$), 47.6 (N-<u>C</u>H$_A$), 37.2 (<u>C</u>H$_{2L}$), 24.9 (<u>C</u>H$_L$), 23.1 and 22.1 (<u>C</u>H$_{3L}$), 17.0 (<u>C</u>H$_{3A}$). HRMS (ESI) m/z: [M+H]+ calculated 637.2504, found 637.2503.

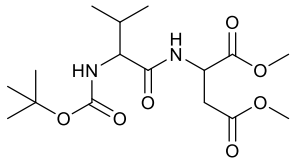
Synthesised and Characterised by Dr Bart Dietrich

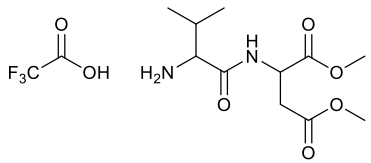
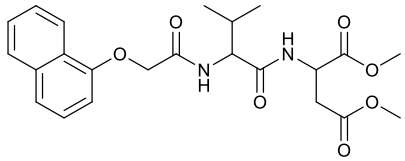
Compound Number	Name	Structure	Characterisation
56	NDI-VI		δ_H (400 MHz, DMSO-d_6) 12.53 (2H, br s, COO<u>H</u>), 8.71 (4H, s, Core <u>H</u>$_{Ar}$), 8.22 (2H, d, *J* 8.38, N<u>H</u>), 5.04-5.00 (2H, m, NH-C<u>H</u>), 4.23-4.19 (2H, m, NH-C<u>H</u>), 2.81-2.66 (2H, m, aliphatic C<u>H</u>), 1.67-1.55 (2H, m, aliphatic C<u>H</u>), 1.23-1.10 (8H, m, aliphatic C<u>H</u>), 0.98-0.84 (2H, m, aliphatic C<u>H</u>), 0.81-0.56 (18H, m, aliphatic C<u>H</u>). δ_C (100 MHz, DMSO-d_6) 173.53, 167.91, and 162.92 (<u>C</u>=O), 130.54, 126.64, and 126.46 (Core <u>C</u>$_{Ar}$), 58.72 and 56.32 (NH-<u>C</u>H), 35.51, 26.93, 24.70, 22.69, 18.63, 15.50, and 10.53 (aliphatic <u>C</u>). HRMS (ESI) m/z: [M+H]+ calculatedd 693.313020; found 693.312668.

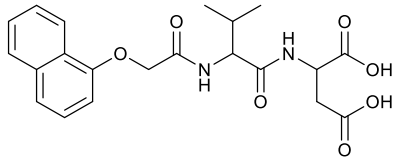
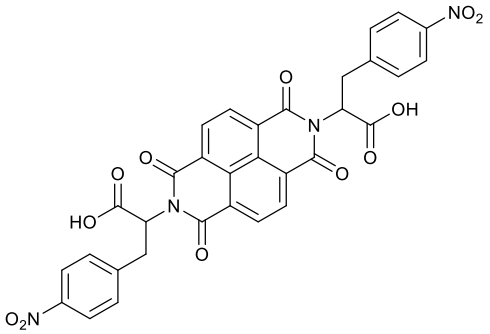
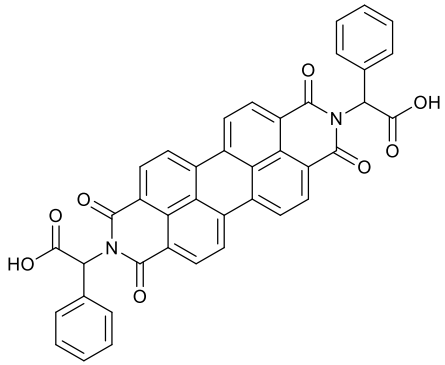
49	NDI-YI		δ_H (400 MHz, DMSO-d_6) 12.54 (2H, s, COO<u>H</u>), 9.00 (2H, s, O<u>H</u>), 8.57 (4H, s, Core <u>H</u>$_{Ar}$), 8.20 (1.7H, d, *J* 8.44, N<u>H</u>), 8.13 (0.3H, d, *J* 8.89, N<u>H</u>), 6.92-6.85 (4H, m, <u>H</u>$_{ArY}$), 6.47-6.39 (4H, m, <u>H</u>$_{ArY}$), 5.69 (0.3H, dd, *J* 9.85, 5.72, NH-C<u>H</u>), 5.61 (1.7H, dd, *J* 9.08, 5.84, NH-C<u>H</u>), 4.31-4.21 (2H, m, C<u>H</u>$_L$), 3.59-3.48 (2H, m, C<u>H</u>$_{2Y}$), 3.21 (0.3H, dd, *J* 14.19, 10.04, C<u>H</u>$_{2Y}$), 3.10 (1.7H, dd, *J* 13.94, 9.40, C<u>H</u>$_{2Y}$), 1.77-1.70 (0.3H, m, <u>H</u>$_I$), 1.70-1.59 (1.7H, m, <u>H</u>$_I$), 1.45-1.33 (0.3H, m, <u>H</u>$_I$), 1.33-1.17 (1.7H, m, <u>H</u>$_I$), 1.16-1.05 (0.3H, m, <u>H</u>$_I$), 1.05-0.90 (1.7H, m, <u>H</u>$_I$), 0.87-0.64 (12H, m, C<u>H</u>$_{3I}$). δ_C (100 MHz, DMSO-d_6) 173.36, 168.16, and 162.51 (<u>C</u>=O), 155.48, 130.43, 130.13, 130.05, 129.88, 128.25, 128.01, 126.67, 126.44, 126.20, 126.07, and 114.89 (<u>C</u>$_{Ar}$), 56.48 and 55.46 (NH-<u>C</u>H), 36, 04, 35.70, 33.47, 24.76, 24.66, 15.51, 10.84, and 10.66 (aliphatic <u>C</u>). HRMS (ESI) m/z: [M-H]- calculated 819.2883; found 819.2888.

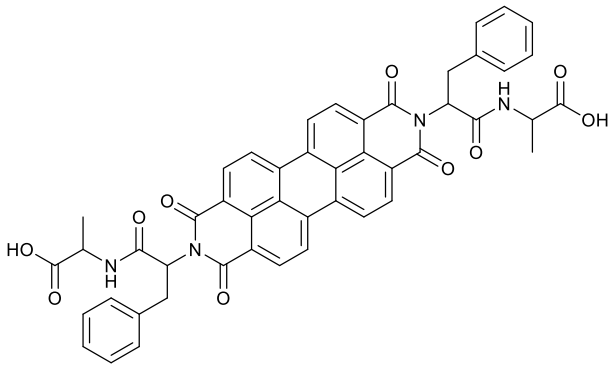
A 3.1.7 Validation Products and Precursors

<u>Synthesised and Characterised by Dr Bart Dietrich</u>

Compound Number	Name	Structure	Characterisation
G1	Boc-VD-OMe		δ_H (400 MHz, DMSO-d_6) 8.30 (1H, d, *J* 7.7, N<u>H</u>), 6.66 (0.85H, d, *J* 8.9, N<u>H</u>), 6.30 (0.15H, d, *J* 7.0, N<u>H</u>), 4.66–4.61 (1H, m, C<u>H</u>-NH), 3.81–3.77 (1H, m, C<u>H</u>-NH), 3.61 (3H, s, O-C<u>H</u>$_3$), 3.60 (3H, s, O-C<u>H</u>$_3$), 2.82 (1H, dd, *J* 16.5, 5.9, C<u>H</u>$_2$), 2.71 (1H, dd, *J* 16.5, 7.3, C<u>H</u>$_2$), 1.96–1.87 (1H, C<u>H</u>(CH$_3$)$_2$), 1.38 (9H, s, C(C<u>H</u>$_3$)$_3$), 0.84 (3H, d, *J* 6.7, CH(C<u>H</u>$_3$)$_2$), 0.80 (3H, d, *J* 6.7, CH(C<u>H</u>$_3$)$_2$). δ_C (100 MHz, DMSO-d_6) 171.3, 171.0, 170.3, and 155.3 (<u>C</u>=O), 78.0 (<u>C</u>(CH$_3$)$_3$), 59.4 (<u>C</u>H-NH), 52.0 and 51.6 (O-<u>C</u>H$_3$), 48.4 (<u>C</u>H-NH), 35.5 (<u>C</u>H$_2$), 30.4 (CH(<u>C</u>H$_3$)$_2$), 28.1 (C(<u>C</u>H$_3$)$_3$), 19.0 and 17.9 (CH(<u>C</u>H$_3$)$_2$). HRMS (ESI/QTOF) m/z: $[M+Na]^+$ calculated 383.178872; found: 383.180287.

G2	VD-OMe ·TFA		δ_H (400 MHz, DMSO-d_6) 8.90 (1H, d, *J* 7.6, N<u>H</u>), 8.09 (3H, br s, N<u>H</u>$_3^+$), 4.75-4.70 (1H, m, C<u>H</u>*), 3.66-3.63 (1H, m, C<u>H</u>*), 3.64 (3H, s, OC<u>H</u>$_3$), 3.63 (3H, s, OC<u>H</u>$_3$), 2.84 (1H, d, *J* 1.0, C<u>H</u>$_2$), 2.82 (1H, d, *J* 2.1, C<u>H</u>$_2$), 2.14-2.05 (1H, m, C<u>H</u>(CH$_3$)$_2$), 0.96 (3H, d, *J* 7.1, CH(C<u>H</u>$_3$)$_2$), 0.94 (3H, d, *J* 7.1, CH(C<u>H</u>$_3$)$_2$). δ_C (100 MHz, DMSO-d_6) *δ*: 170.56, 170.27, and 167.99 (<u>C</u>=O), 158.16 (q, *J* 30.5, O<u>C</u>-CF$_3$), 117.24 (q, *J* 300.05, <u>C</u>F$_3$), 57.10 (<u>C</u>H*), 52.26 and 51.77 (O<u>C</u>H$_3$), 48.54 (<u>C</u>H*), 35.21 (<u>C</u>H$_2$), 29.83 (<u>C</u>H(CH$_3$)$_2$), 18.02 and 17.39 (CH(<u>C</u>H$_3$)$_2$). HRMS (ESI/QTOF) m/z: $[M\text{-TFA}+H]^+$ calculated 261.144498; found 261.144725.
G3	1Nap-VD-OMe		δ_H (400 MHz, DMSO-d_6) 8.63 (1H, d, *J* 7.7, N<u>H</u>), 8.24–8.19 (1H, m, <u>H</u>$_{Ar}$), 8.00 (1H, d, *J* 9.0, N<u>H</u>), 7.92–7.86 (1H, m, <u>H</u>$_{Ar}$), 7.59–7.48 (3H, m, <u>H</u>$_{Ar}$), 7.40 (1H, t, *J* 8.0, <u>H</u>$_{Ar}$), 6.91 (1H, d, *J* 7.4, <u>H</u>$_{Ar}$), 4.82 (1H, d, *J* 14.6, O-C<u>H</u>$_2$), 4.78 (1H, d, *J* 14.5, O-C<u>H</u>$_2$), 4.69–4.63 (1H, m, C<u>H</u>-NH), 4.34 (1H, dd, *J* 9.0, 6.2, C<u>H</u>-NH), 3.62 (3H, s, O-C<u>H</u>$_3$), 3.60 (3H, s, O-C<u>H</u>$_3$), 2.84 (1H, dd, *J* 16.5, 5.9, CH-C<u>H</u>$_2$), 2.73 (1H, dd, *J* 16.5, 7.5, CH-C<u>H</u>$_2$), 2.09–1.97 (1H, m, C<u>H</u>(CH$_3$)$_2$), 0.87 (3H, d, *J* 6.8, CH(C<u>H</u>$_3$)$_2$), 0.83 (3H, d, *J* 6.8, CH(C<u>H</u>$_3$)$_2$). δ_C (100 MHz, DMSO-d_6) 170.9, 170.6, 170.3, and 167.3 (<u>C</u>=O), 153.1, 134.1, 127.5, 126.6, 126.0, 125.5, 124.8, 121.4, 120.6, and 105.6 (<u>C</u>$_{Ar}$), 66.7 (O-<u>C</u>H$_2$), 56.7 (<u>C</u>H-NH), 52.1 and 51.7 (O-<u>C</u>H$_3$), 48.5 (<u>C</u>H-NH), 35.3 (CH-<u>C</u>H$_2$), 31.1 (<u>C</u>H(CH$_3$)$_2$), 19.0 and 17.6 (CH(<u>C</u>H$_3$)$_2$). HRMS (ESI/QTOF) m/z: $[M+Na]^+$ calculated 467.178872; found: 467.180725.

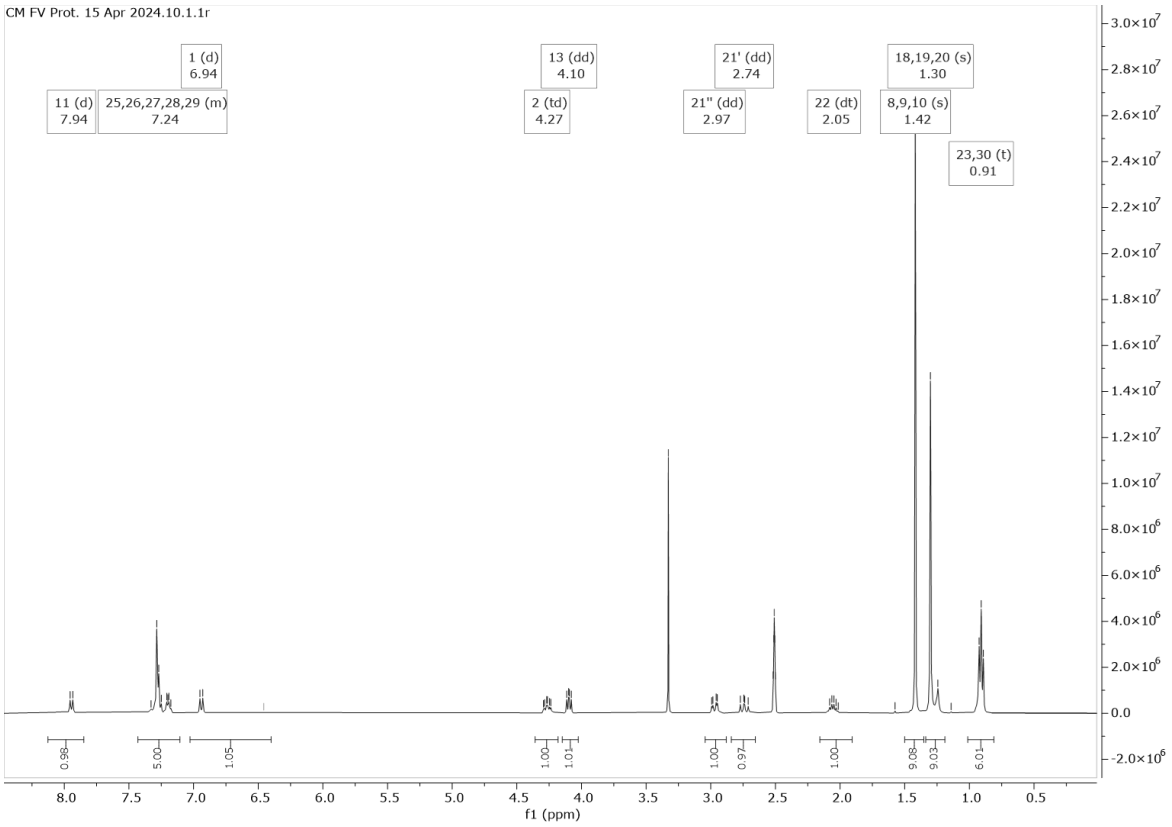
54	1Nap-VD		δ_H (400 MHz, DMSO-d_6) 12.59 (1H, br, COO<u>H</u>), 8.47 (1H, d, *J* 7.8, N<u>H</u>), 8.25-8.19 (1H, m, <u>H</u>$_{Ar}$), 7.98 (1H, d, *J* 9.1, N<u>H</u>), 7.92-7.86 (1H, m, <u>H</u>$_{Ar}$), 7.58-7.49 (3H, m, <u>H</u>$_{Ar}$), 7.40 (1H, t, *J* 8.0, <u>H</u>$_{Ar}$), 6.92 (1H, dd, *J* 7.6, 0.5, <u>H</u>$_{Ar}$), 4.82 (1H, d, *J* 14.5, OC<u>H</u>$_2$), 4.77 (1H, d, *J* 14.5, OC<u>H</u>$_2$), 4.58-4.52 (1H, m, C<u>H</u>*), 4.37 (1H, dd, *J* 9.1, 5.9, C<u>H</u>*), 2.72 (1H, dd, *J* 16.6, 5.7, CH*C<u>H</u>$_2$), 2.59 (1H, dd, *J* 16.6, 7.4, CH*C<u>H</u>$_2$), 2.11-1.97 (1H, m, C<u>H</u>(CH$_3$)$_2$), 0.87 (3H, d, *J* 6.8, CH(C<u>H</u>$_3$)$_2$), 0.82 (3H, d, *J* 6.8, CH(C<u>H</u>$_3$)$_2$). δ_C NMR (101 MHz, DMSO) 172.2, 171.5, 170.4, and 167.2 (<u>C</u>=O), 153.1, 134.1, 127.5, 126.6, 126.0, 125.5, 124.8, 121.4, 120.7, and 105.7 (<u>C</u>$_{Ar}$), 67.0 (O-<u>C</u>H$_2$), 56.6 (<u>C</u>H-NH), 48.7 (<u>C</u>H-NH), 35.8 (CH-<u>C</u>H$_2$), 31.2 (<u>C</u>H(CH$_3$)$_2$), 19.1 and 17.6 (CH(<u>C</u>H$_3$)$_2$). HRMS (ESI/QTOF) m/z: $[M+H]^+$ calculated 651.1005; found: 651.0969.
55	NDI-Phe-NO_2		δ_H (400 MHz, DMSO-d_6) 13.15 (2H, br, COO<u>H</u>), 8.66 (4H, s, <u>H</u>$_{Ar}$), 8.03 (4H, d, *J* 8.7, <u>H</u>$_{Ar}$), 7.52 (4H, d, *J* 8.7, <u>H</u>$_{Ar}$), 5.92 (2H, dd, *J* 8.7, 5.7, C<u>H</u>-NH), 3.76 (2H, dd, *J* 14.1, 5.6, Ph-C<u>H</u>$_2$), 3.44 (2H, dd, *J* 14.1, 8.8, Ph-C<u>H</u>$_2$). δ_C (101 MHz, DMSO) 169.9 and 162.1 (<u>C</u>=O), 146.4, 146.2, 131.3, 130.4, 126.2, 125.7, and 123.3 (<u>C</u>$_{Ar}$), 54.2 (<u>C</u>H-NH), 34.3 (<u>C</u>H$_2$). HRMS (ESI/QTOF) m/z: $[M-H]^-$ calculated 417.1656; found: 417.1670.
G4			Could not be isolated based on literature method [47]

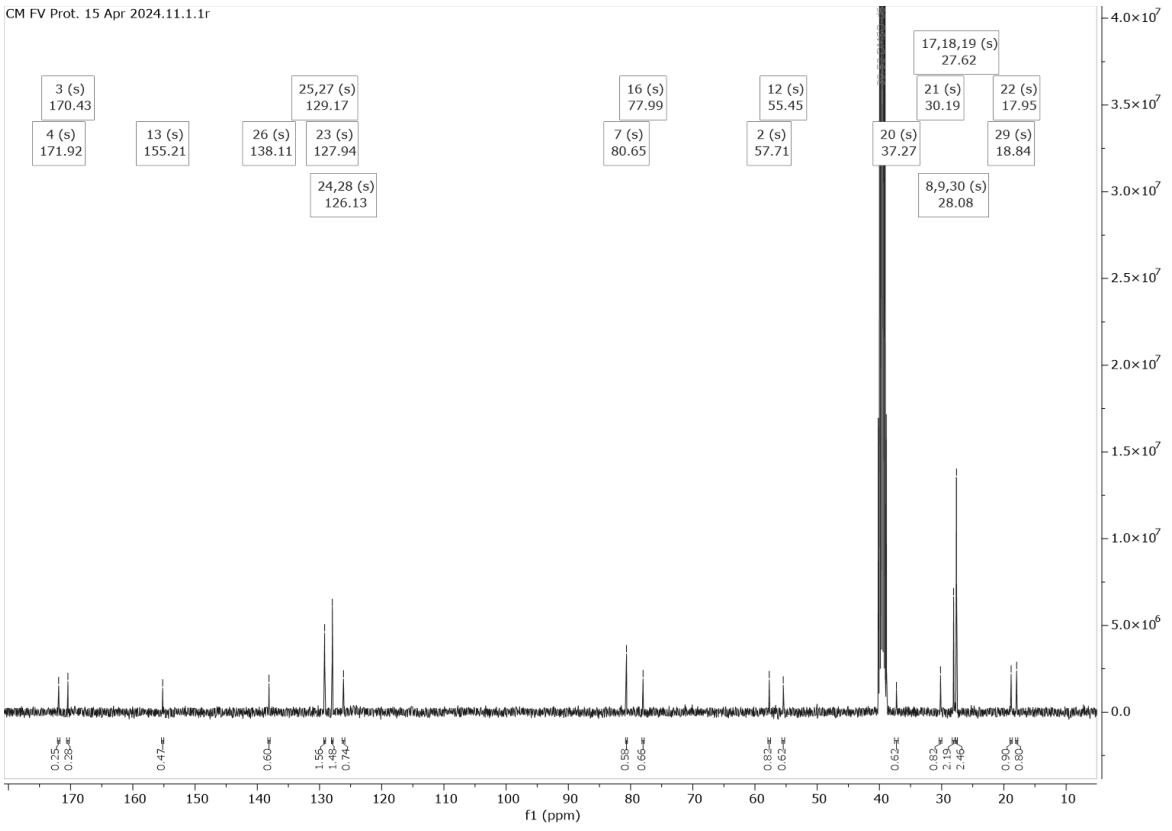
G5			Could not be isolated based on literature method [48]

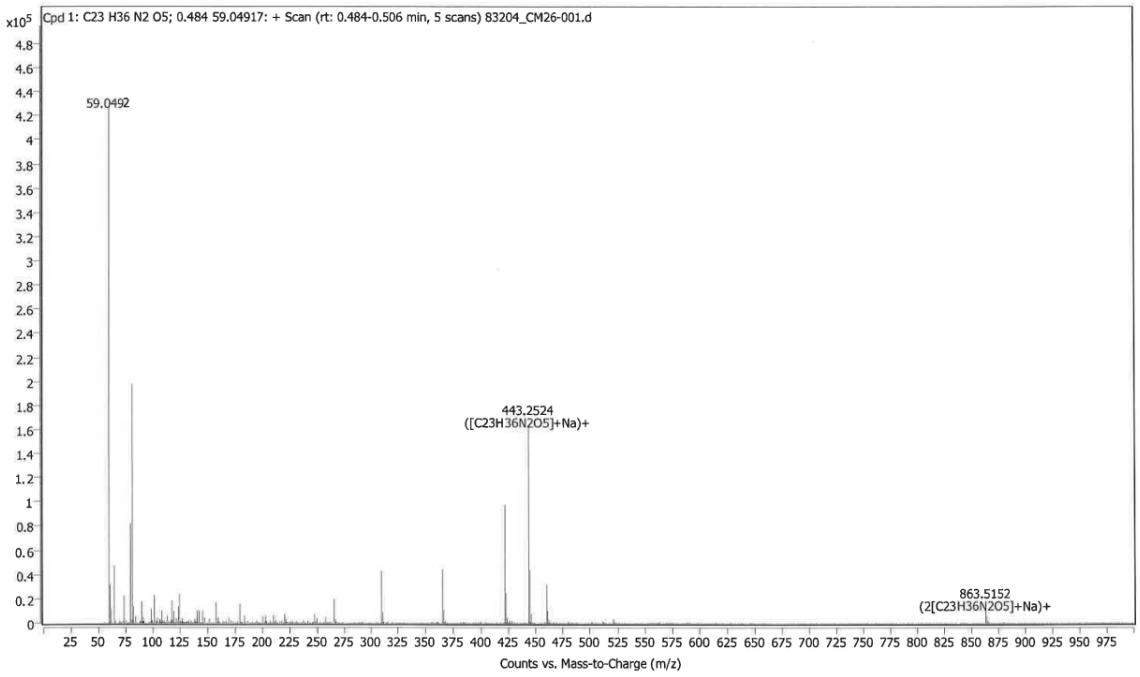
A 3.2 NMR and HRMS Spectra

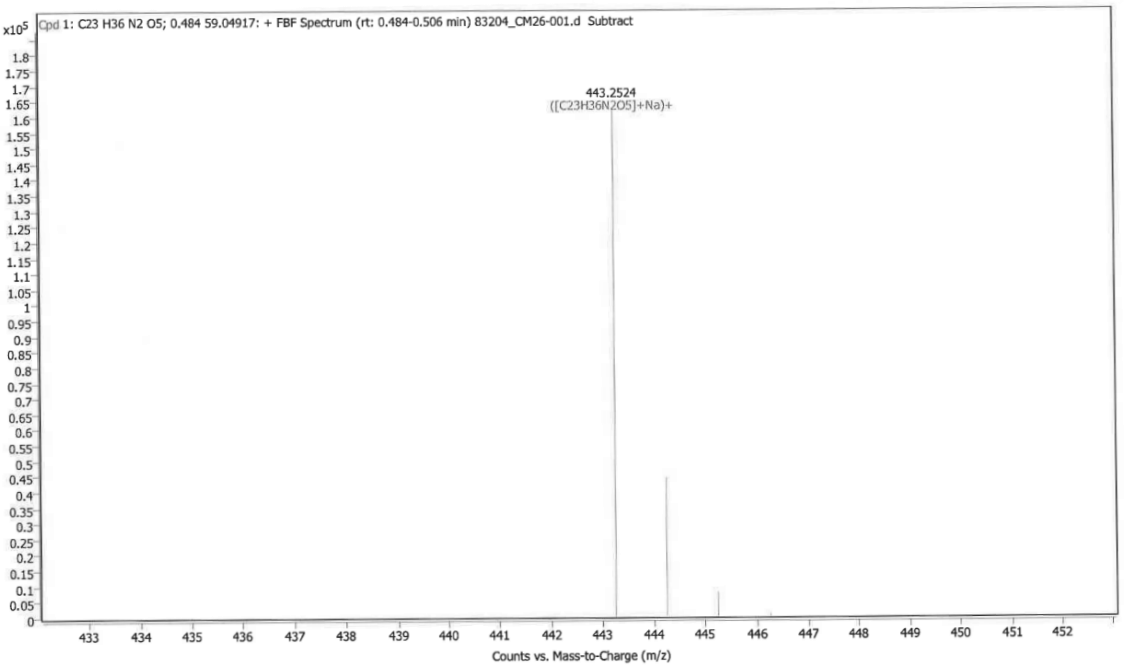
A 3.2.1 Protected Dipeptide Products

Boc-FV-O^{t}Bu **C1**

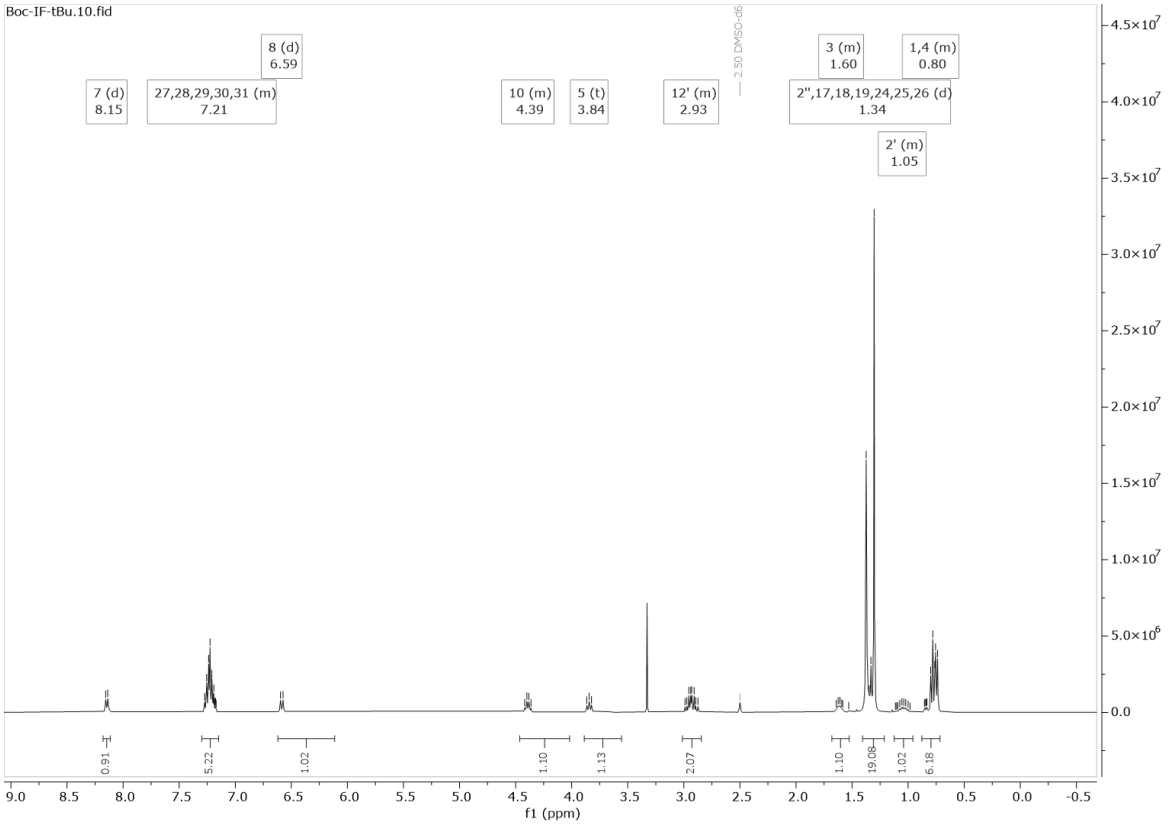


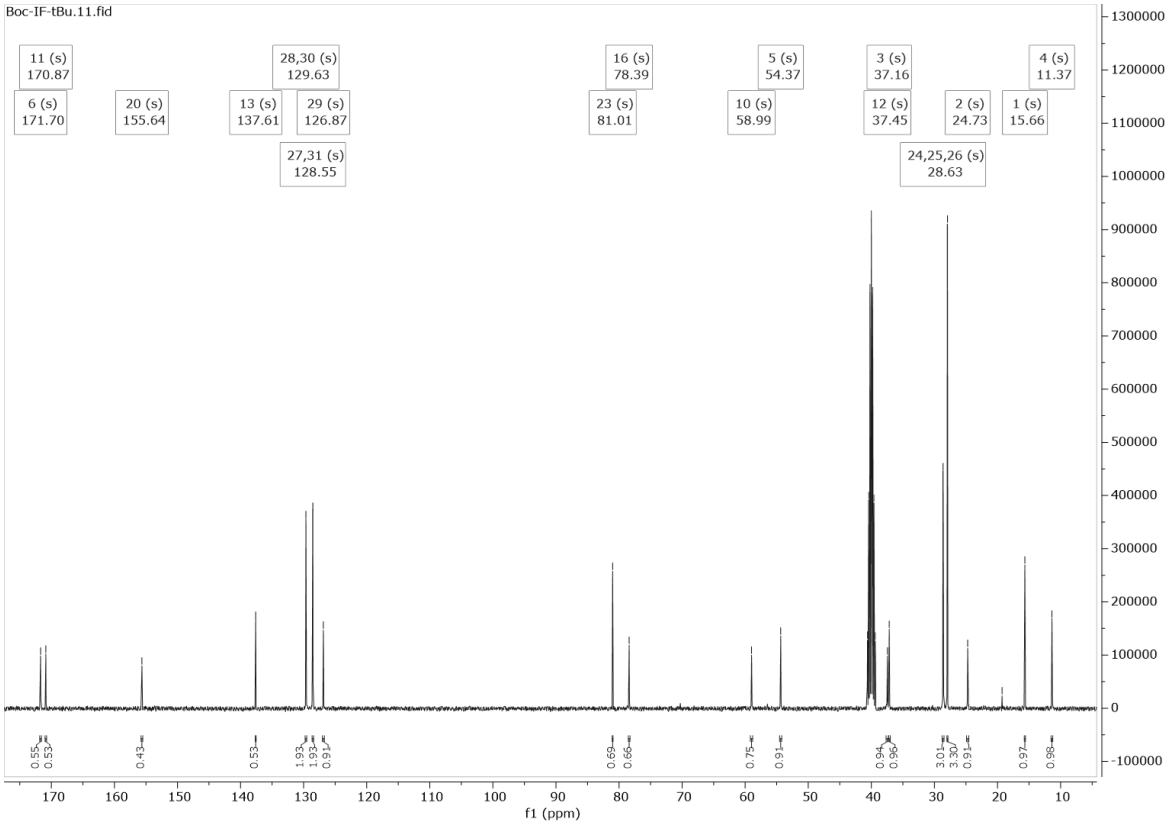


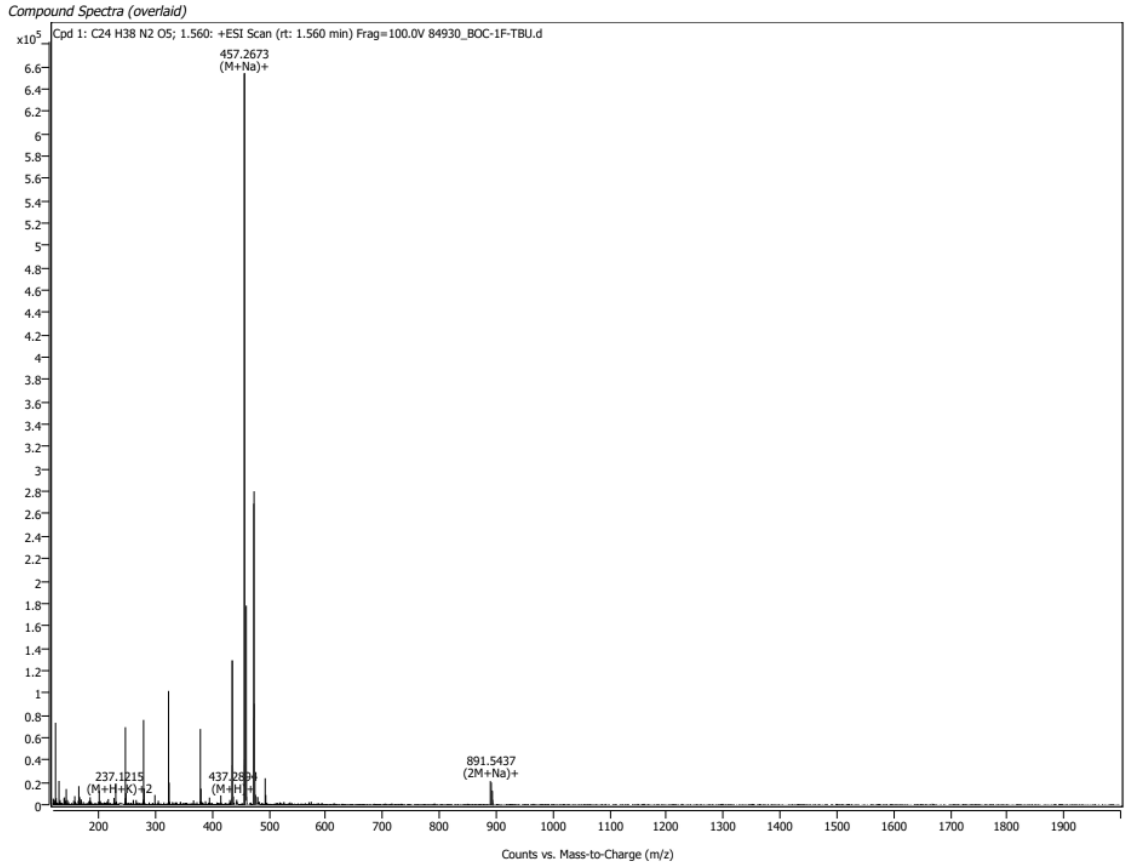

Cpd 1: C23 H36 N2 O5; 0.484 59.04917: + Scan (rt: 0.484-0.506 min, 5 scans) 83204_CM26-001.d
x10^5
59.0492
443.2524
([C23H36N2O5]+Na)+
863.5152
(2[C23H36N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

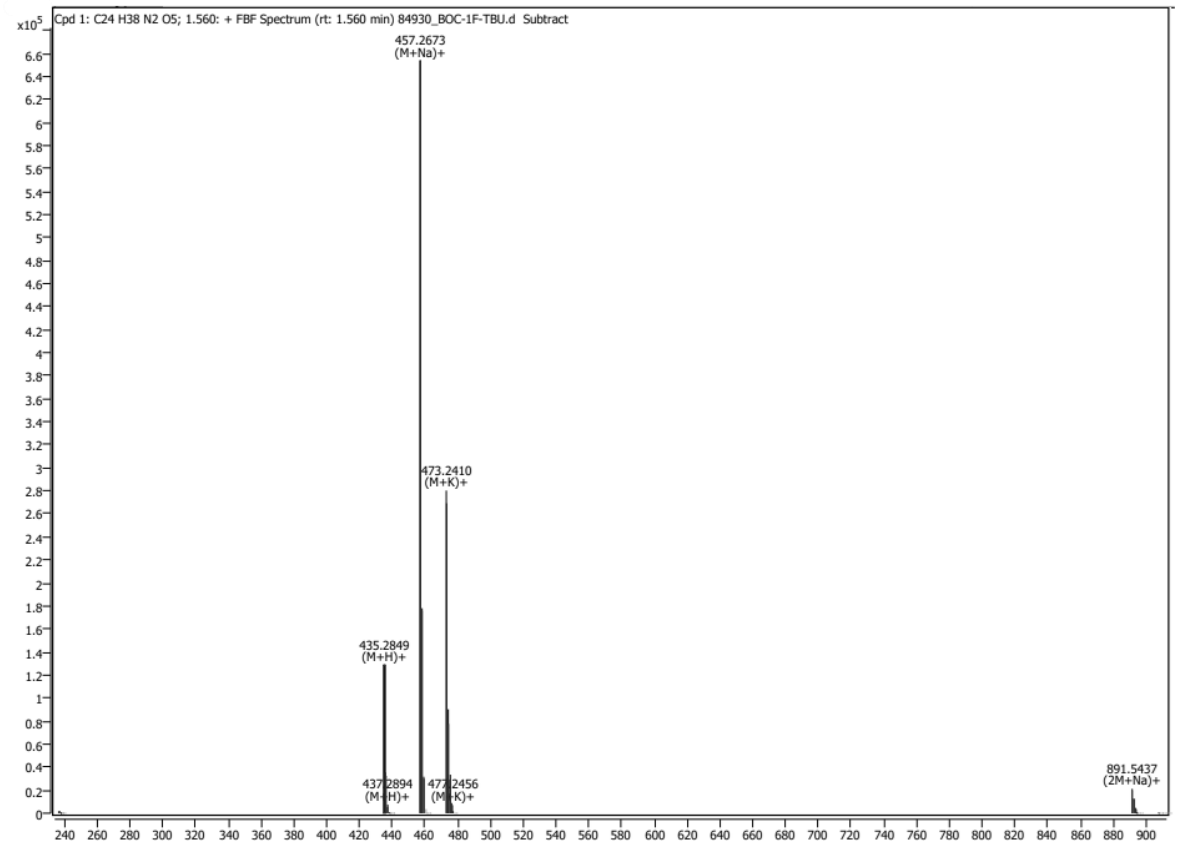

Cpd 1: C23 H36 N2 O5; 0.484 59.04917: + FBF Spectrum (rt: 0.484-0.506 min) 83204_CM26-001.d Subtract
x10^5
443.2524
([C23H36N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-IF-O^{t}Bu **C2**

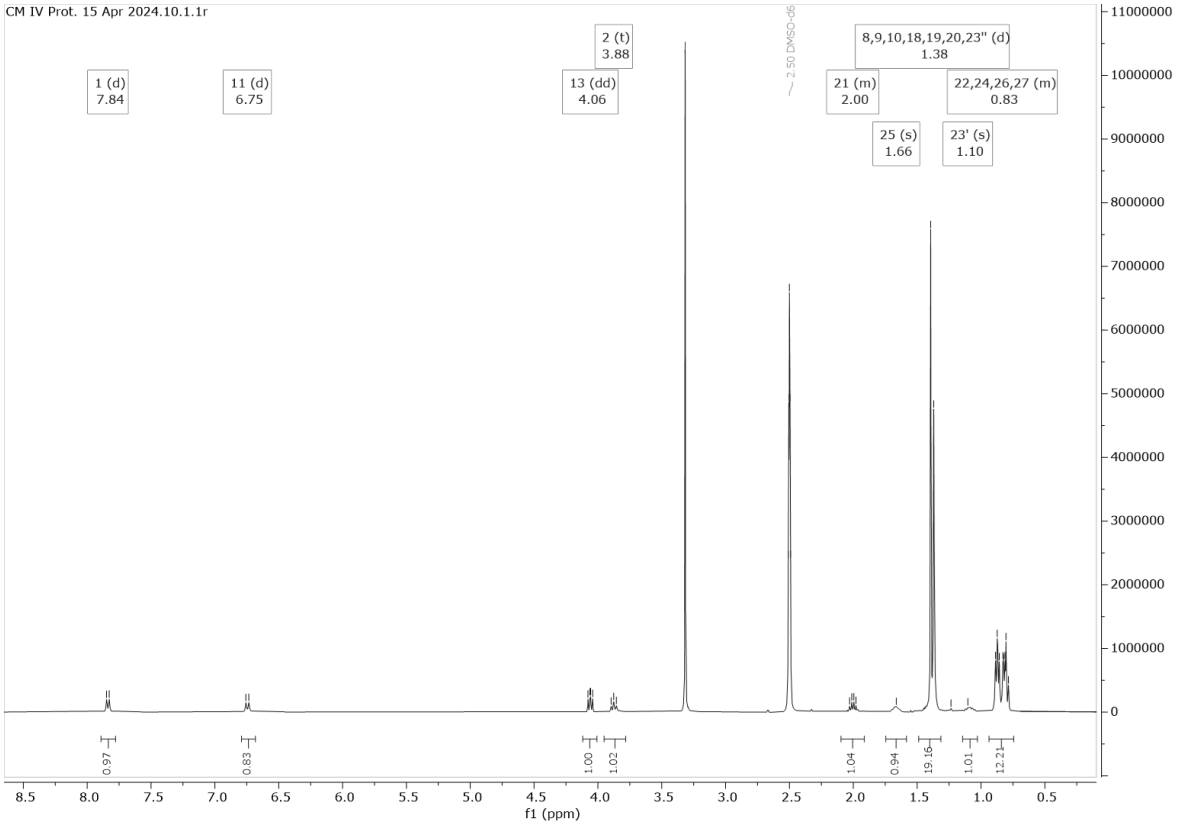


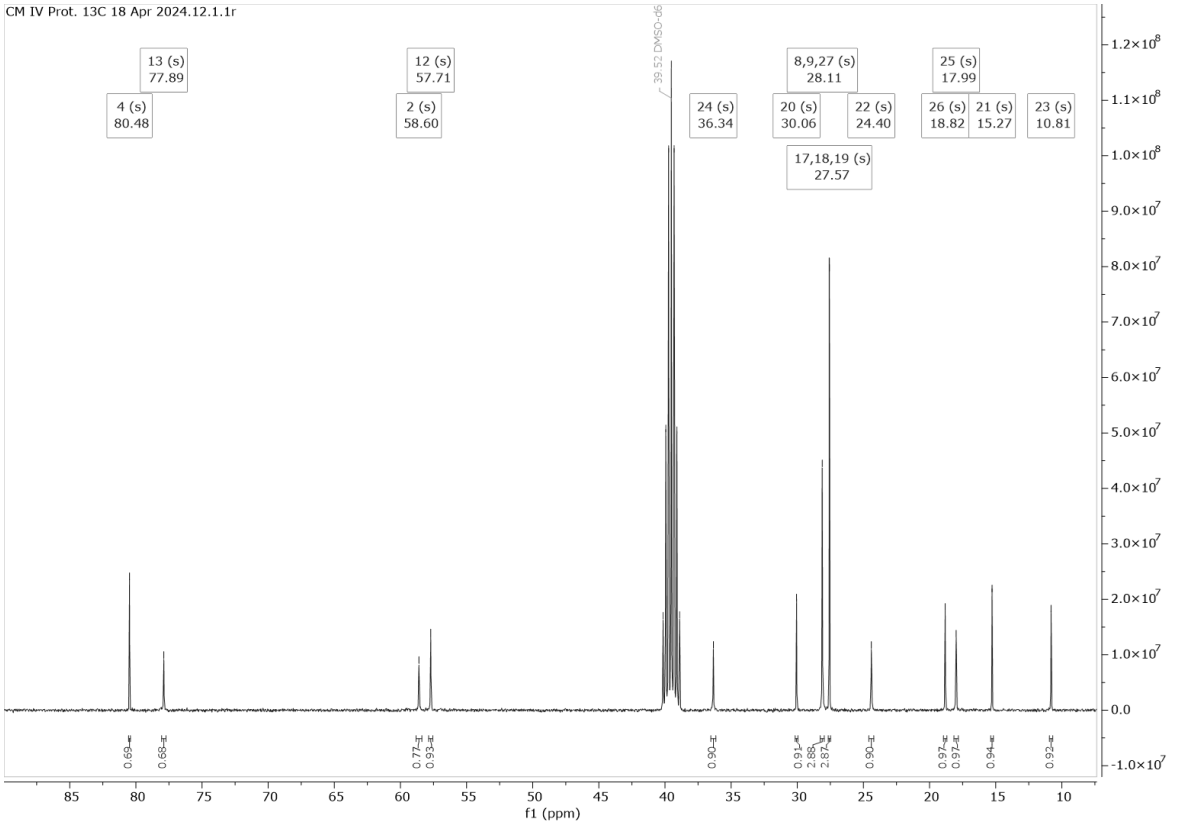


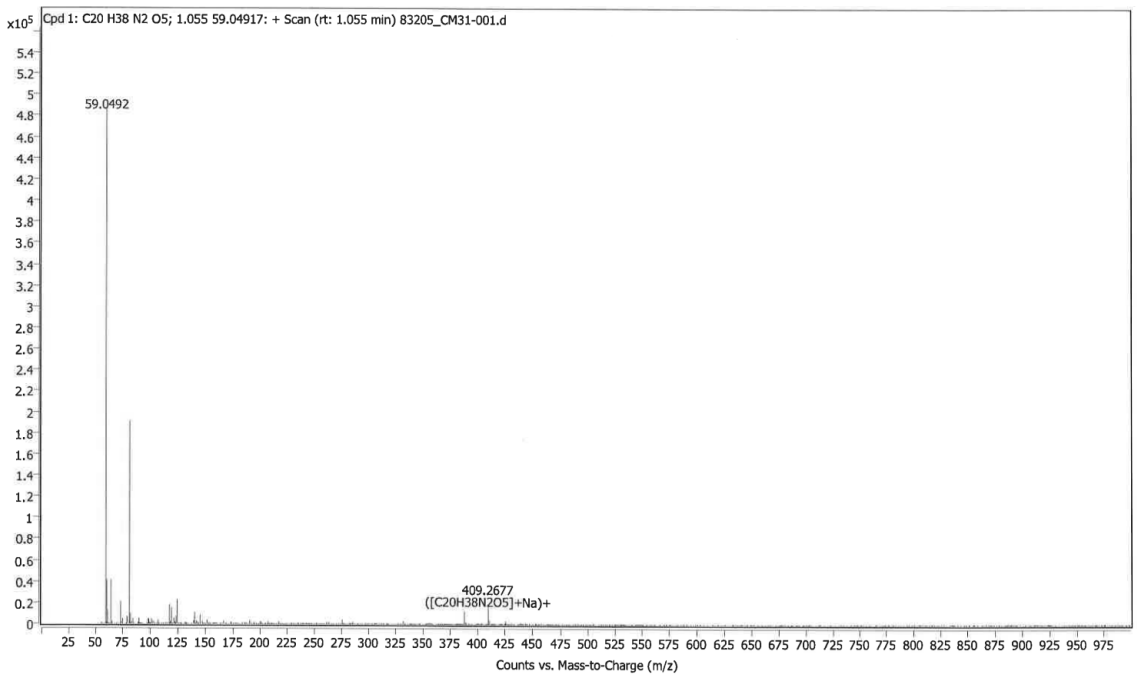

Compound Spectra (overlaid)
Cpd 1: C24 H38 N2 O5; 1.560: +ESI Scan (rt: 1.560 min) Frag=100.0V 84930_BOC-1F-TBU.d
x10^5
457.2673
(M+Na)+
891.5437
(2M+Na)+
237.1215
(M+H+K)+2
437.2894
(M+H)+
Counts vs. Mass-to-Charge (m/z)

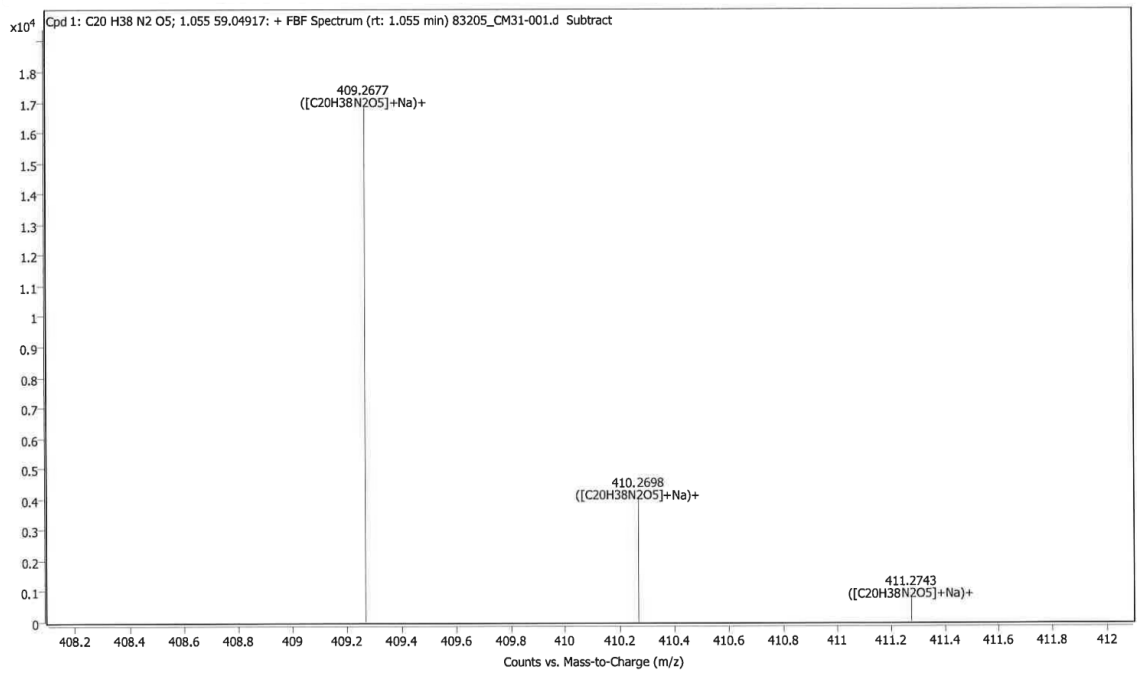

Cpd 1: C24 H38 N2 O5; 1.560: + FBF Spectrum (rt: 1.560 min) 84930_BOC-1F-TBU.d Subtract
x10^5
457.2673
(M+Na)+
473.2410
(M+K)+
435.2849
(M+H)+
437.2894
(M+H)+
473.2456
(M+K)+
891.5437
(2M+Na)+

Boc-IV-O^{t}Bu **C3**

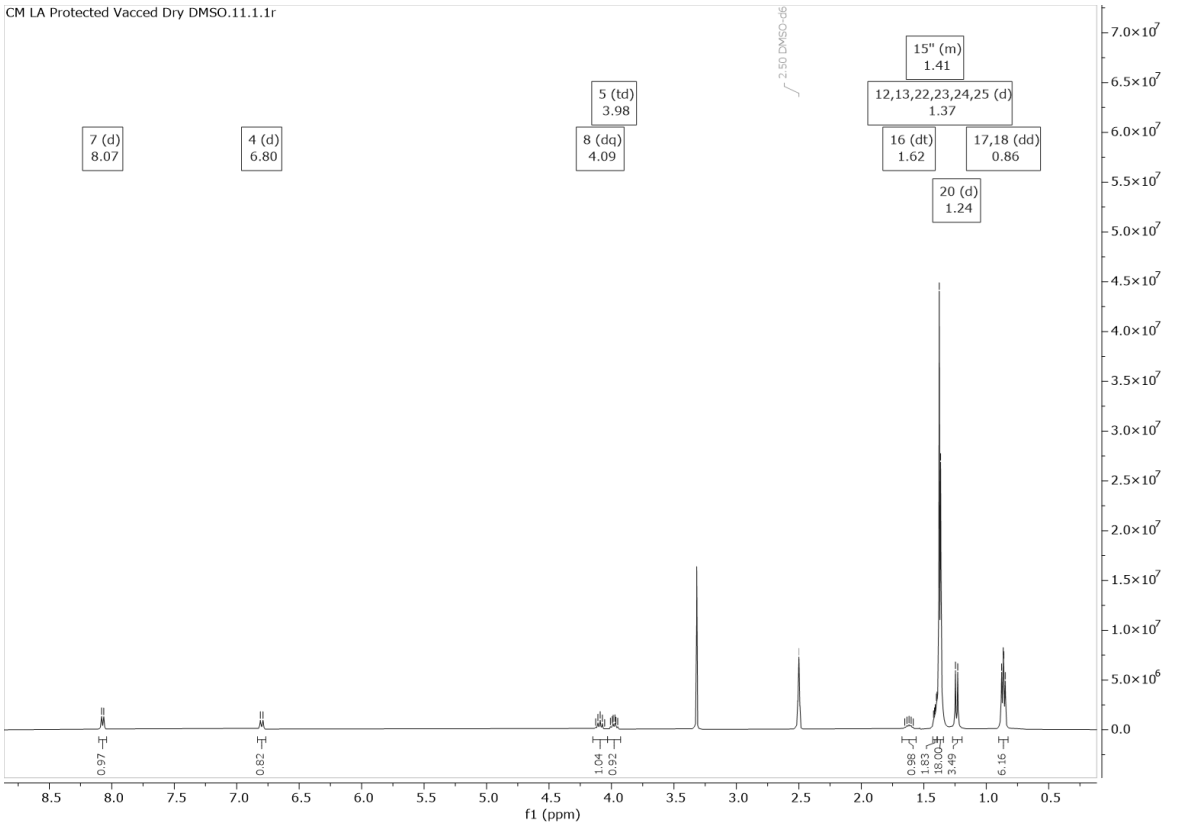


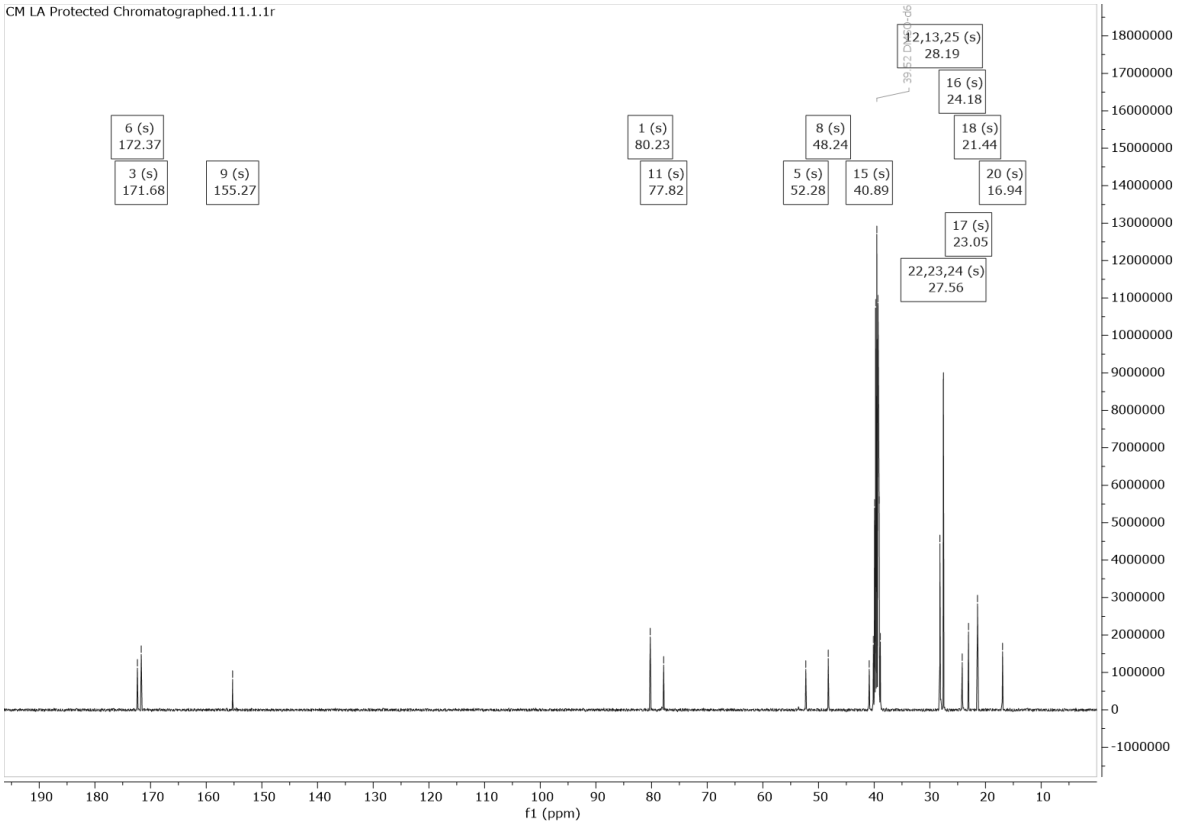


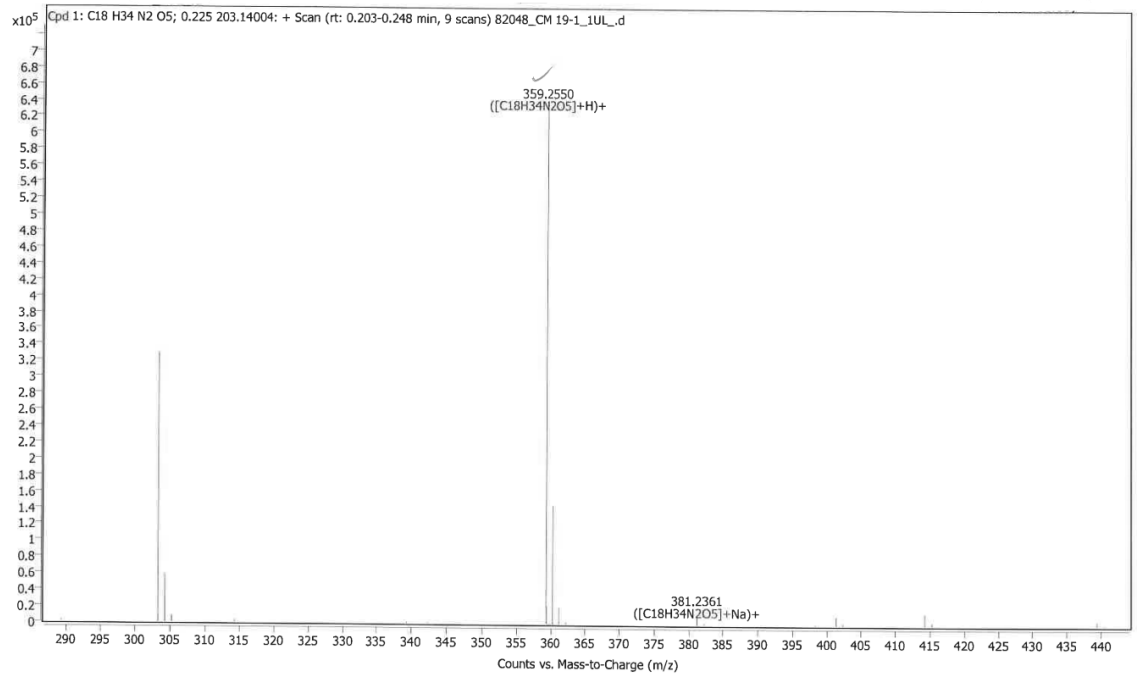

Cpd 1: C20 H38 N2 O5; 1.055 59.04917: + Scan (rt: 1.055 min) 83205_CM31-001.d
x10^5
59.0492
409.2677
([C20H38N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

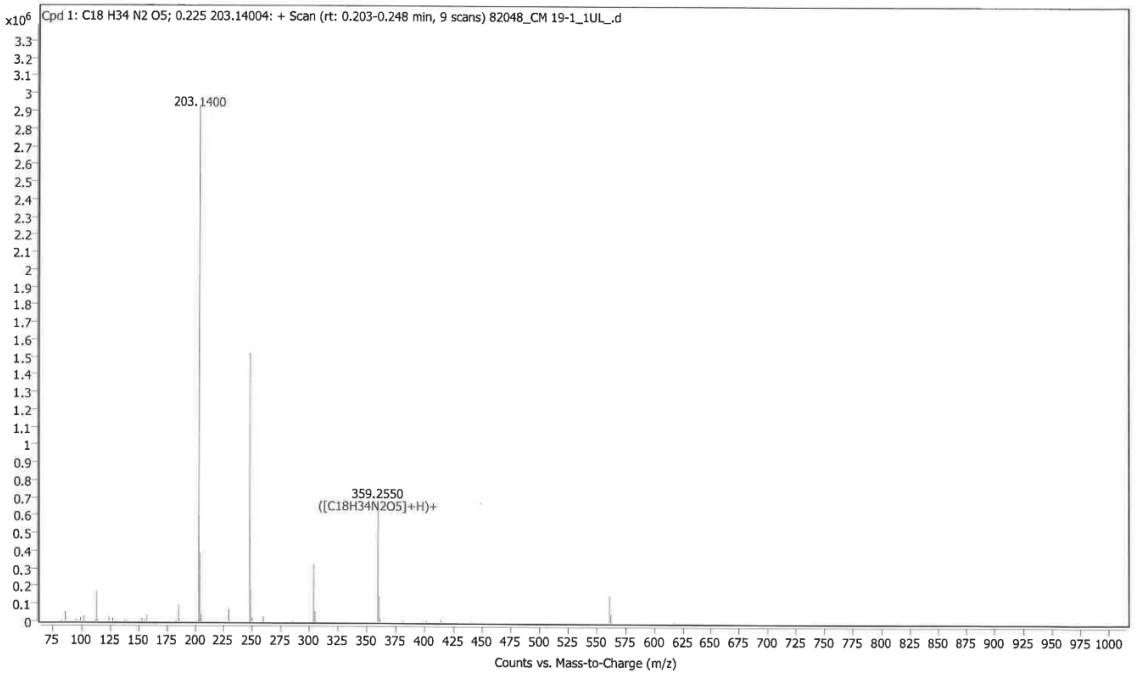

Cpd 1: C20 H38 N2 O5; 1.055 59.04917: + FBF Spectrum (rt: 1.055 min) 83205_CM31-001.d Subtract
x10^4
409.2677
([C20H38N2O5]+Na)+
410.2698
([C20H38N2O5]+Na)+
411.2743
([C20H38N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-LA-O^{t}Bu **C4**

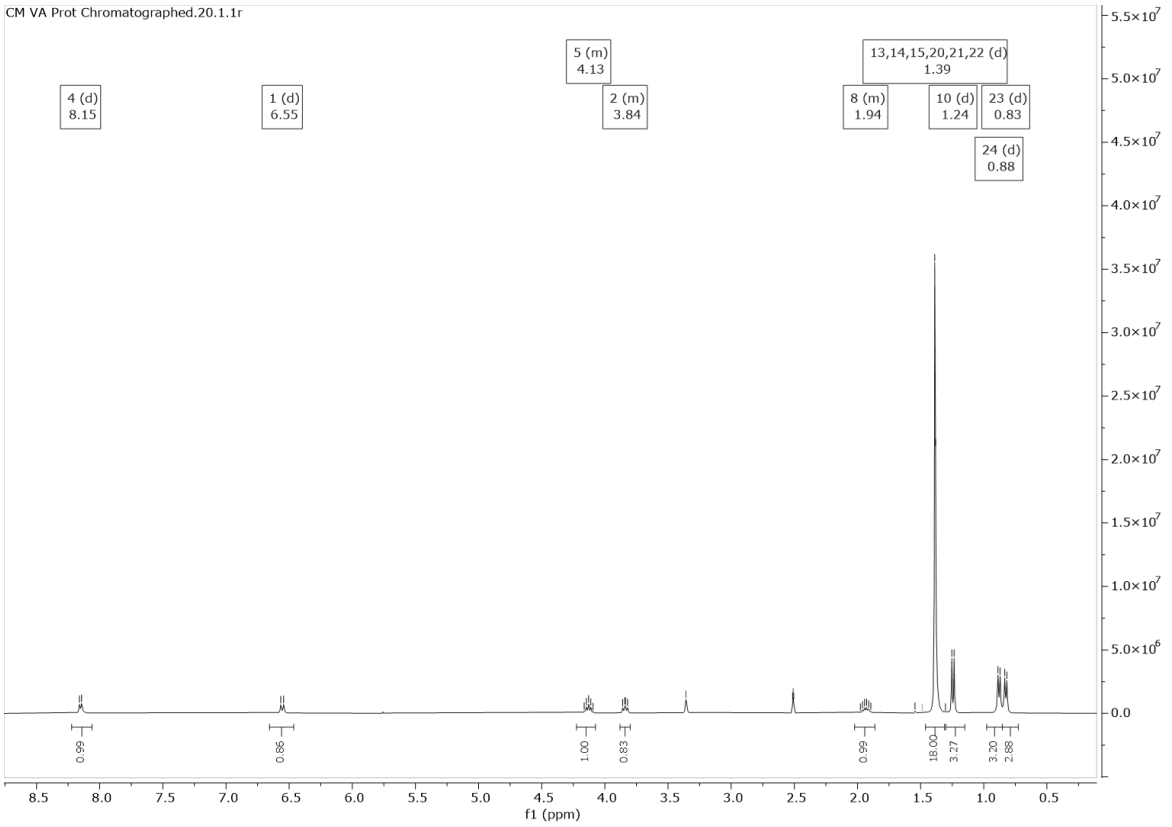


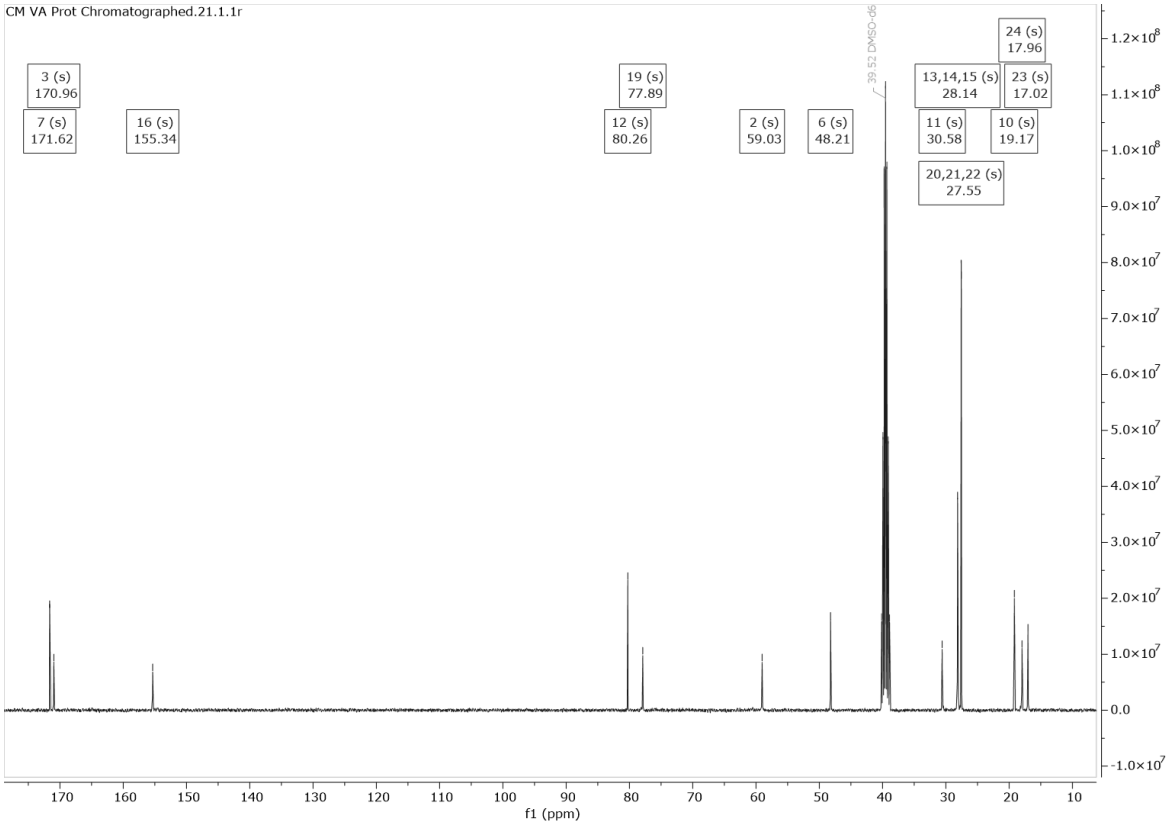


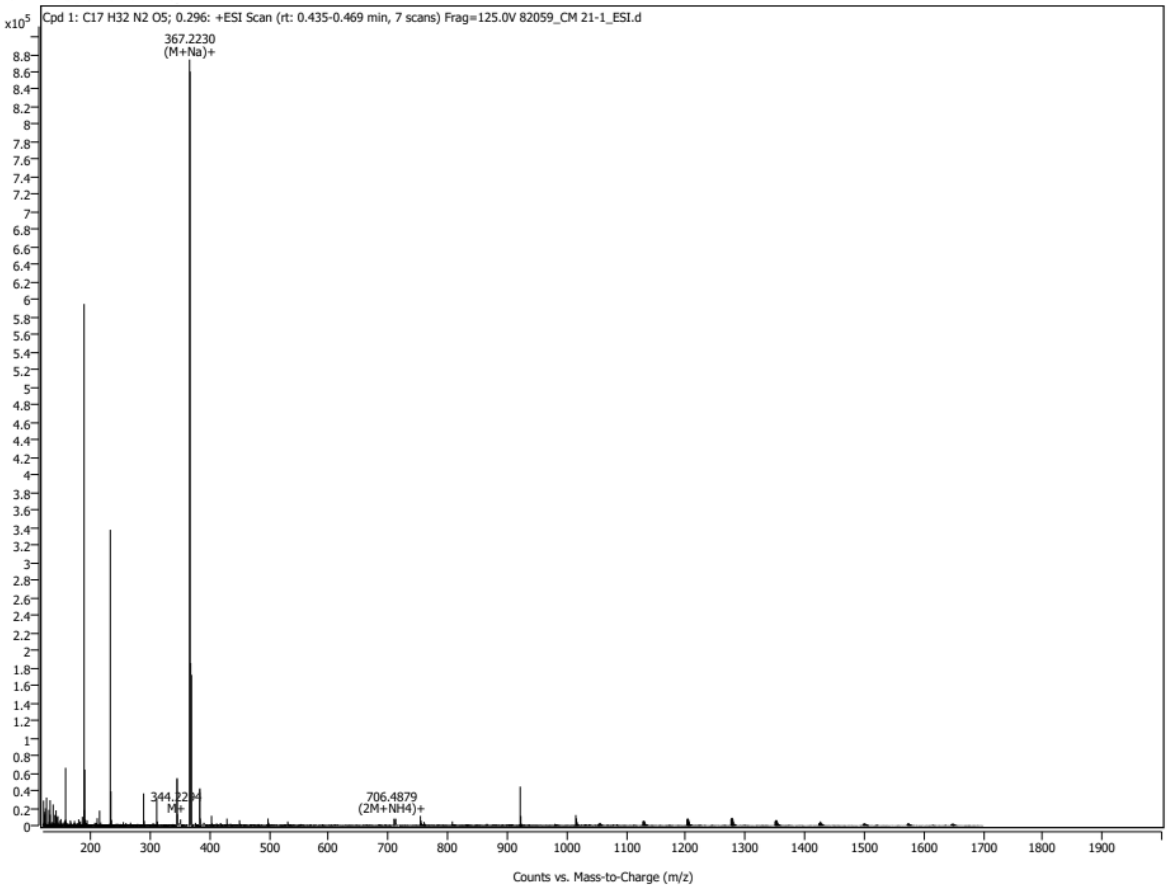

Cpd 1: C18 H34 N2 O5; 0.225 203.14004: + Scan (rt: 0.203-0.248 min, 9 scans) 82048_CM 19-1_1UL_.d
x10^5
359.2550
([C18H34N2O5]+H)+
381.2361
([C18H34N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

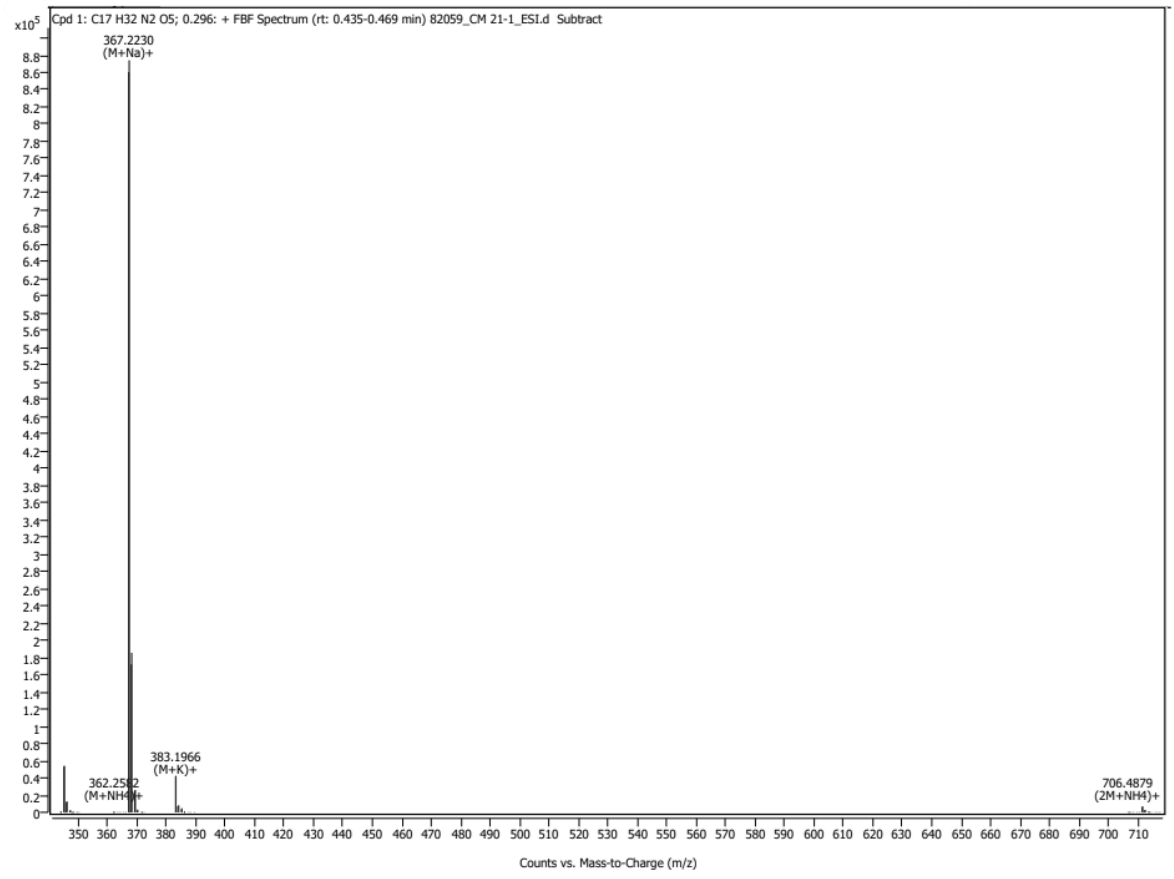

Cpd 1: C18 H34 N2 O5; 0.225 203.14004: + Scan (rt: 0.203-0.248 min, 9 scans) 82048_CM 19-1_1UL_.d
x10^6
203.1400
359.2550
([C18H34N2O5]+H)+
Counts vs. Mass-to-Charge (m/z)

Boc-VA-O^{t}Bu **C5**

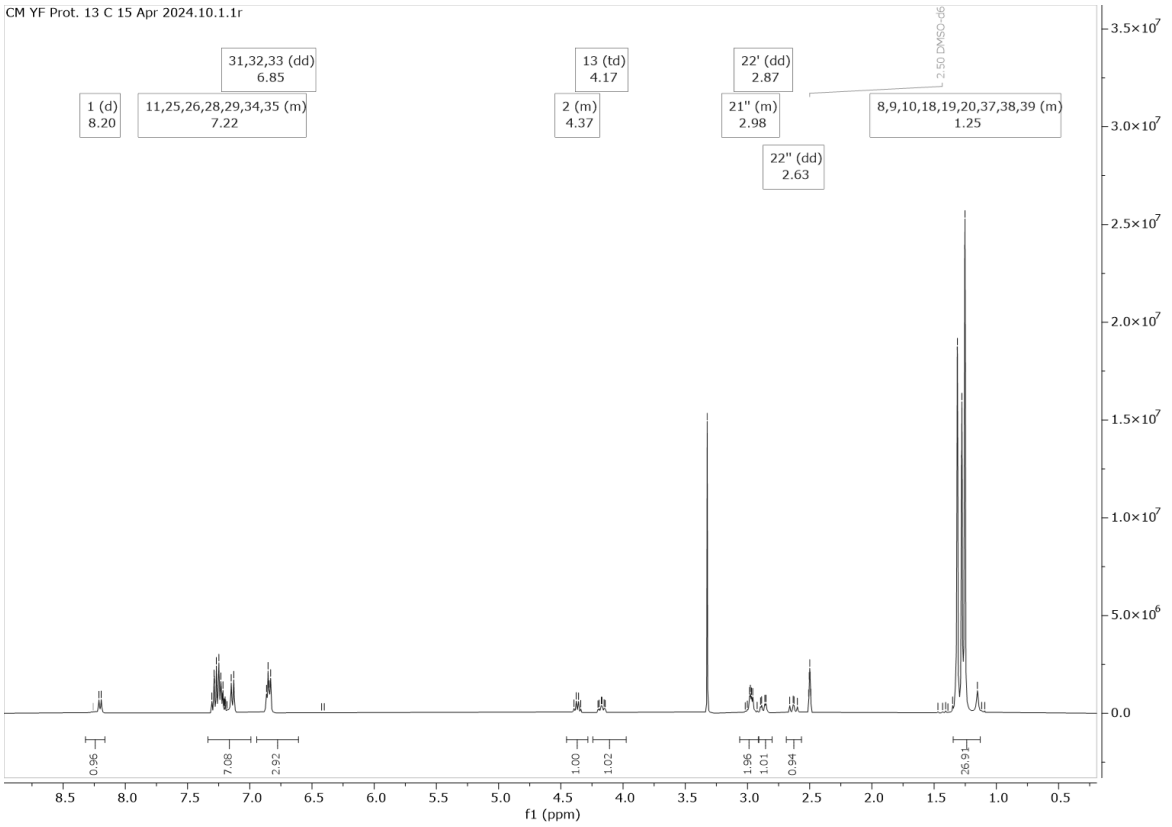


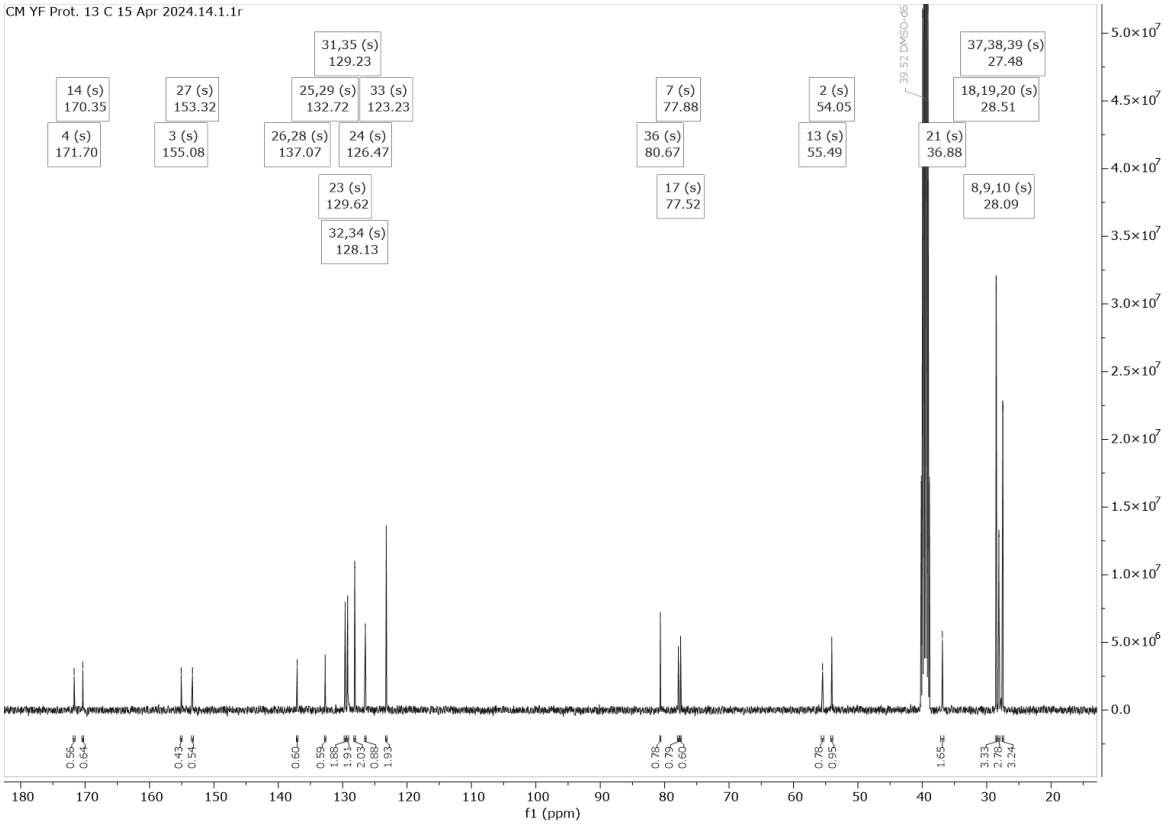


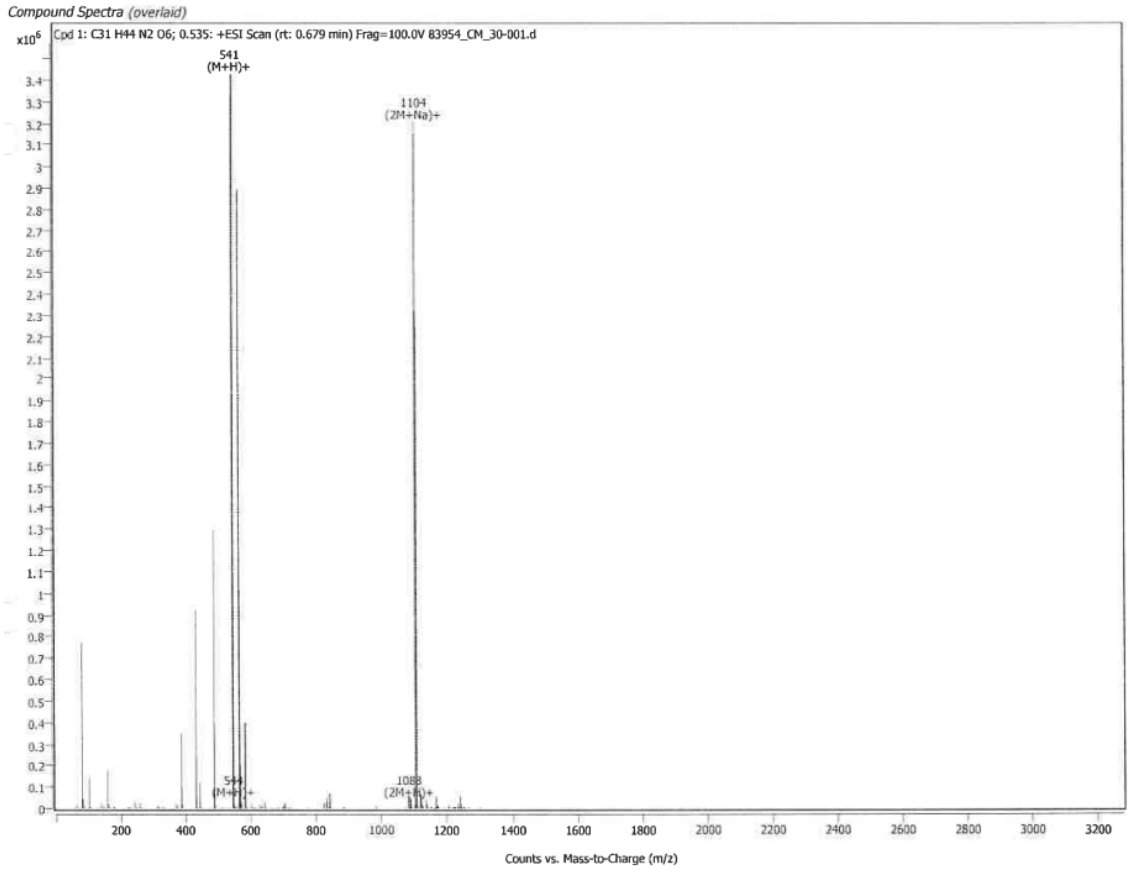

Cpd 1: C17 H32 N2 O5; 0.296: +ESI Scan (rt: 0.435-0.469 min, 7 scans) Frag=125.0V 82059_CM 21-1_ESI.d
x10^5
367.2230
(M+Na)+
344.2294
M+
706.4879
(2M+NH4)+
Counts vs. Mass-to-Charge (m/z)

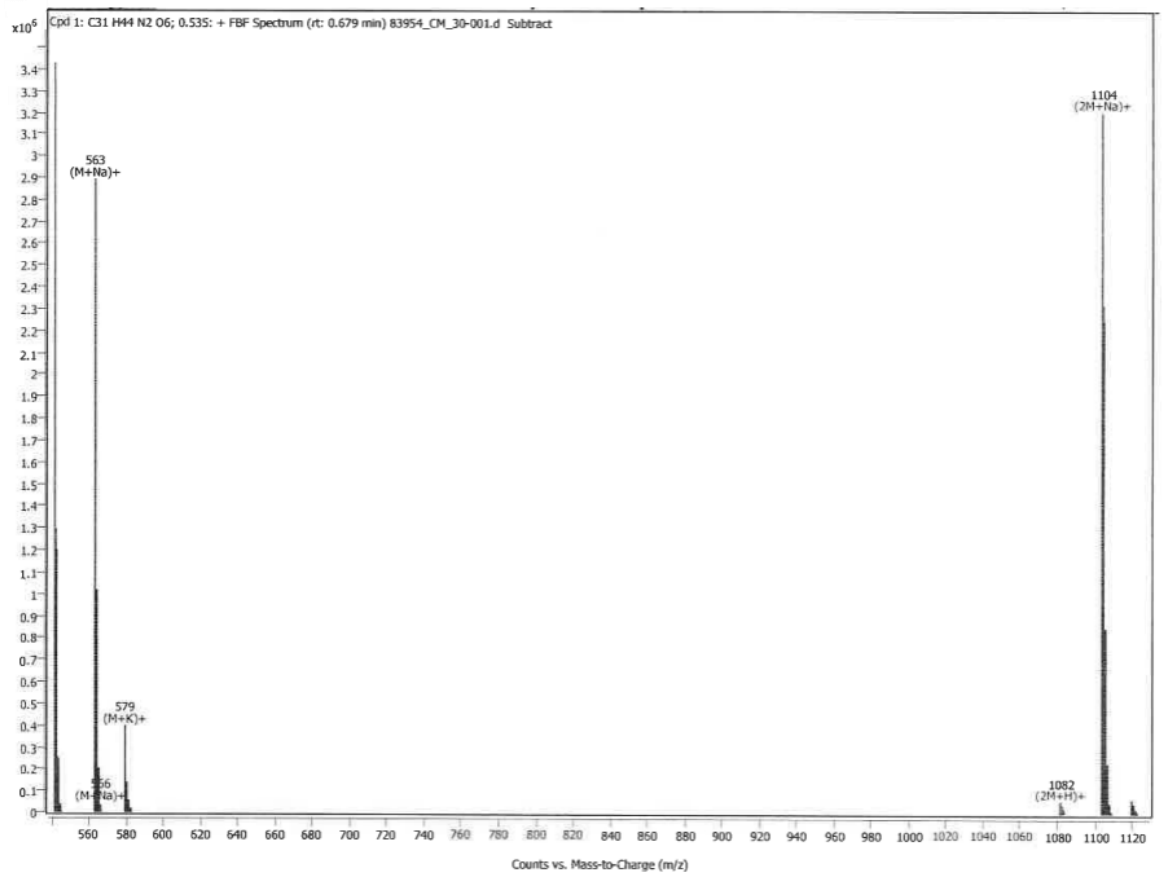

Cpd 1: C17 H32 N2 O5; 0.296: + FBF Spectrum (rt: 0.435-0.469 min) 82059_CM 21-1_ESI.d Subtract
x10^5
367.2230
(M+Na)+
383.1966
(M+K)+
362.2502
(M+NH4)+
706.4879
(2M+NH4)+
Counts vs. Mass-to-Charge (m/z)

Boc-YF-O^{t}Bu **C6**

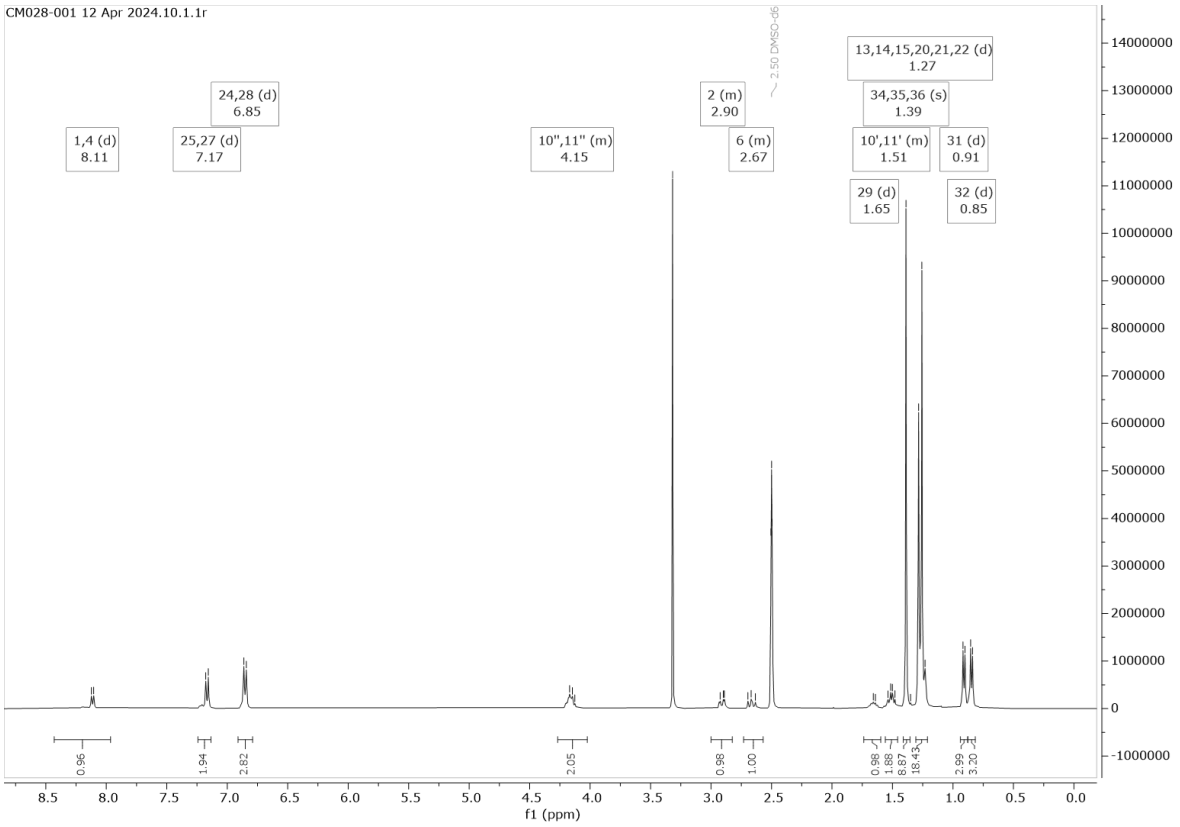


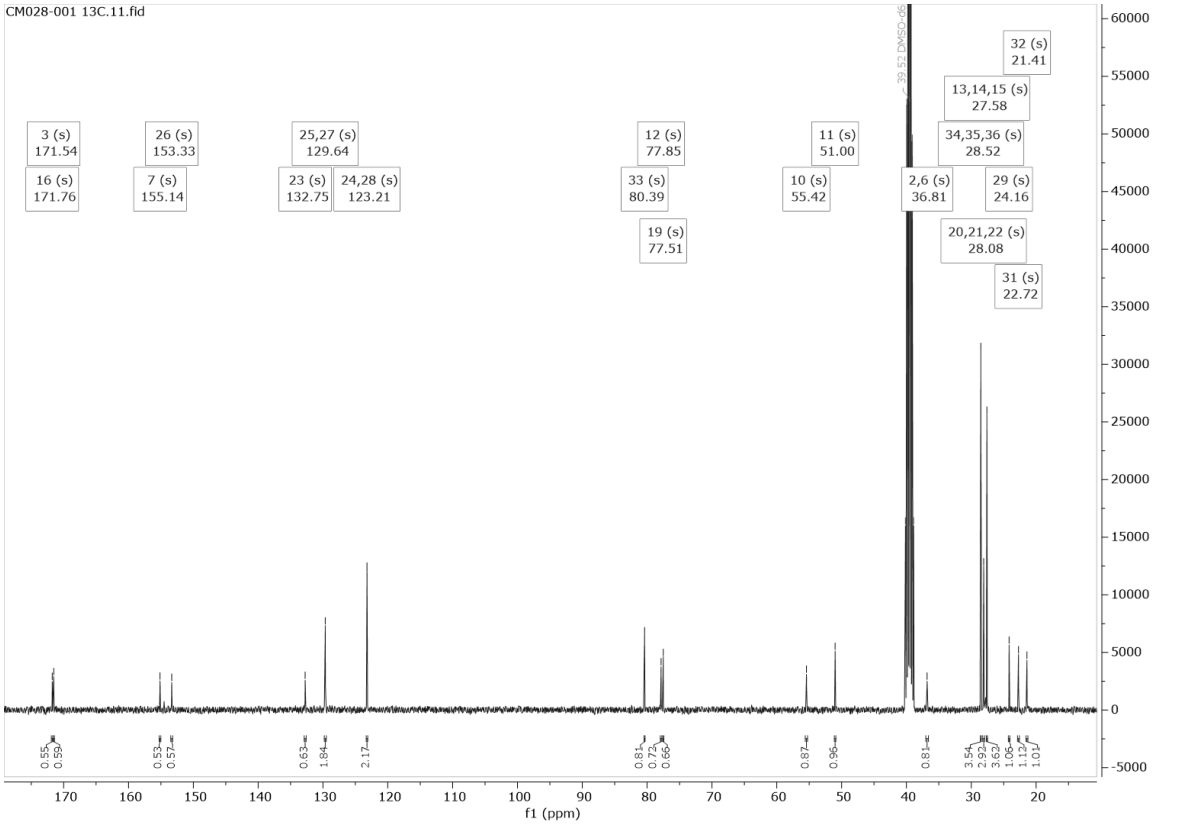


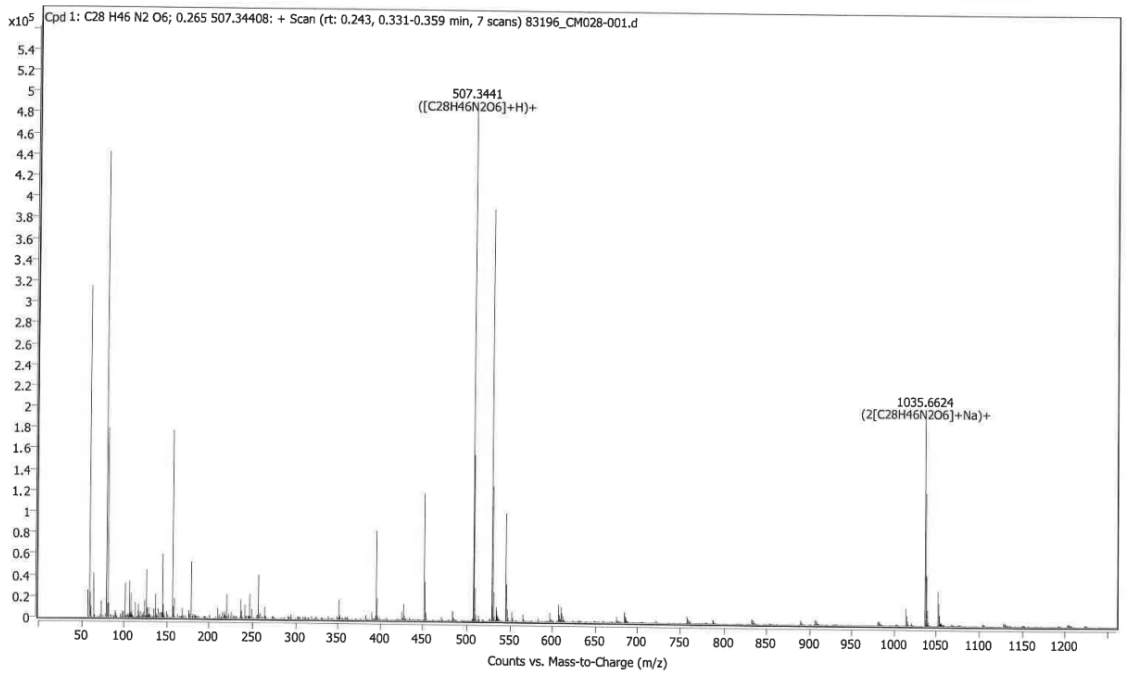

Compound Spectra (overlaid)
Cpd 1: C31 H44 N2 O6; 0.535: +ESI Scan (rt: 0.679 min) Frag=100.0V 83954_CM_30-001.d
x10^6
541
(M+H)+
1104
(2M+Na)+
Counts vs. Mass-to-Charge (m/z)

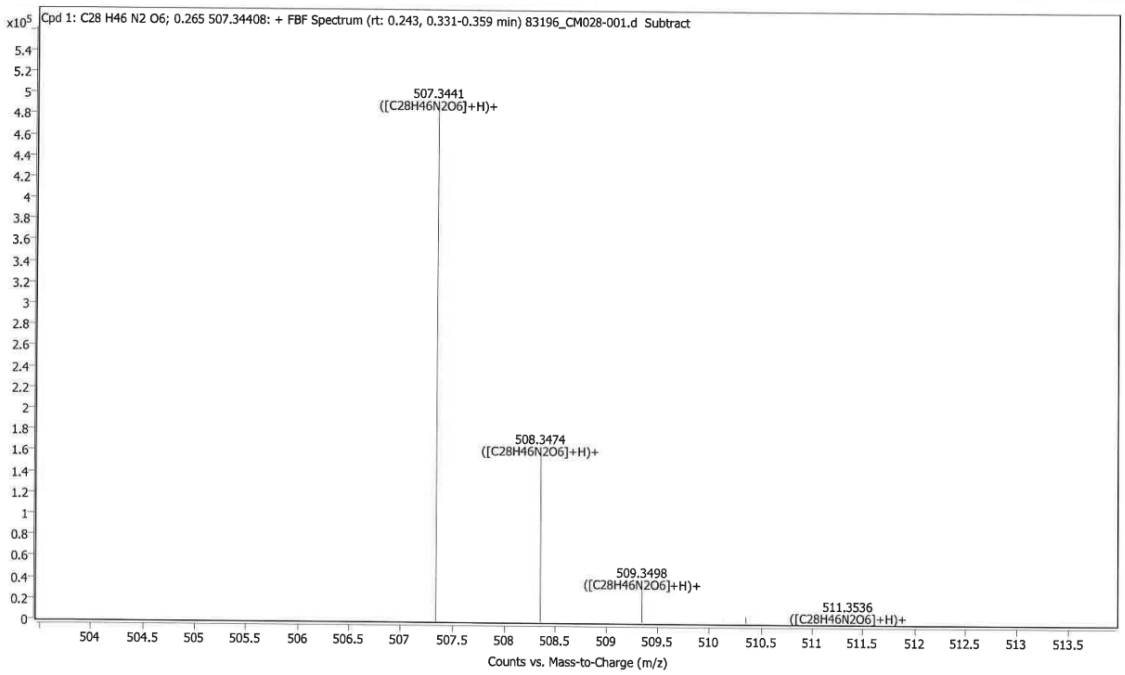

Cpd 1: C31 H44 N2 O6; 0.535: + FBF Spectrum (rt: 0.679 min) 83954_CM_30-001.d Subtract
x10^6
563
(M+Na)+
1104
(2M+Na)+
579
(M+K)+
1082
(2M+H)+
Counts vs. Mass-to-Charge (m/z)

Boc-YL-O^{t}Bu **C7**

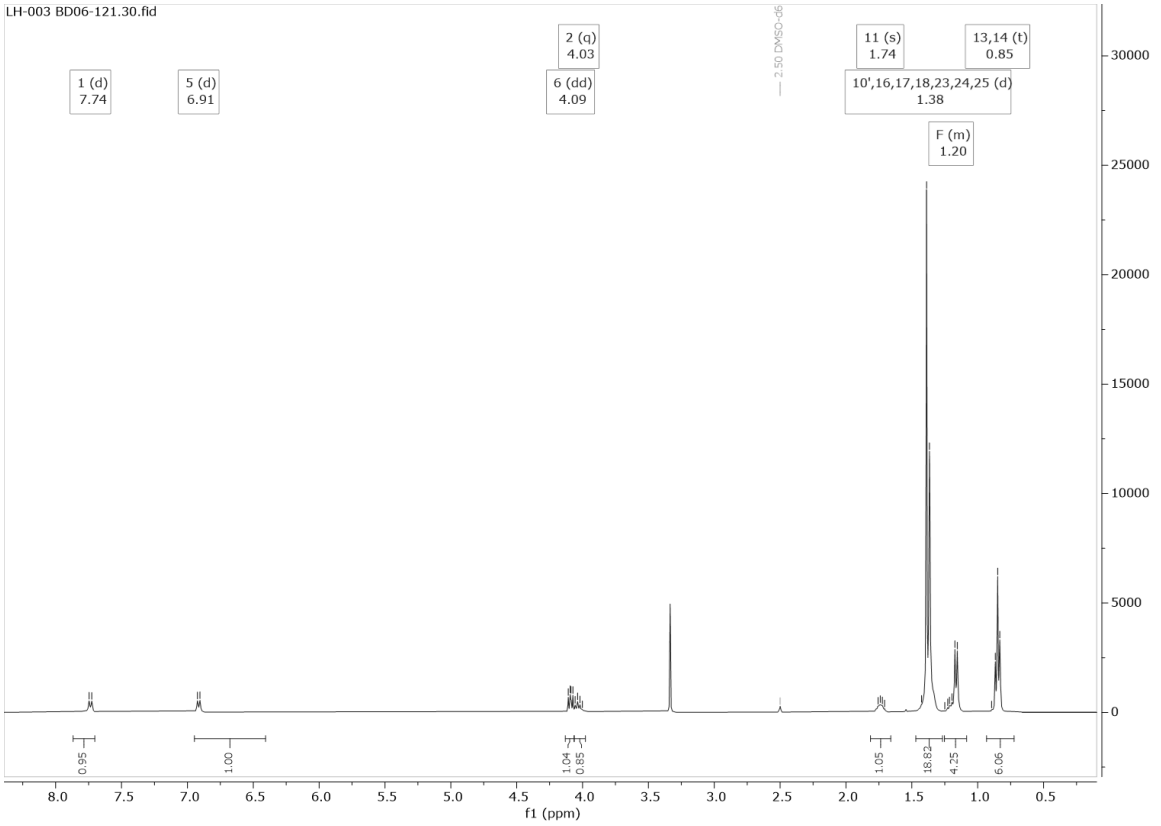


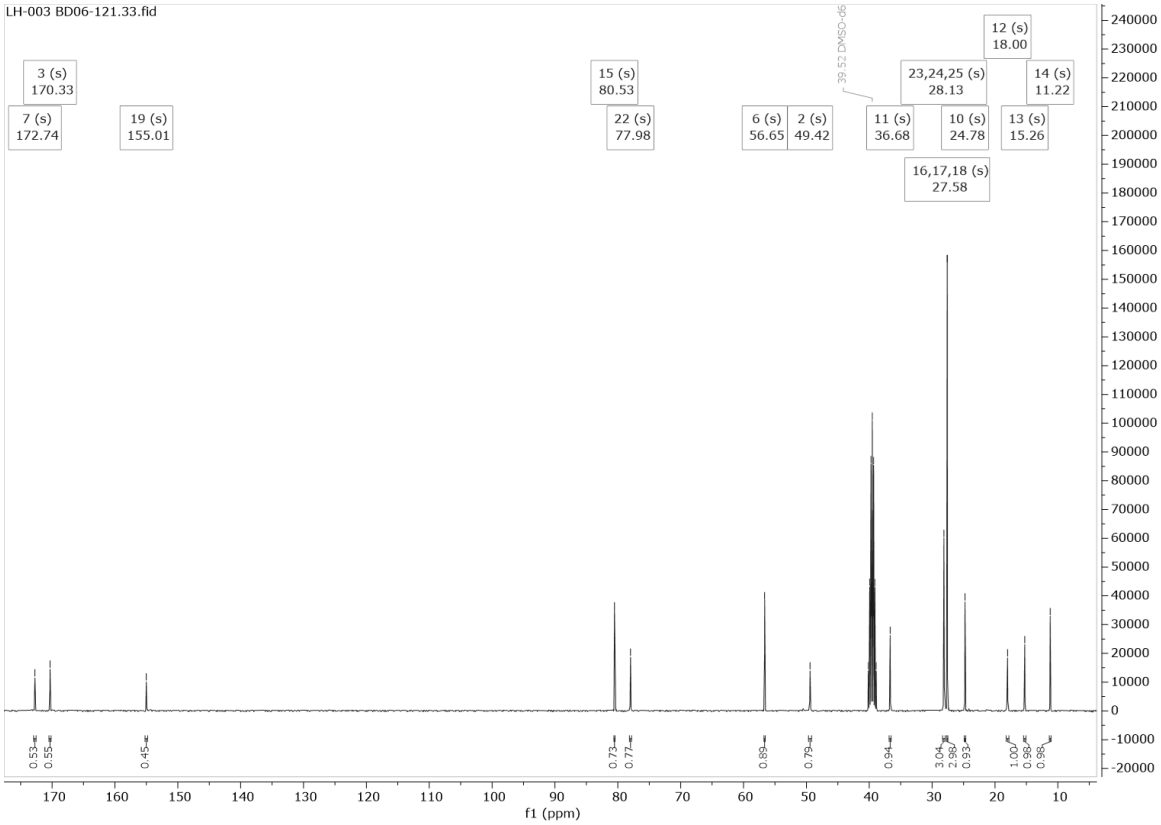


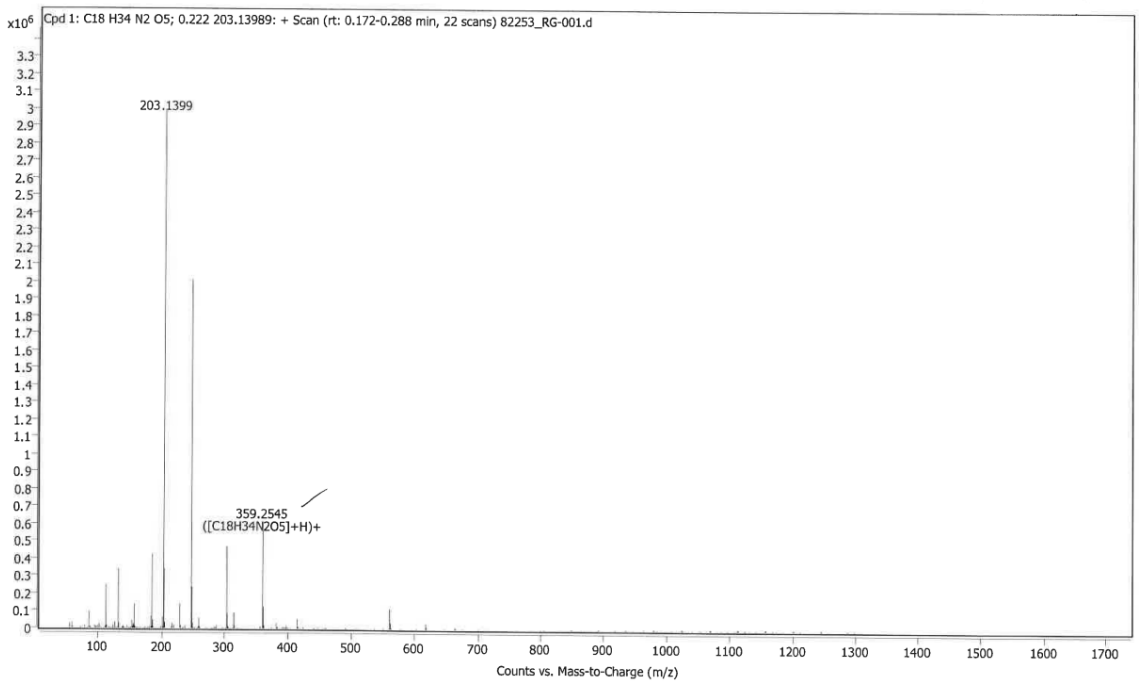

x10^5
Cpd 1: C28 H46 N2 O6; 0.265 507.34408: + Scan (rt: 0.243, 0.331-0.359 min, 7 scans) 83196_CM028-001.d
507.3441
([C28H46N2O6]+H)+
1035.6624
(2[C28H46N2O6]+Na)+
Counts vs. Mass-to-Charge (m/z)

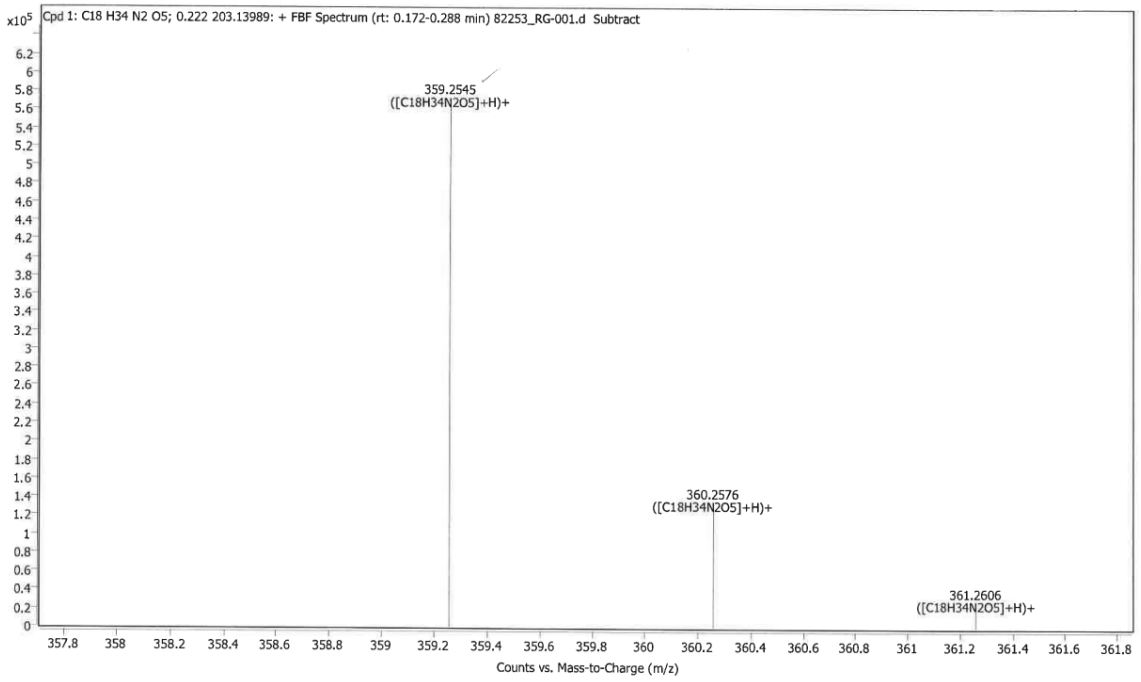

x10^5
Cpd 1: C28 H46 N2 O6; 0.265 507.34408: + FBF Spectrum (rt: 0.243, 0.331-0.359 min) 83196_CM028-001.d Subtract
507.3441
([C28H46N2O6]+H)+
508.3474
([C28H46N2O6]+H)+
509.3498
([C28H46N2O6]+H)+
511.3536
([C28H46N2O6]+H)+
Counts vs. Mass-to-Charge (m/z)

Boc-AL-O^{t}Bu **C8**

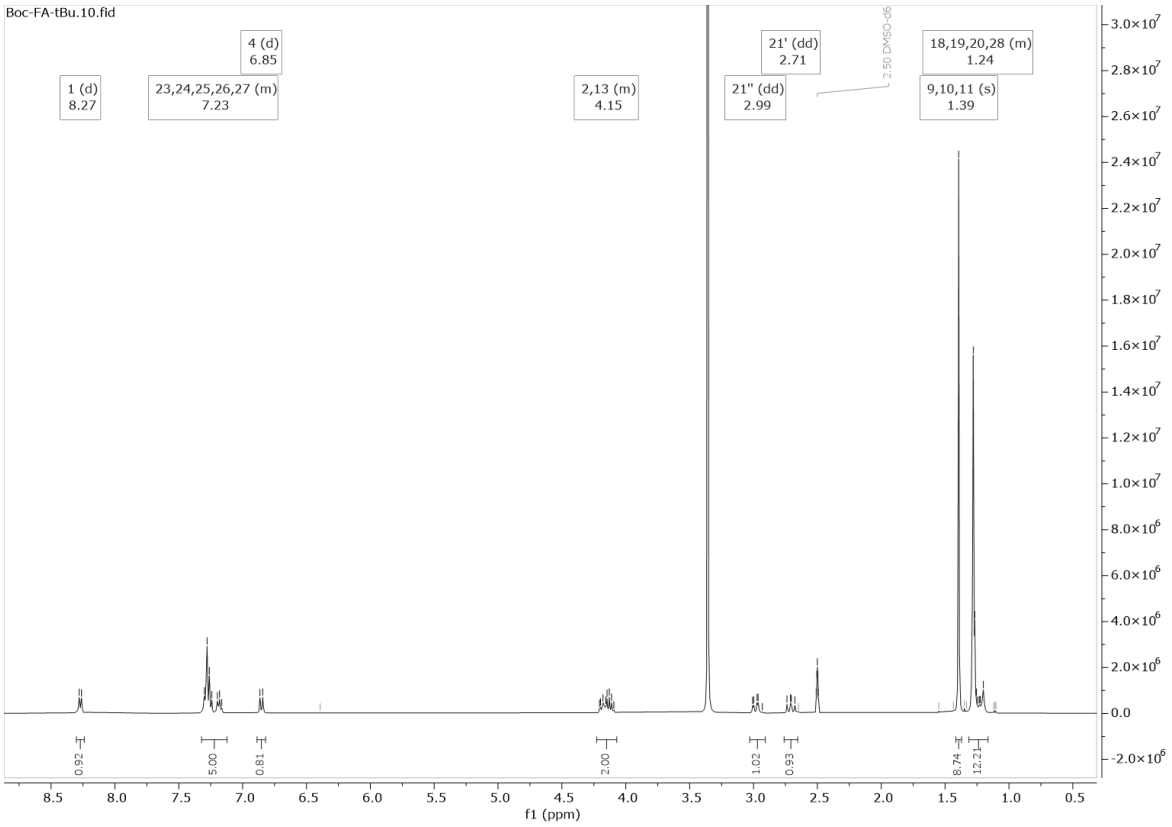


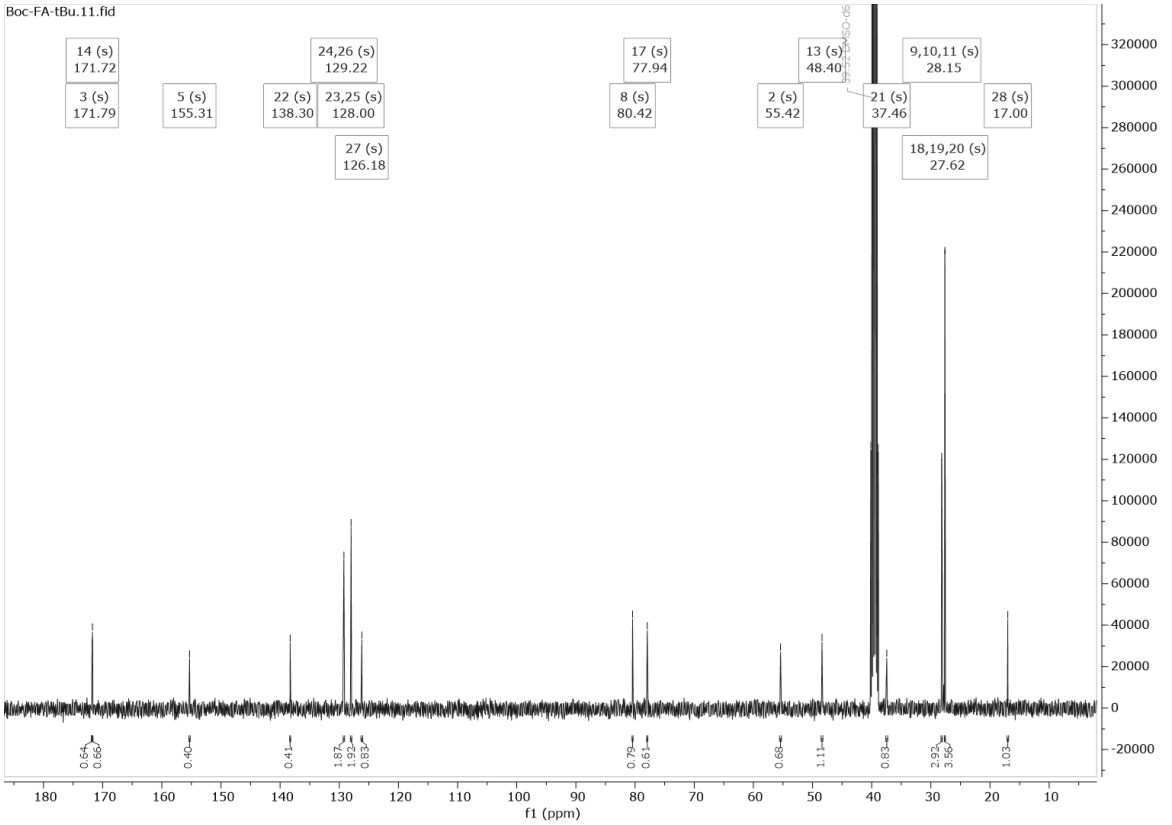


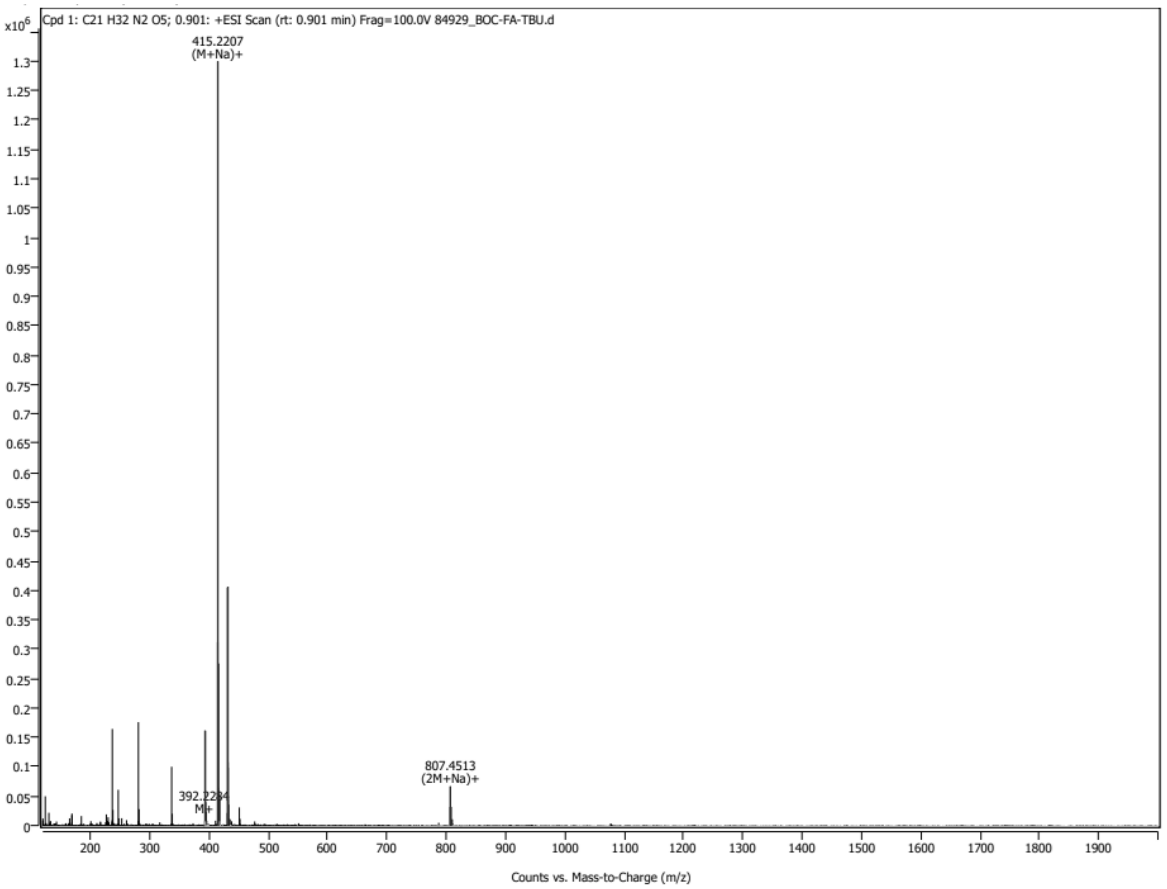

x10^6
Cpd 1: C18 H34 N2 O5; 0.222 203.13989: + Scan (rt: 0.172-0.288 min, 22 scans) 82253_RG-001.d
203.1399
359.2545
([C18H34N2O5]+H)+
Counts vs. Mass-to-Charge (m/z)

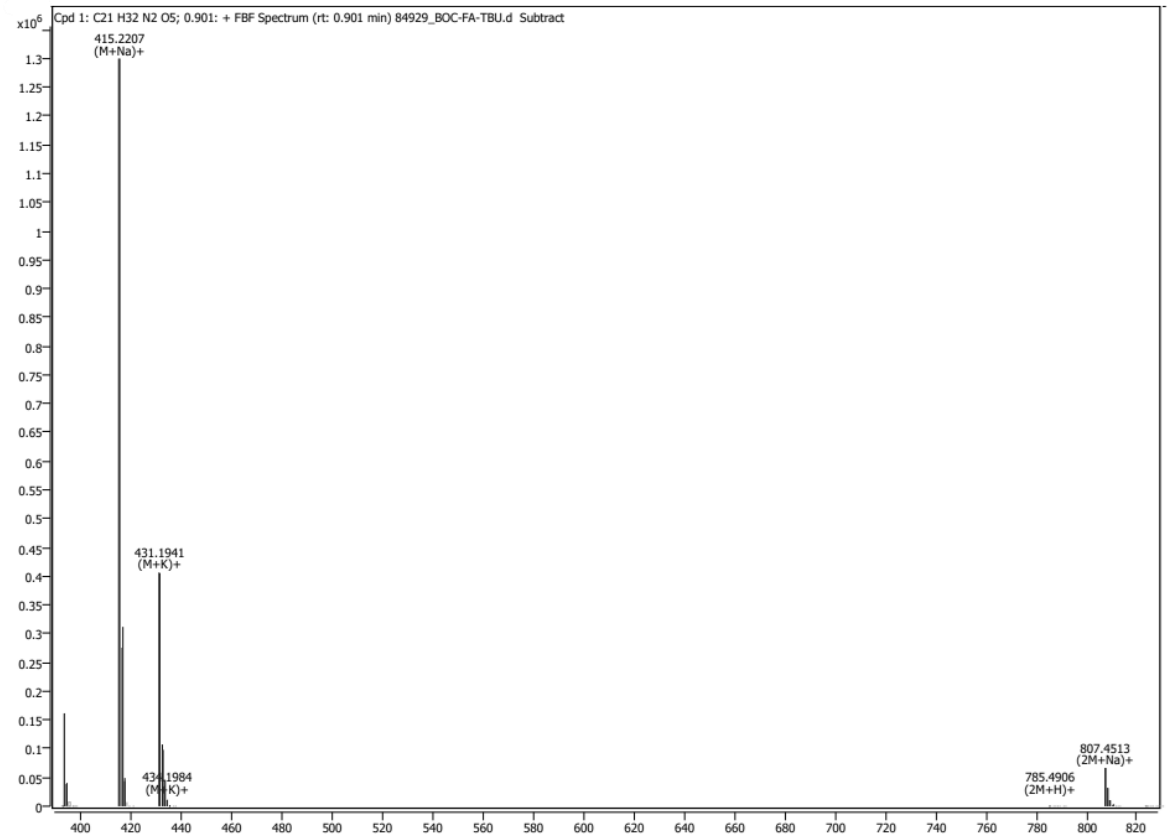

x10^5
Cpd 1: C18 H34 N2 O5; 0.222 203.13989: + FBF Spectrum (rt: 0.172-0.288 min) 82253_RG-001.d Subtract
359.2545
([C18H34N2O5]+H)+
360.2576
([C18H34N2O5]+H)+
361.2606
([C18H34N2O5]+H)+
Counts vs. Mass-to-Charge (m/z)

Boc-FA-O^{t}Bu **C9**

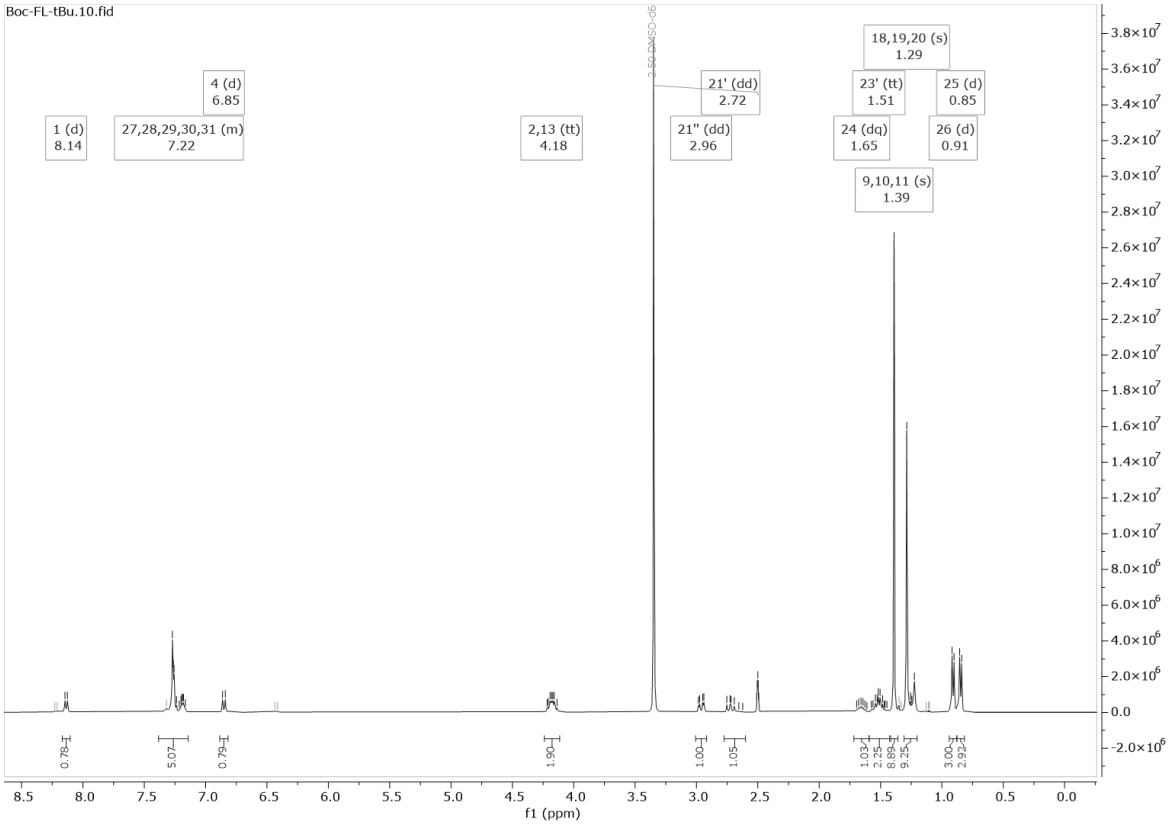


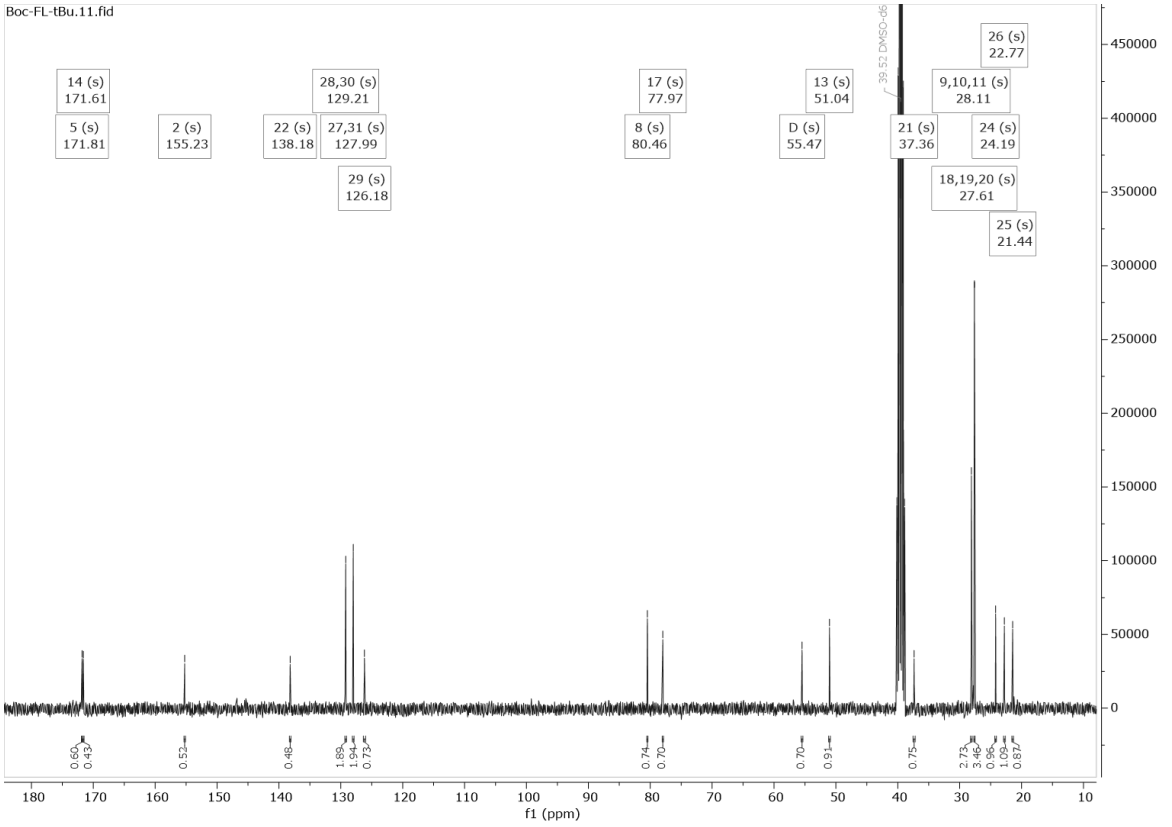


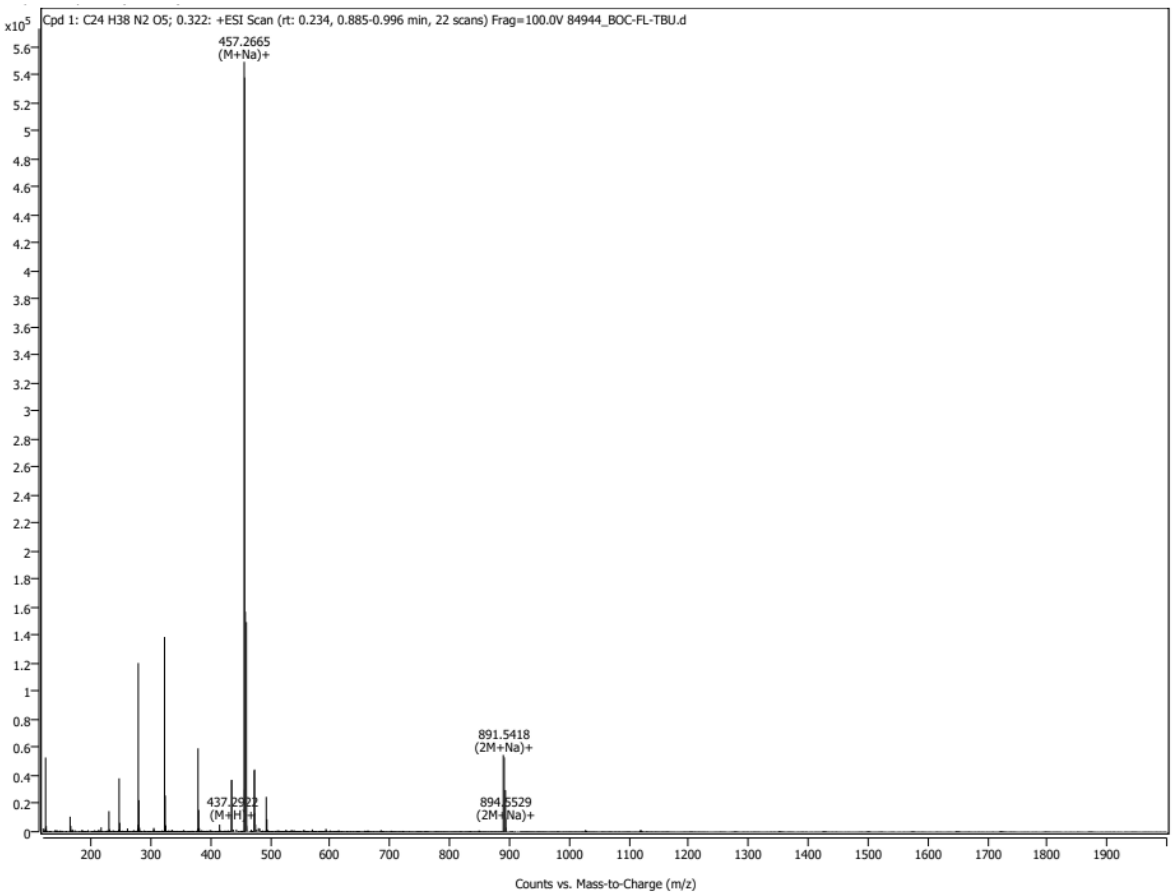

Cpd 1: C21 H32 N2 O5; 0.901: +ESI Scan (rt: 0.901 min) Frag=100.0V 84929_BOC-FA-TBU.d
x10^6
415.2207
(M+Na)+
807.4513
(2M+Na)+
392.2284
M+
Counts vs. Mass-to-Charge (m/z)

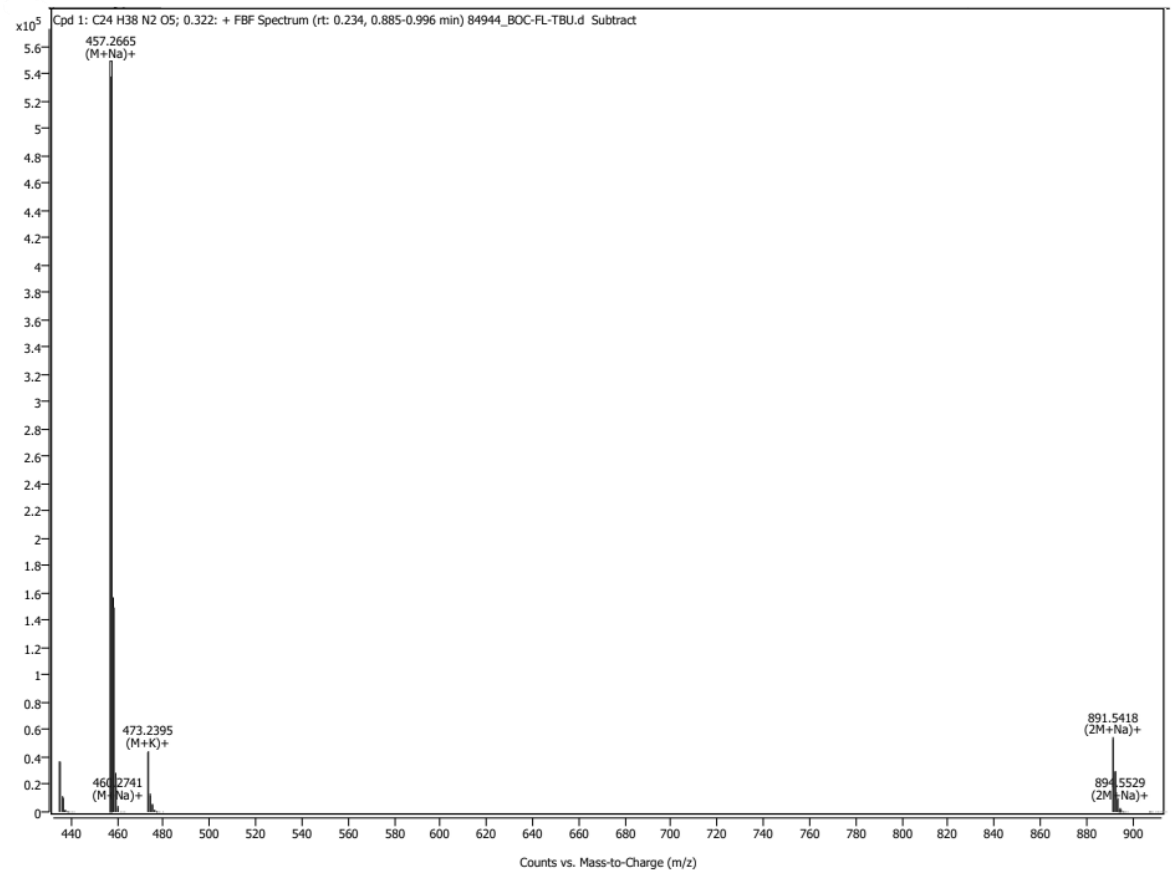

Cpd 1: C21 H32 N2 O5; 0.901: + FBF Spectrum (rt: 0.901 min) 84929_BOC-FA-TBU.d Subtract
x10^6
415.2207
(M+Na)+
431.1941
(M+K)+
807.4513
(2M+Na)+
785.4906
(2M+H)+

Boc-FL-O^{t}Bu **C10**

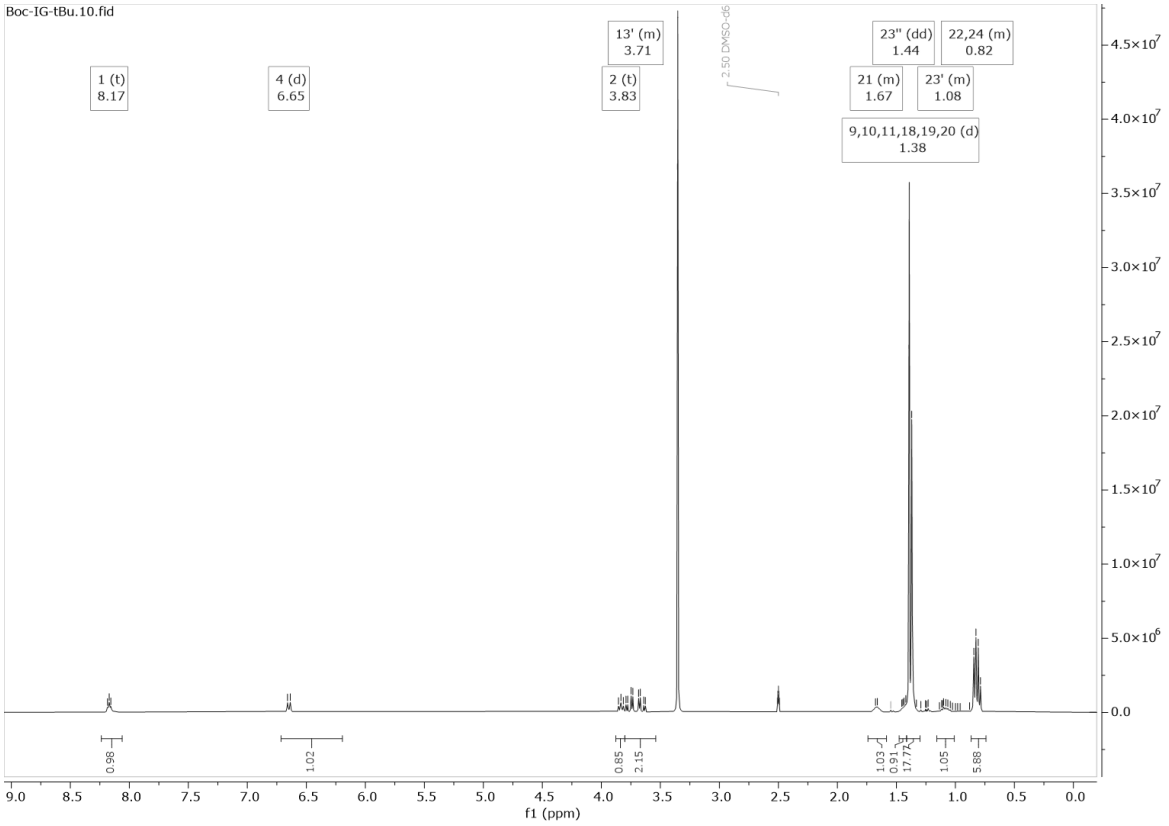


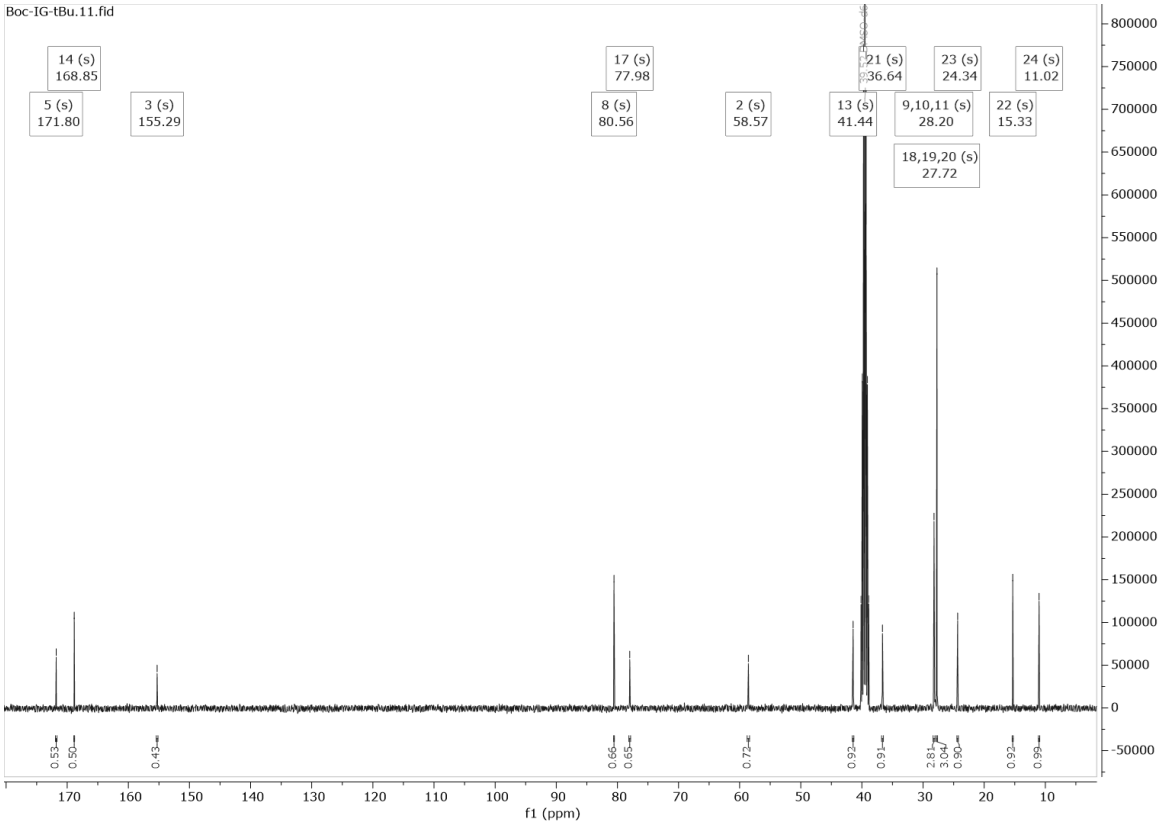


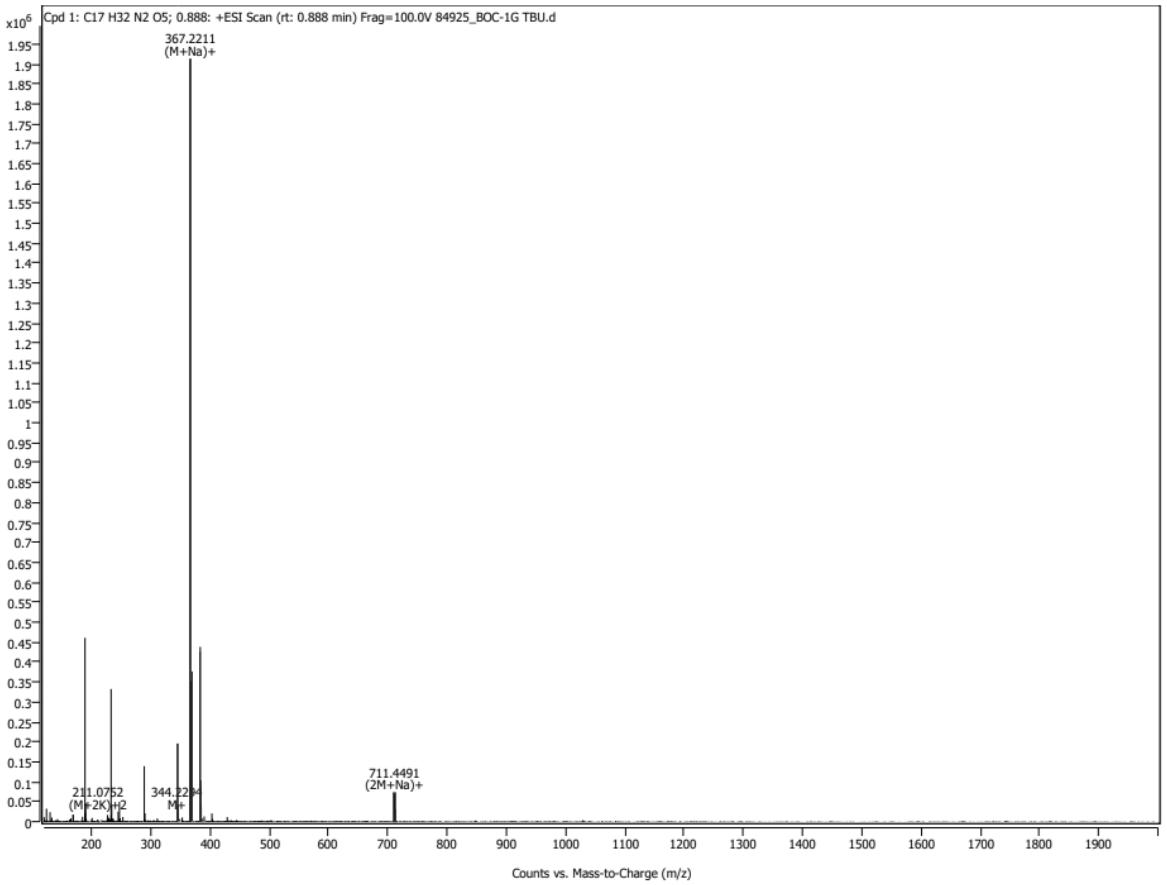

Cpd 1: C24 H38 N2 O5; 0.322: +ESI Scan (rt: 0.234, 0.885-0.996 min, 22 scans) Frag=100.0V 84944_BOC-FL-TBU.d
x10^5
457.2665
(M+Na)+
891.5418
(2M+Na)+
437.2922
(M+H)+
894.5529
(2M+Na)+
Counts vs. Mass-to-Charge (m/z)

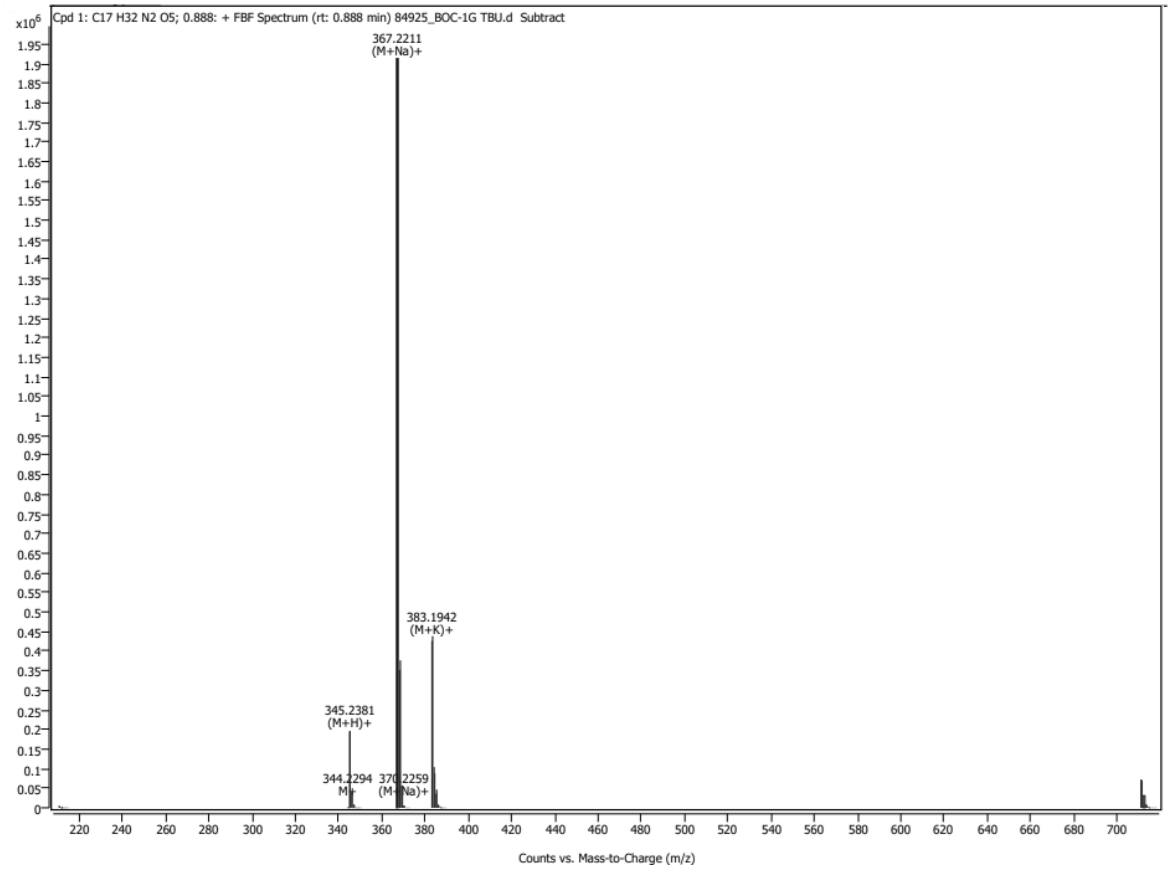

Cpd 1: C24 H38 N2 O5; 0.322: + FBF Spectrum (rt: 0.234, 0.885-0.996 min) 84944_BOC-FL-TBU.d Subtract
x10^5
457.2665
(M+Na)+
473.2395
(M+K)+
460.2741
(M+Na)+
891.5418
(2M+Na)+
894.5529
(2M+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-IG-O^{t}Bu **C11**

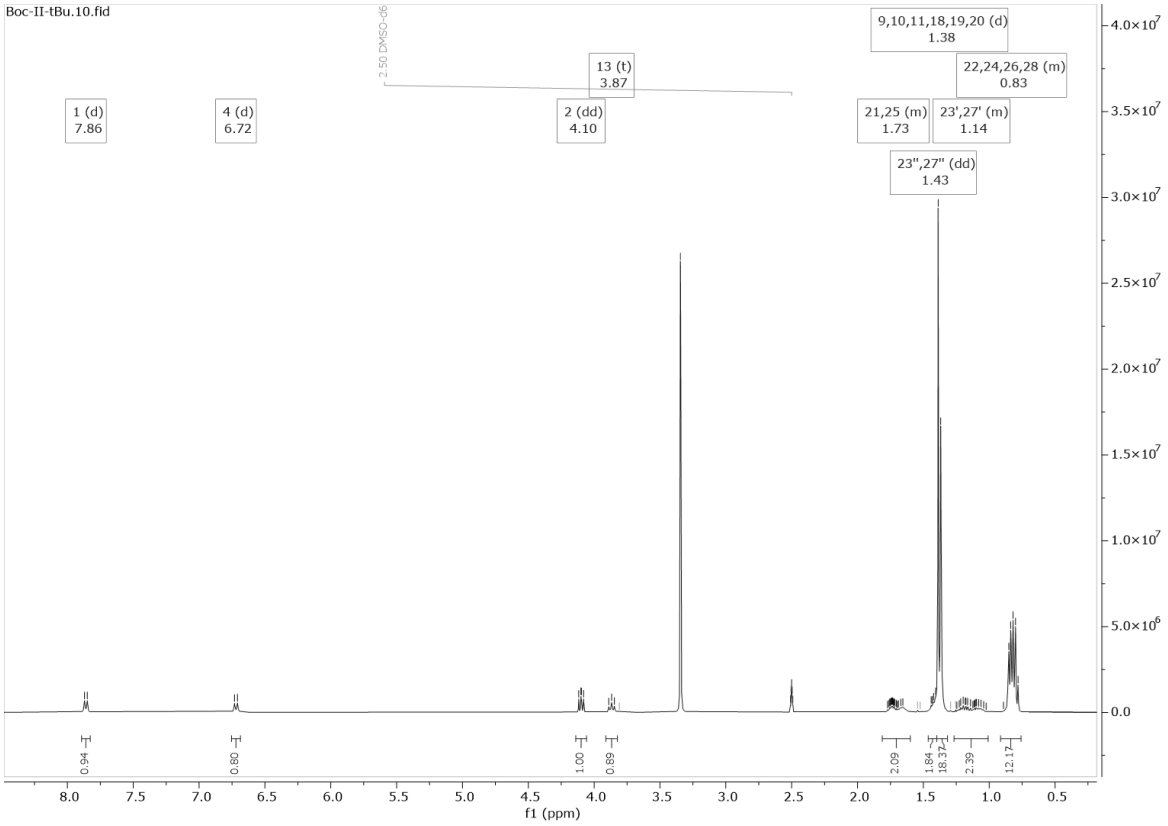


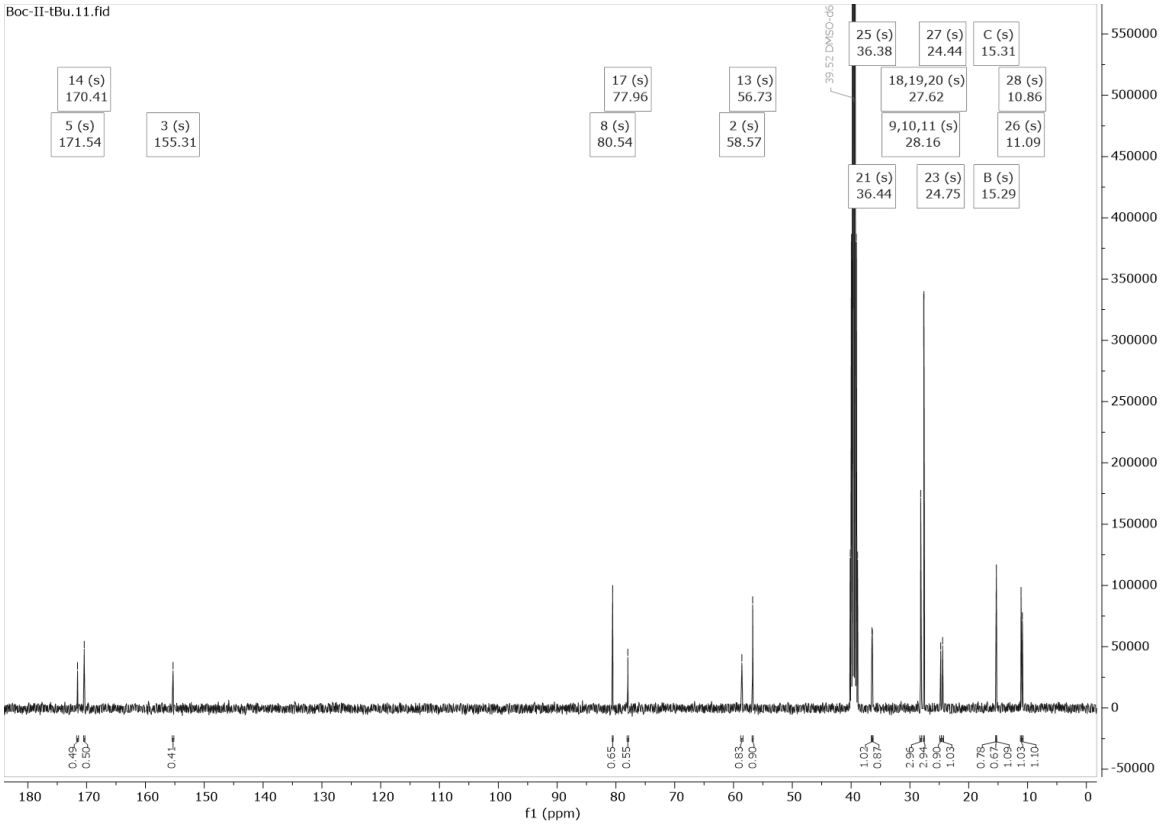


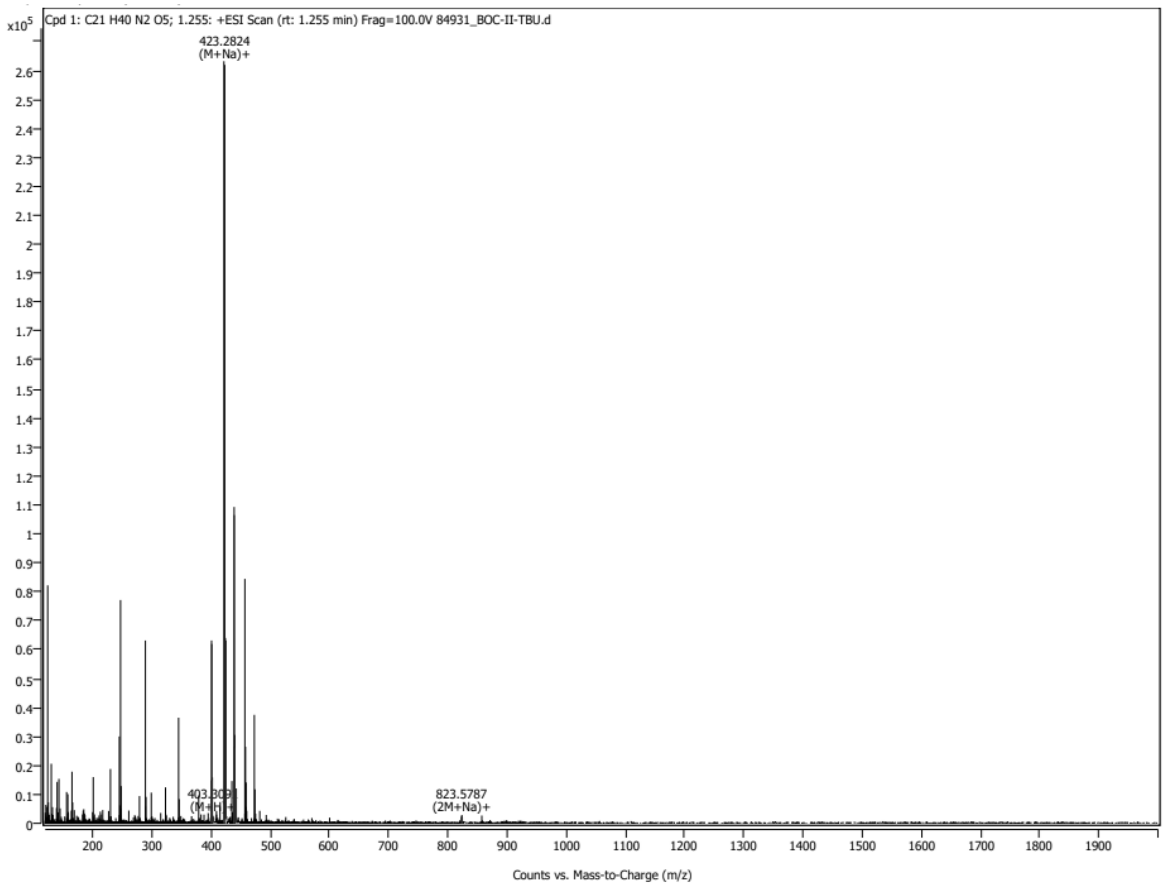

Cpd 1: C17 H32 N2 O5; 0.888: +ESI Scan (rt: 0.888 min) Frag=100.0V 84925_BOC-1G TBU.d
x10^6
367.2211
(M+Na)+
711.4491
(2M+Na)+
211.0752
(M+2K)+2
344.22
M+
Counts vs. Mass-to-Charge (m/z)

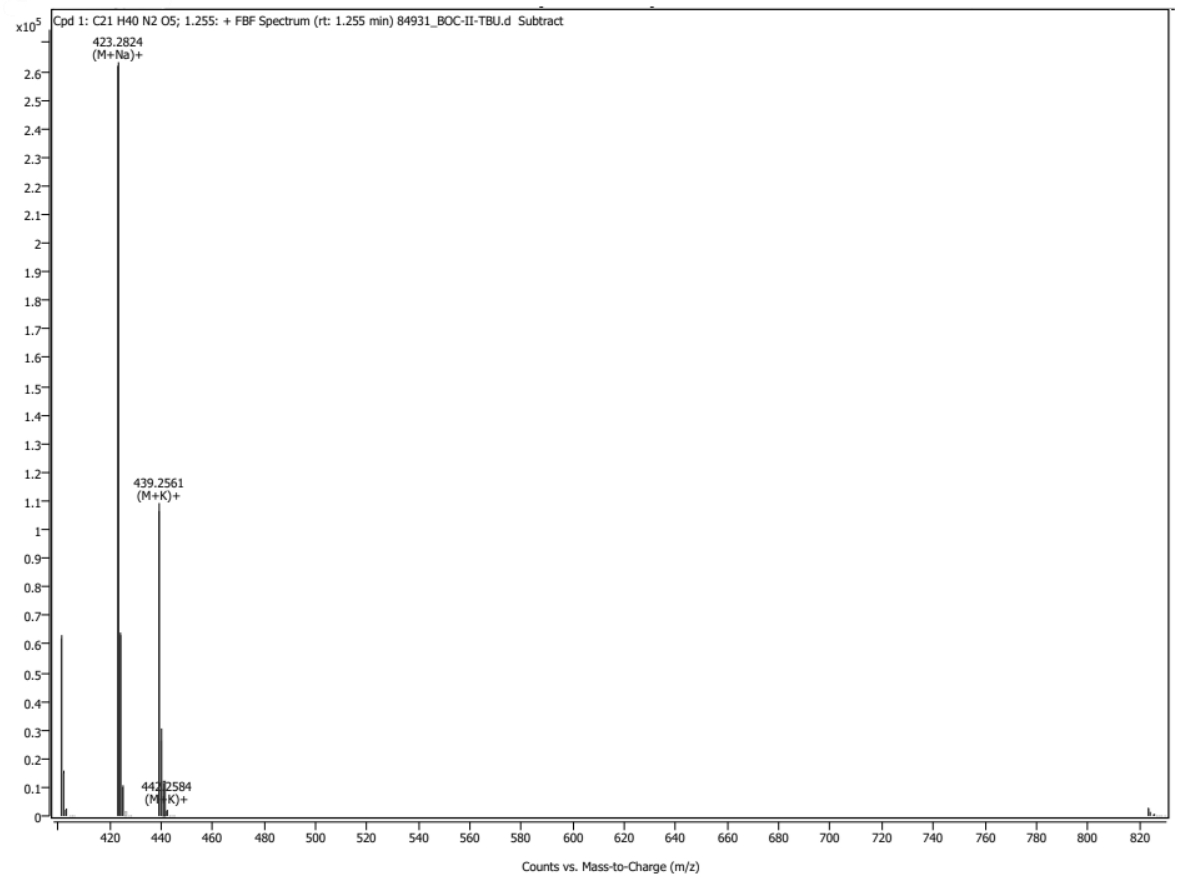

Cpd 1: C17 H32 N2 O5; 0.888: + FBF Spectrum (rt: 0.888 min) 84925_BOC-1G TBU.d Subtract
x10^6
367.2211
(M+Na)+
383.1942
(M+K)+
345.2381
(M+H)+
344.2294
M+
370.2259
(M+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-II-O^{t}Bu **C12**

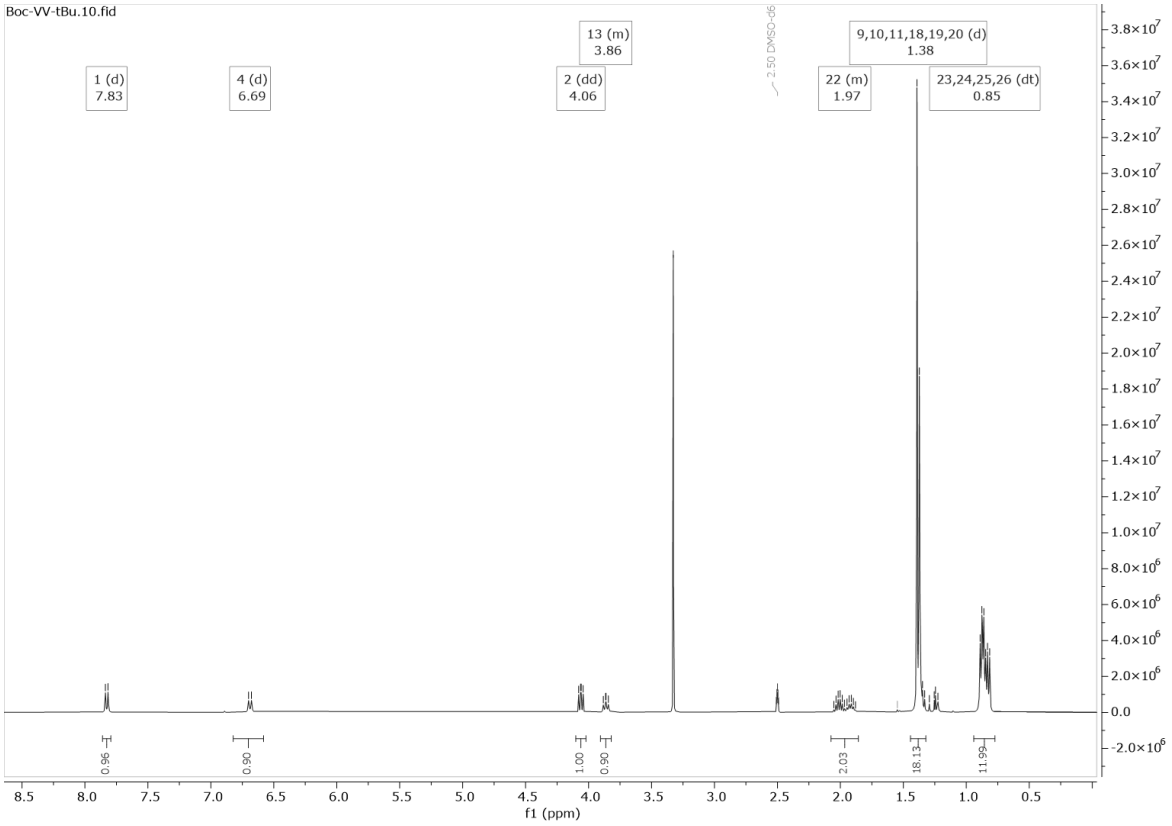


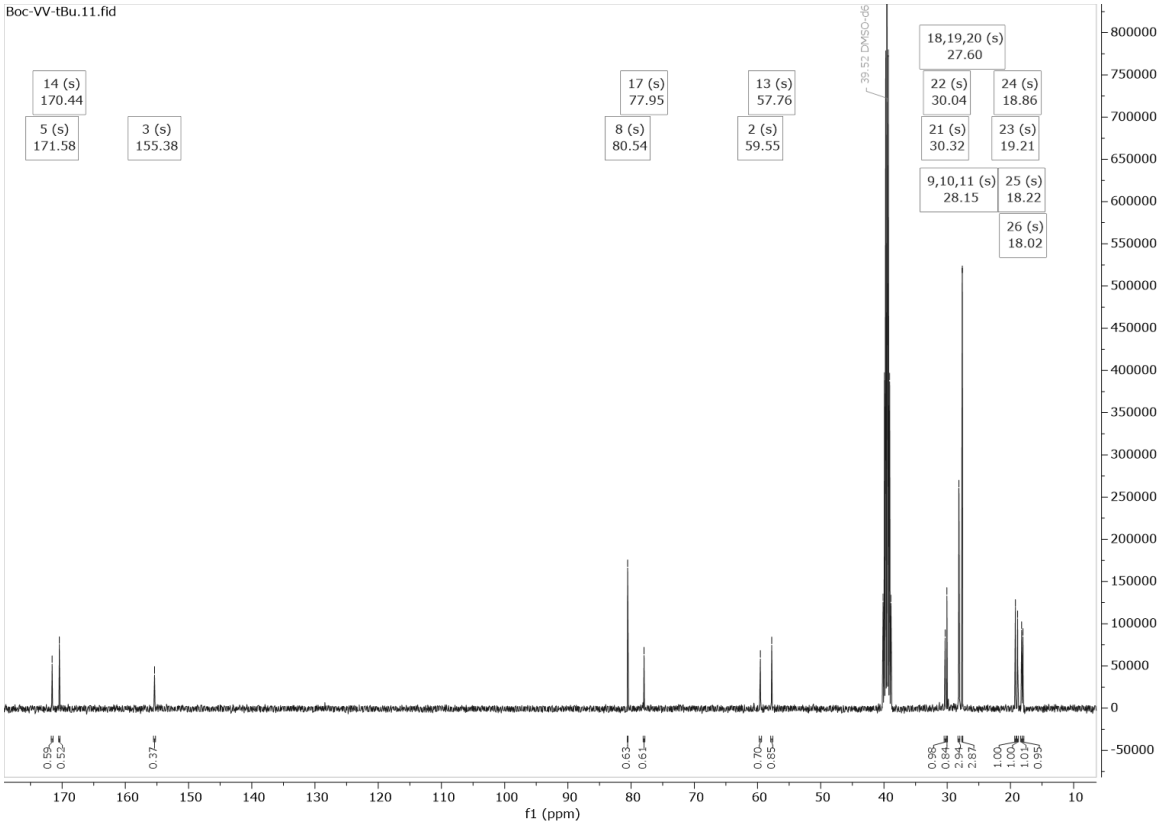


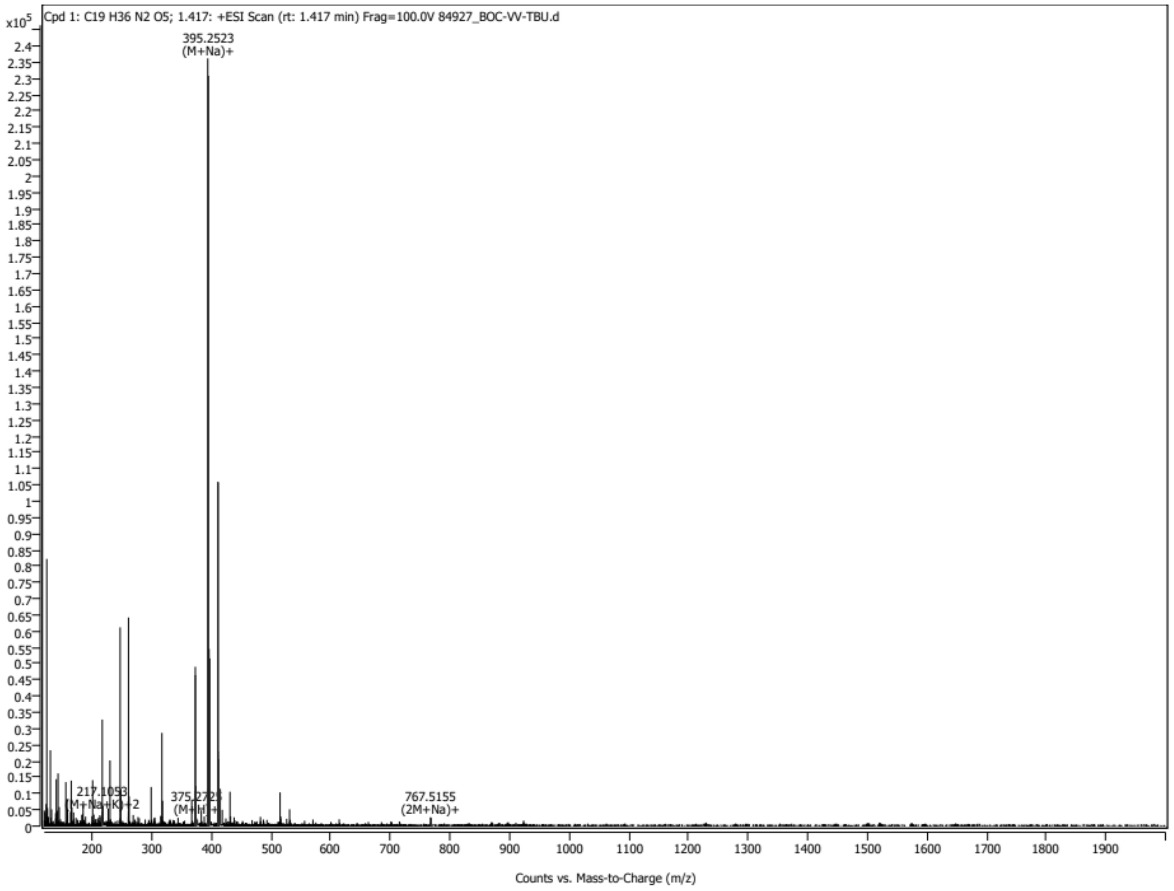

Cpd 1: C21 H40 N2 O5; 1.255: +ESI Scan (rt: 1.255 min) Frag=100.0V 84931_BOC-II-TBU.d
x10^5
423.2824
(M+Na)+
823.5787
(2M+Na)+
Counts vs. Mass-to-Charge (m/z)

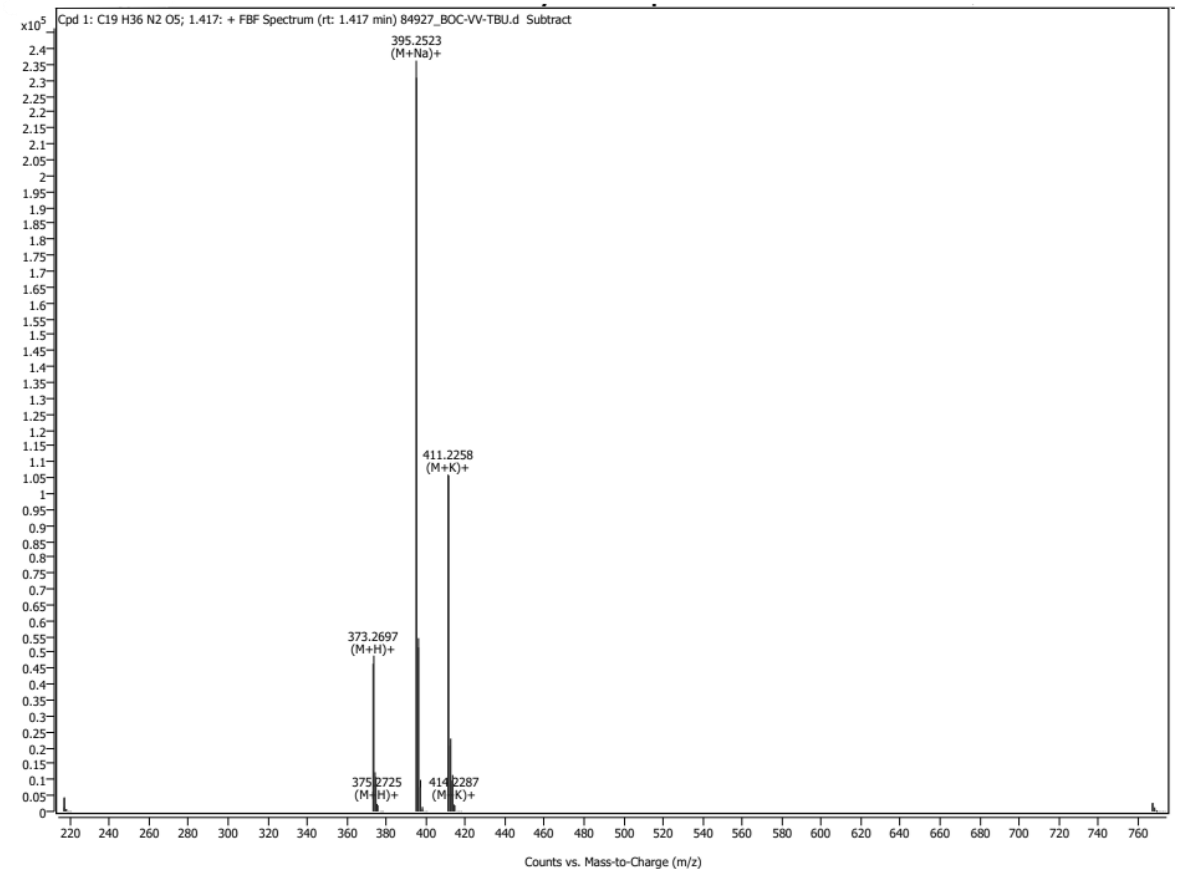

Cpd 1: C21 H40 N2 O5; 1.255: + FBF Spectrum (rt: 1.255 min) 84931_BOC-II-TBU.d Subtract
x10^5
423.2824
(M+Na)+
439.2561
(M+K)+
Counts vs. Mass-to-Charge (m/z)

Boc-VV-O^{t}Bu **C13**

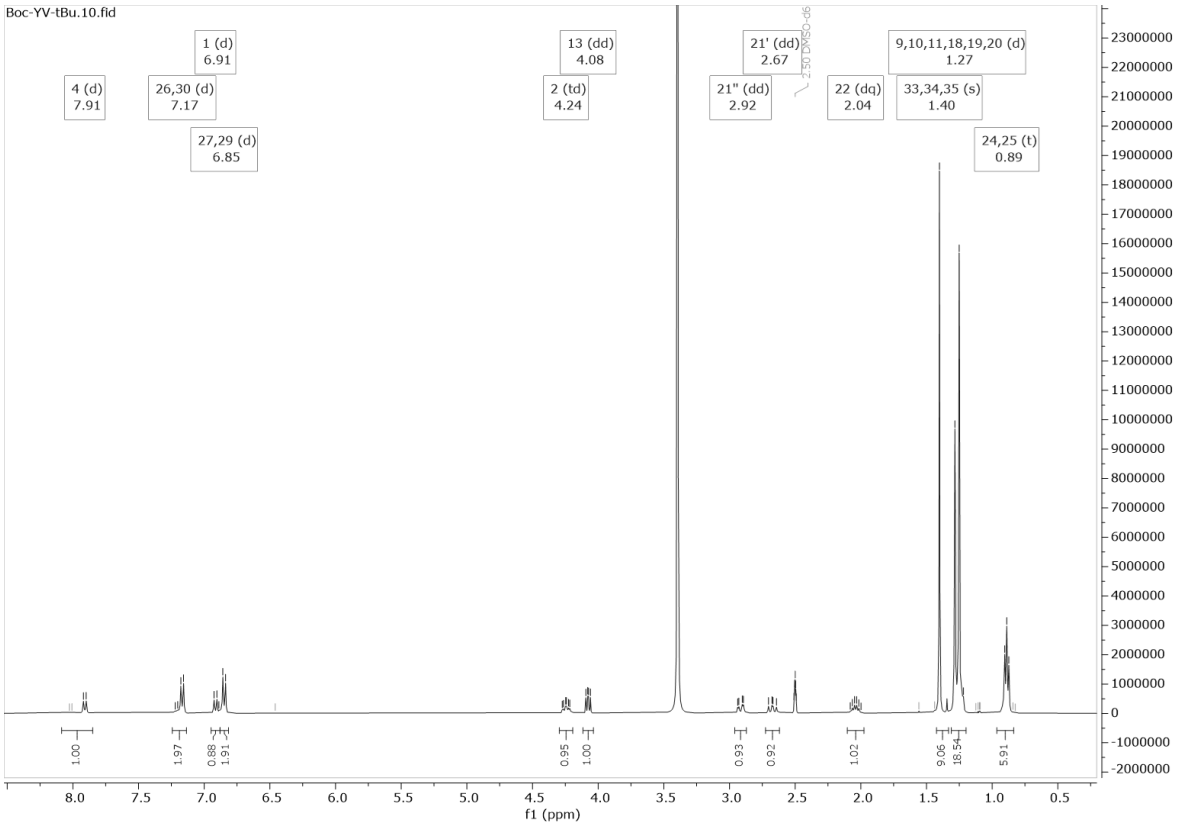


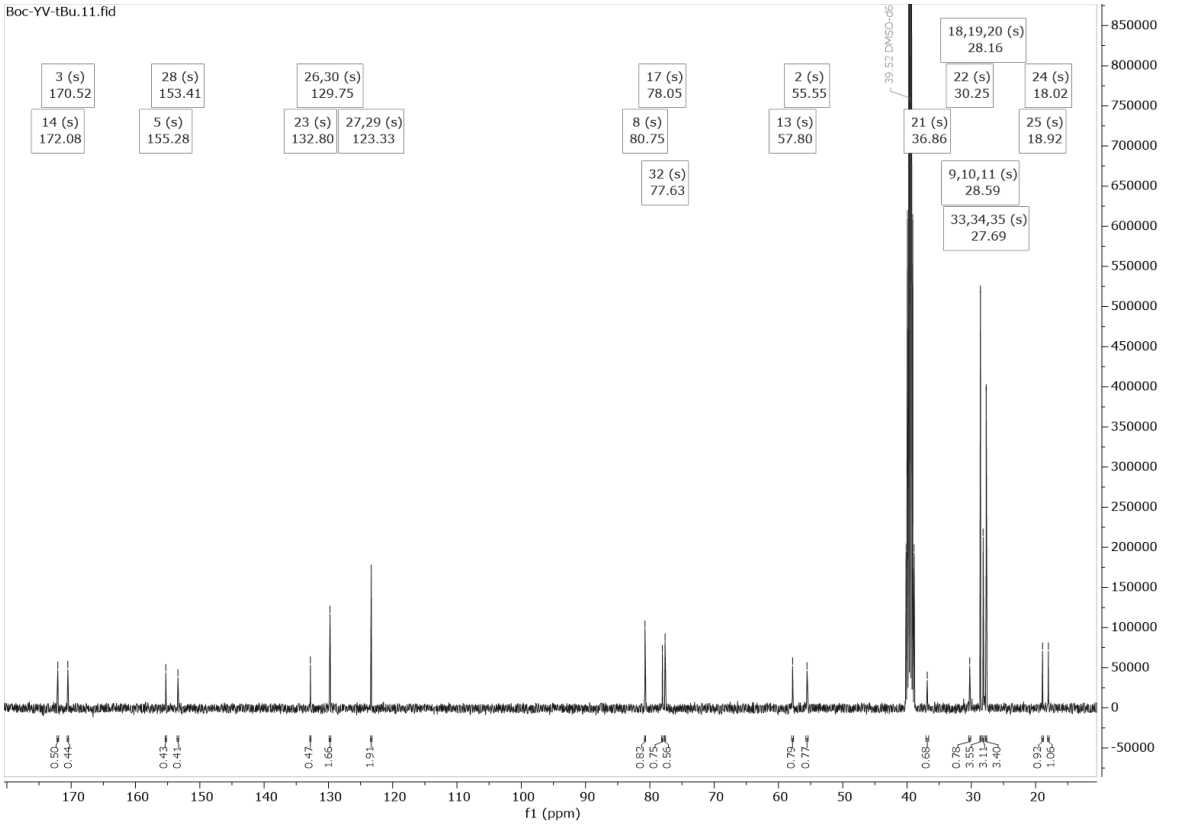


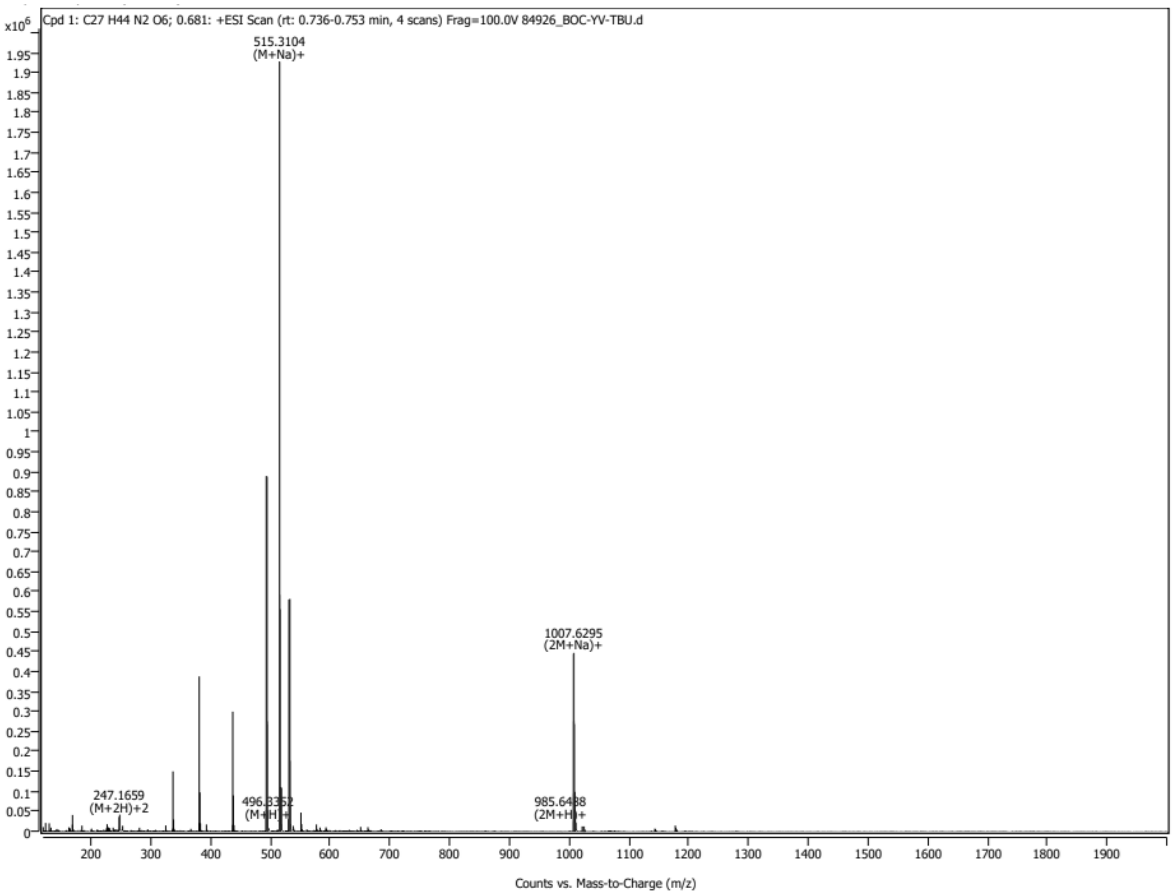

Cpd 1: C19 H36 N2 O5; 1.417: +ESI Scan (rt: 1.417 min) Frag=100.0V 84927_BOC-VV-TBU.d
x10^5
395.2523
(M+Na)+
217.1053
(M+Na+K)+2
375.2725
(M+H)+
767.5155
(2M+Na)+
Counts vs. Mass-to-Charge (m/z)

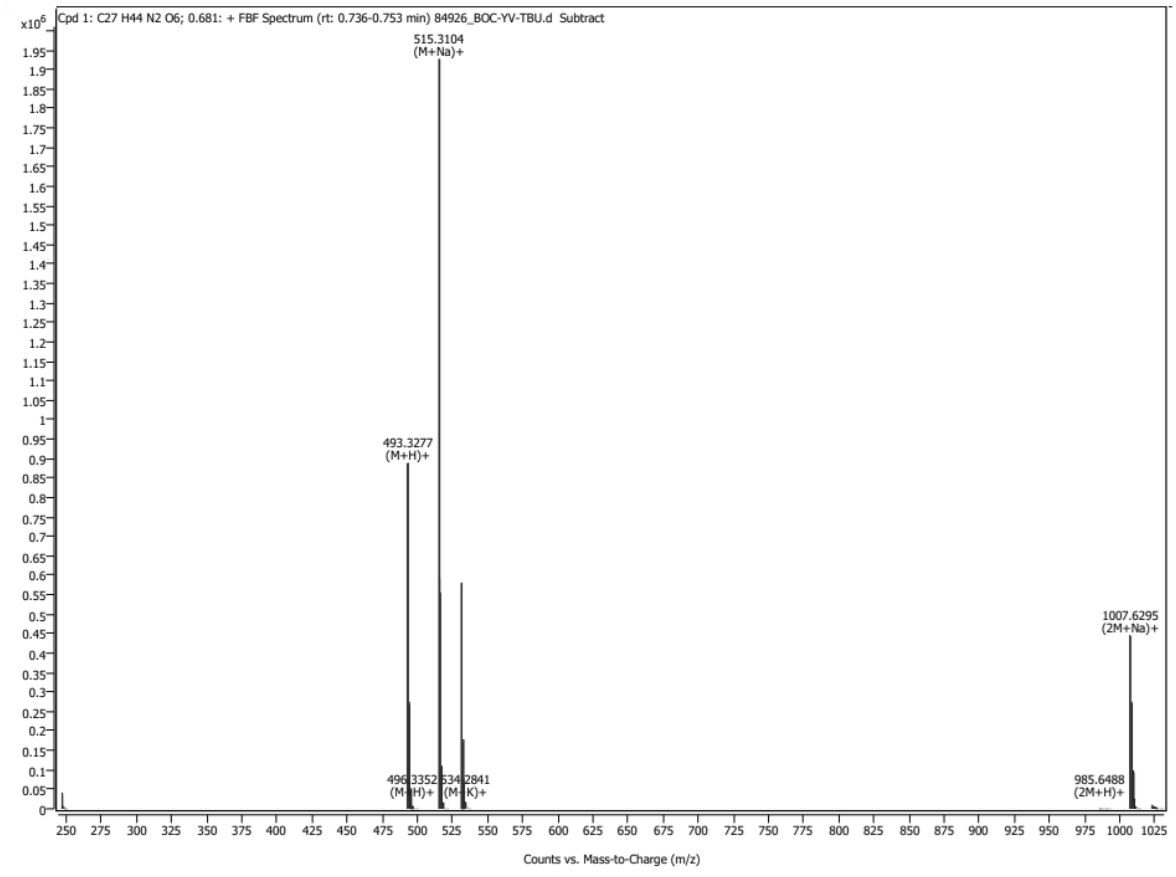

Cpd 1: C19 H36 N2 O5; 1.417: + FBF Spectrum (rt: 1.417 min) 84927_BOC-VV-TBU.d Subtract
x10^5
395.2523
(M+Na)+
411.2258
(M+K)+
373.2697
(M+H)+
375.2725
(M+H)+
Counts vs. Mass-to-Charge (m/z)

Boc-YV-O^{t}Bu **C14**

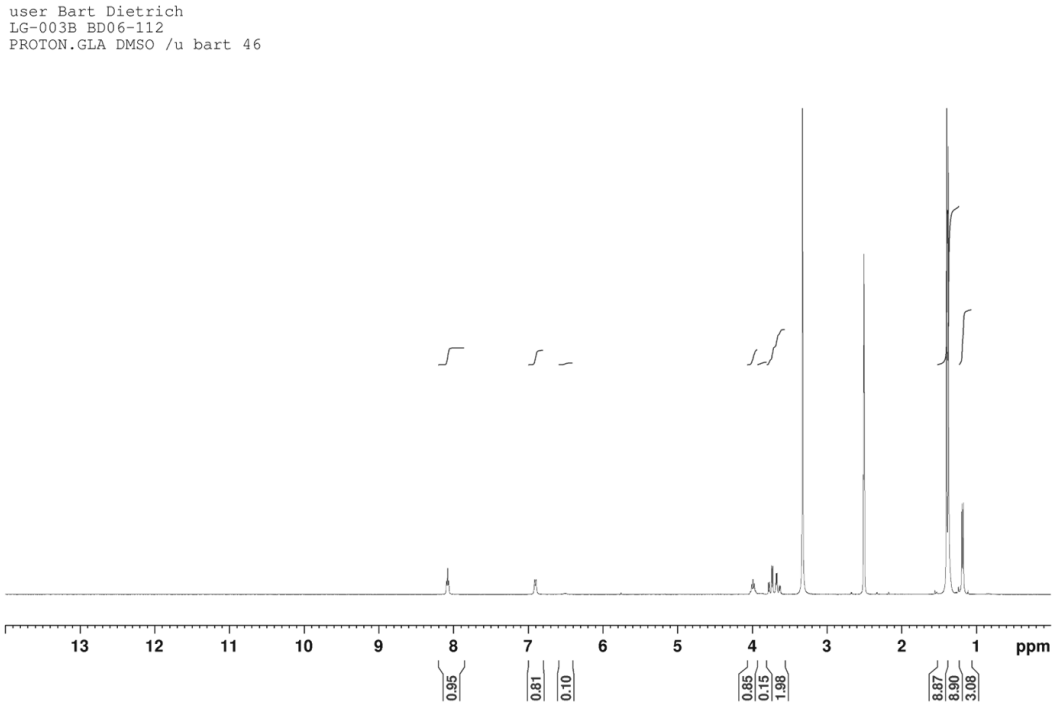


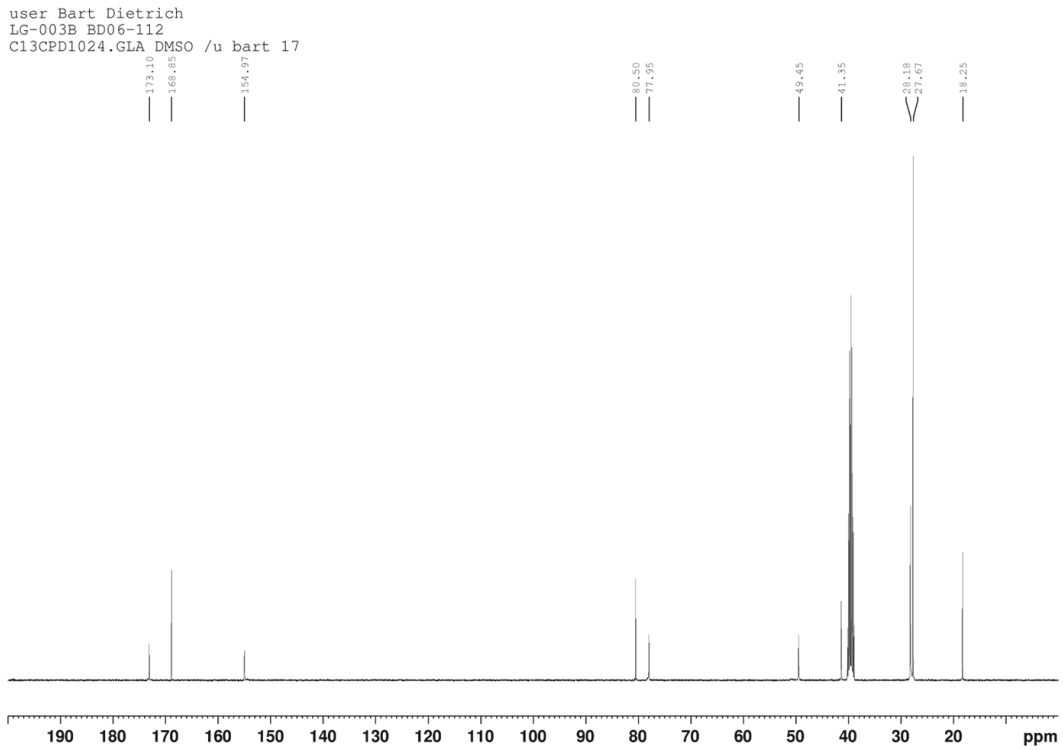


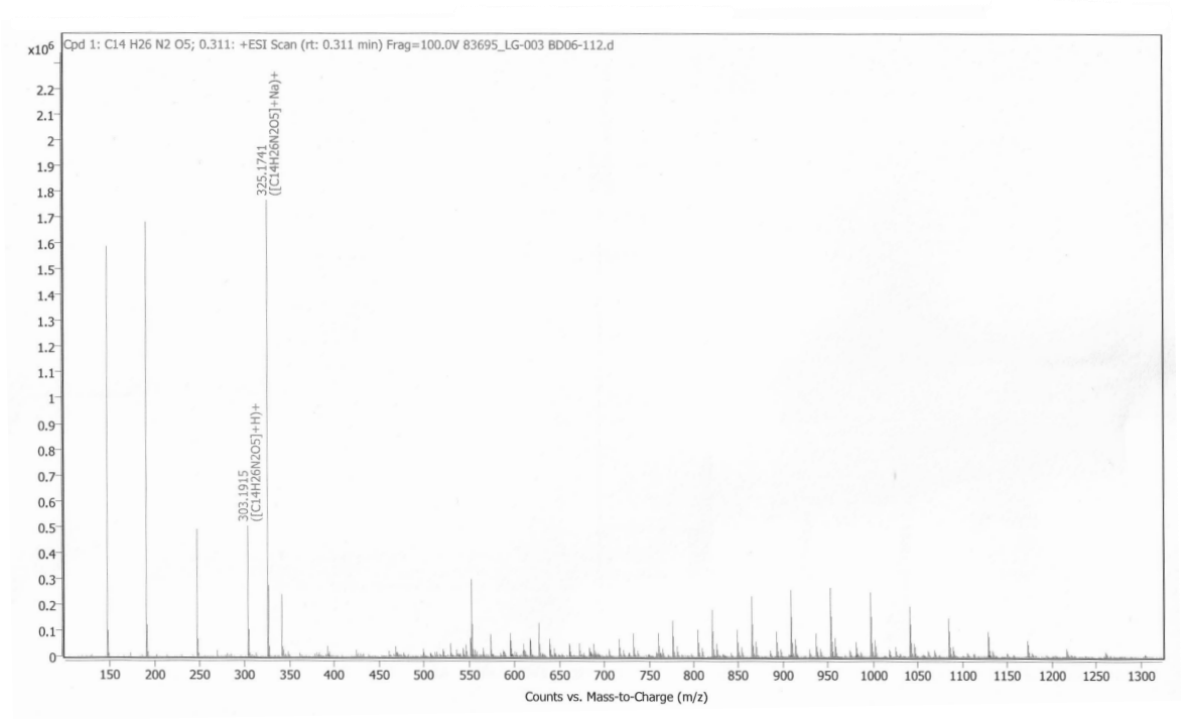

Cpd 1: C27 H44 N2 O6; 0.681: +ESI Scan (rt: 0.736-0.753 min, 4 scans) Frag=100.0V 84926_BOC-YV-TBU.d
x10^6
515.3104
(M+Na)+
1007.6295
(2M+Na)+
247.1659
(M+2H)+2
496.3352
(M+H)+
985.6488
(2M+H)+
Counts vs. Mass-to-Charge (m/z)

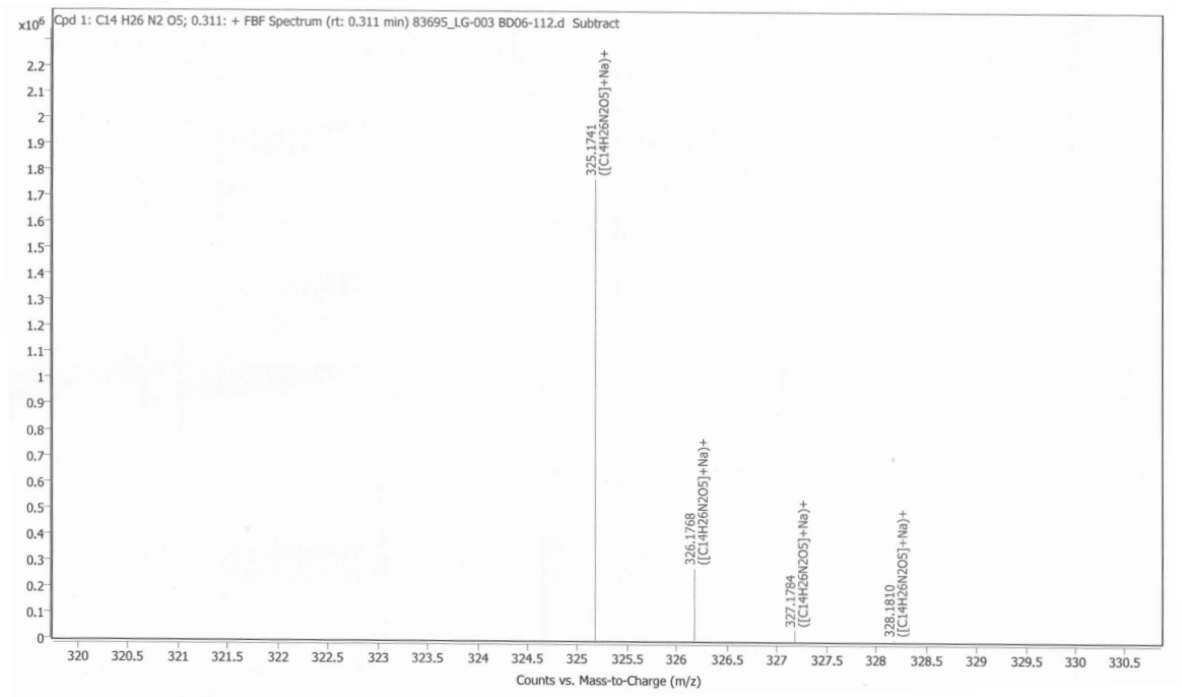

Cpd 1: C27 H44 N2 O6; 0.681: + FBF Spectrum (rt: 0.736-0.753 min) 84926_BOC-YV-TBU.d Subtract
x10^6
515.3104
(M+Na)+
493.3277
(M+H)+
1007.6295
(2M+Na)+
496.3352
(M+H)+
534.2841
(M+K)+
985.6488
(2M+H)+
Counts vs. Mass-to-Charge (m/z)

Boc-AG-O^{t}Bu **C15**

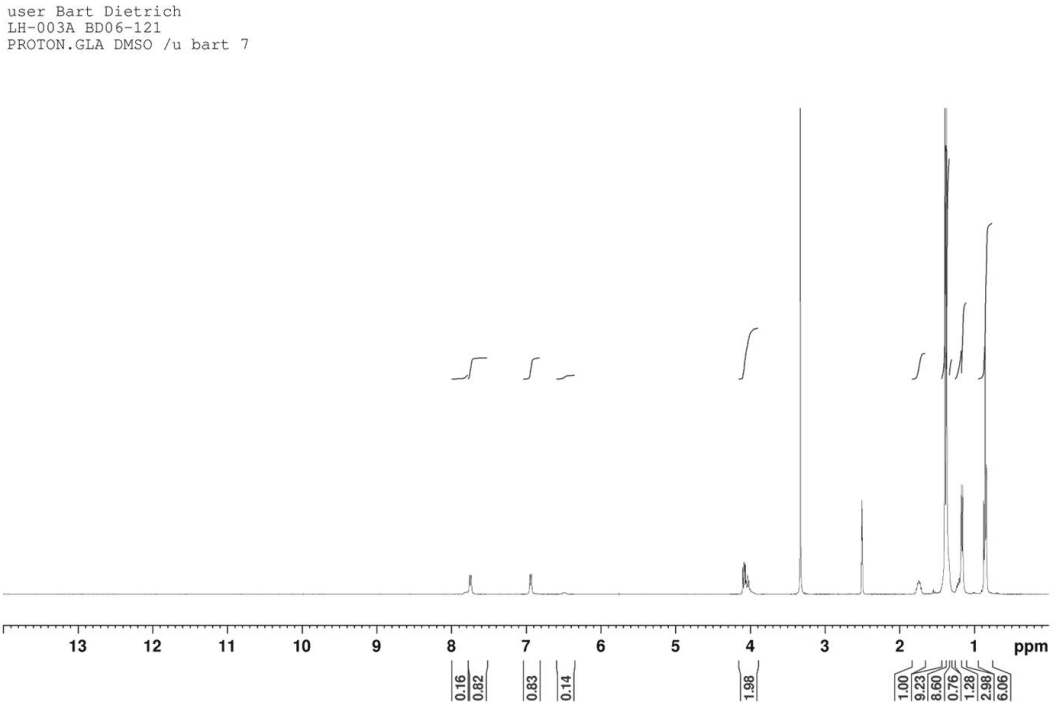


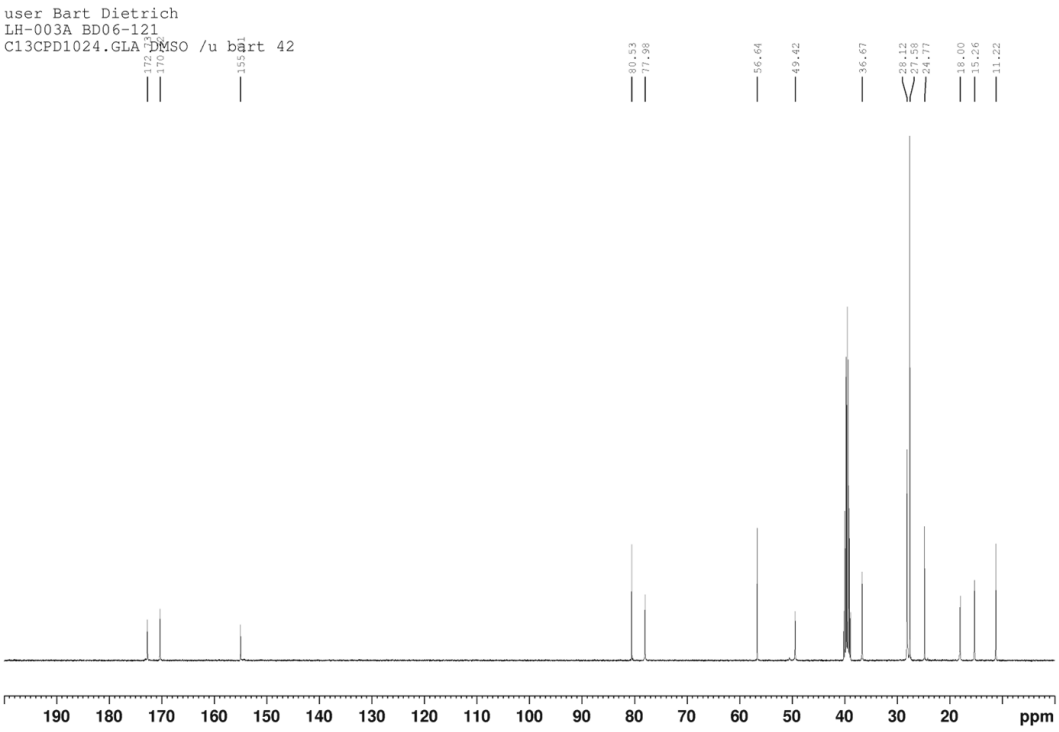


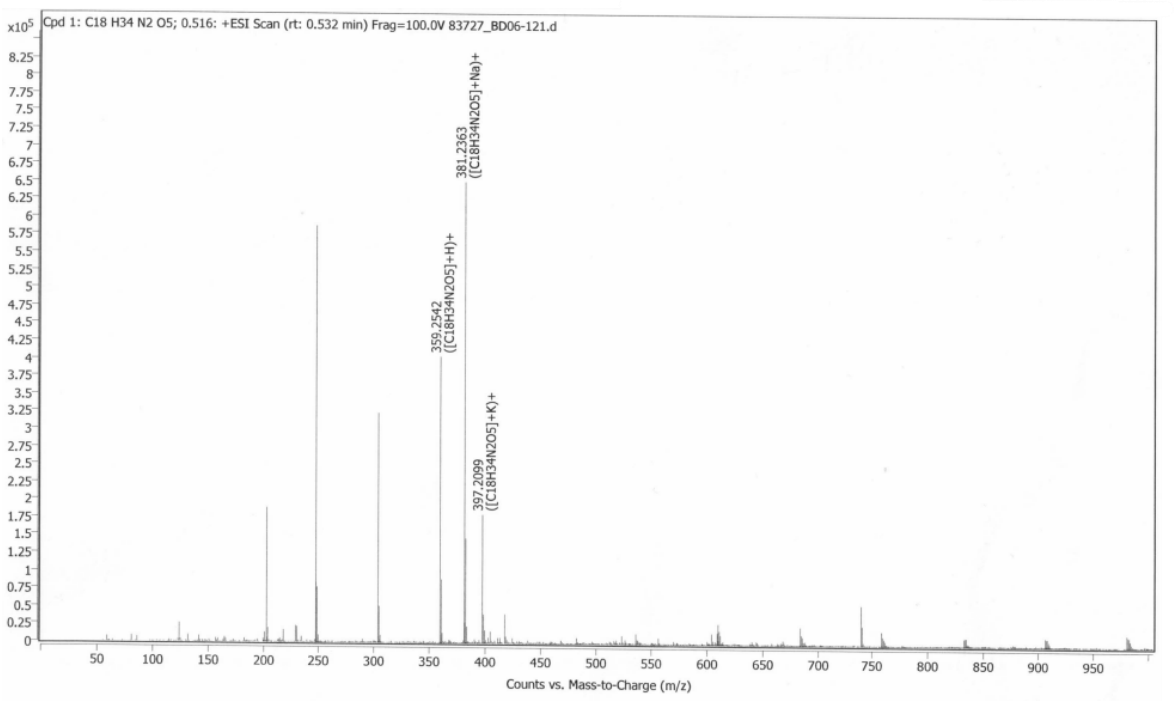

Cpd 1: C14 H26 N2 O5; 0.311: +ESI Scan (rt: 0.311 min) Frag=100.0V 83695_LG-003 BD06-112.d
x10^6
325.1741
([C14H26N2O5]+Na)+
303.1915
([C14H26N2O5]+H)+
Counts vs. Mass-to-Charge (m/z)

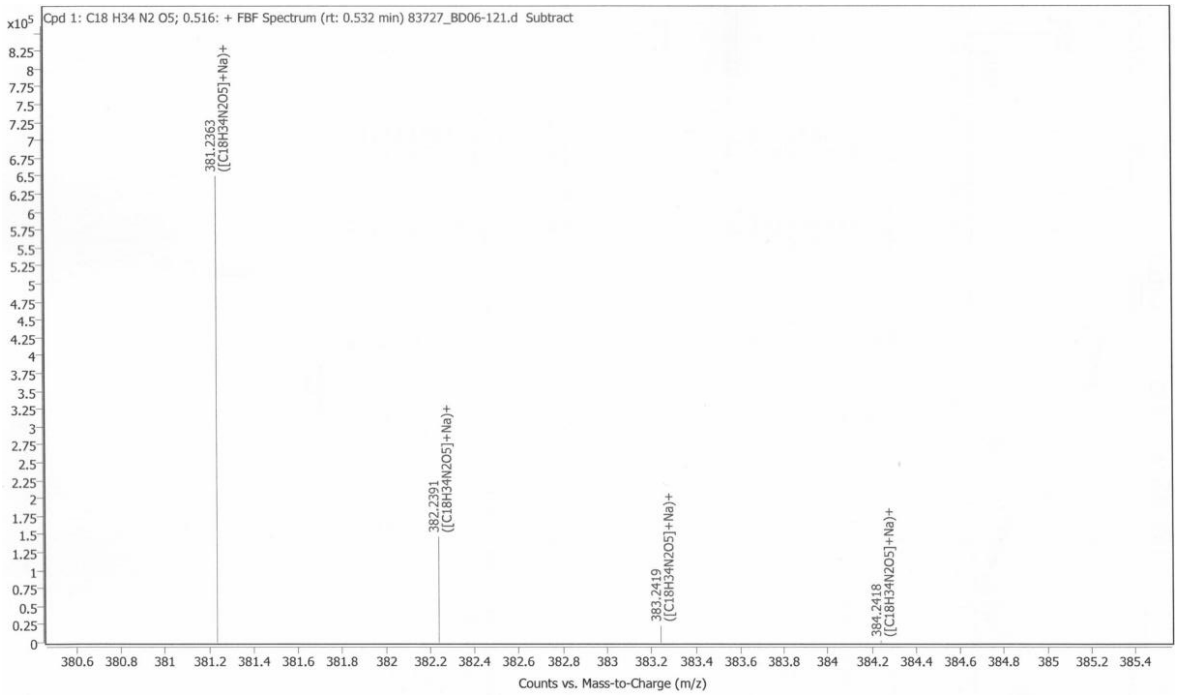

Cpd 1: C14 H26 N2 O5; 0.311: + FBF Spectrum (rt: 0.311 min) 83695_LG-003 BD06-112.d Subtract
x10^6
325.1741
([C14H26N2O5]+Na)+
326.1768
([C14H26N2O5]+Na)+
327.1784
([C14H26N2O5]+Na)+
328.1810
([C14H26N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-AI-O^{t}Bu **C16**

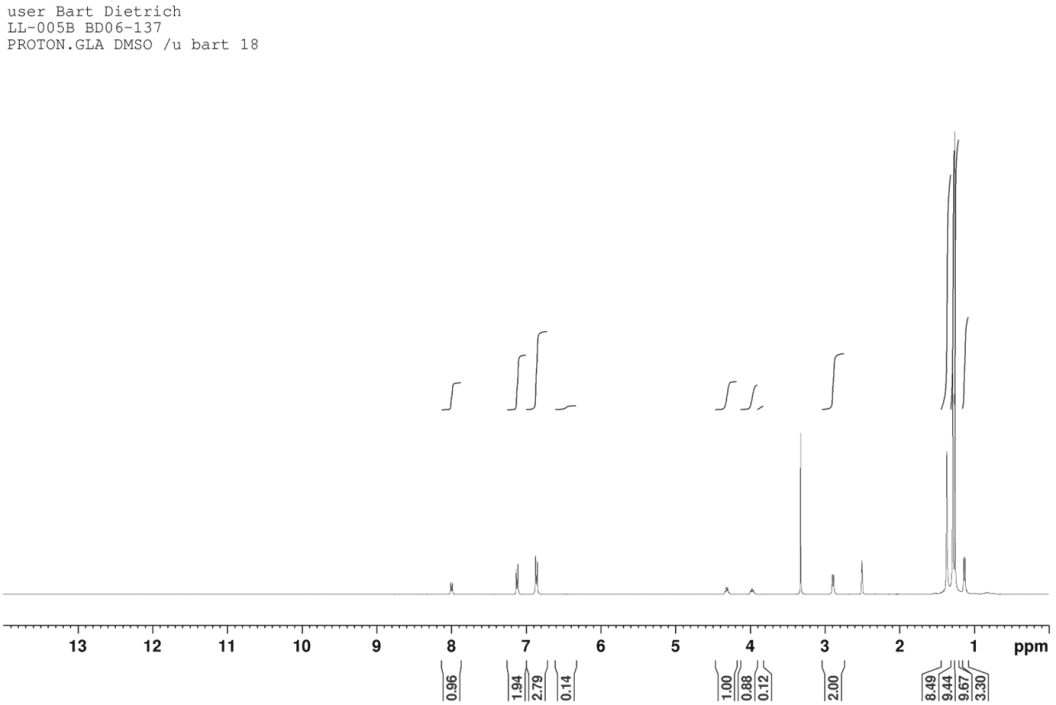


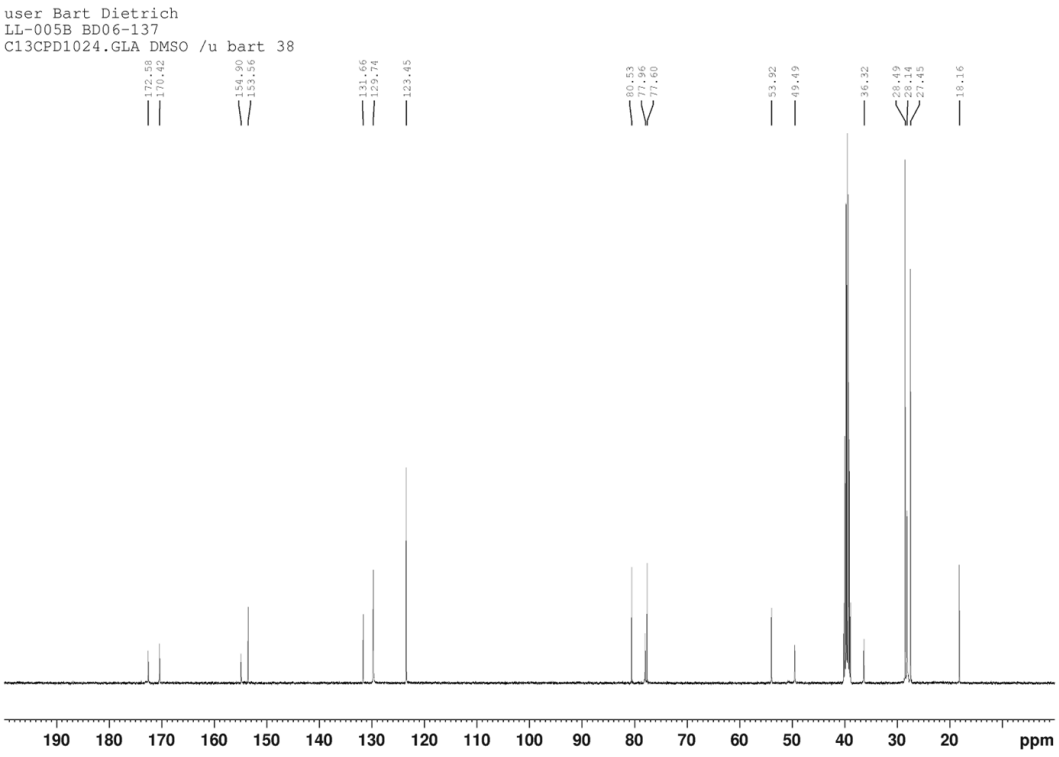


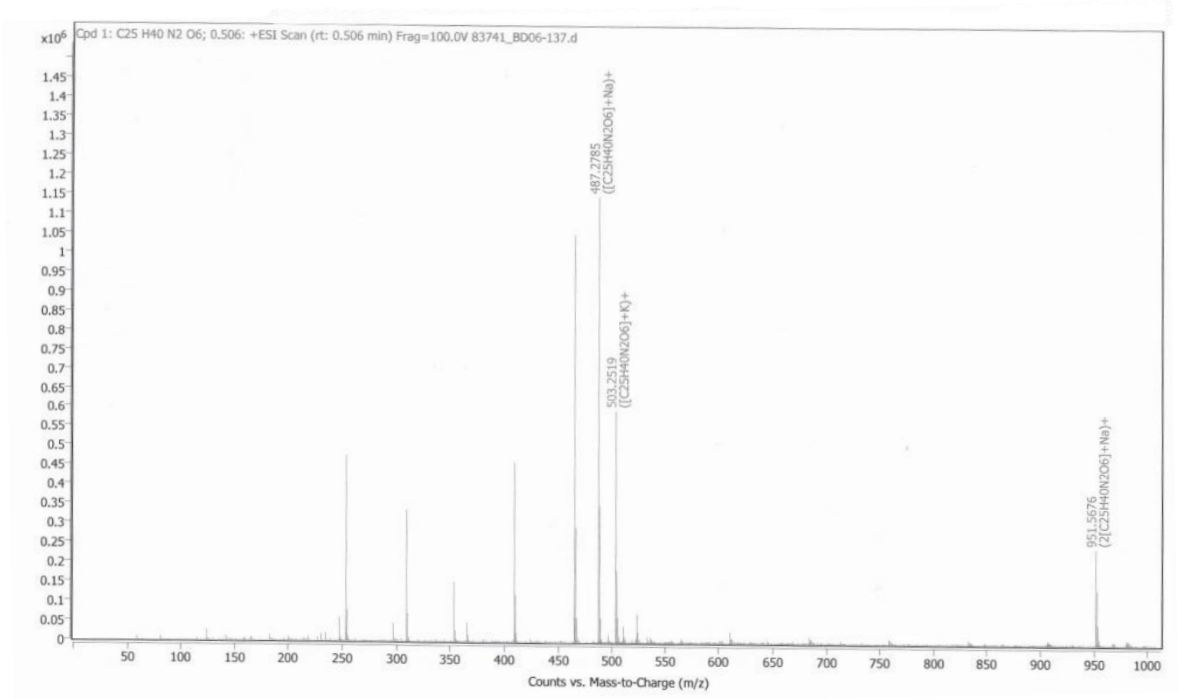

Cpd 1: C18 H34 N2 O5; 0.516: +ESI Scan (rt: 0.532 min) Frag=100.0V 83727_BD06-121.d
x10^5
381.2363 ([C18H34N2O5]+Na)+
359.2542 ([C18H34N2O5]+H)+
397.2099 ([C18H34N2O5]+K)+
Counts vs. Mass-to-Charge (m/z)

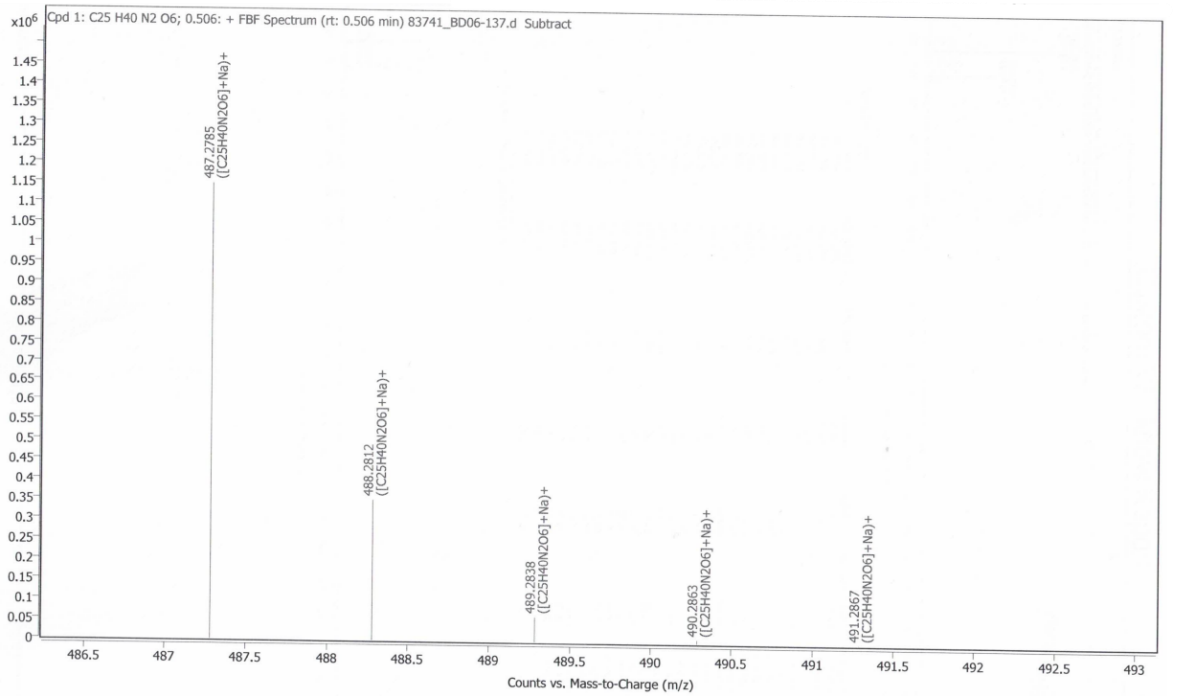

Cpd 1: C18 H34 N2 O5; 0.516: + FBF Spectrum (rt: 0.532 min) 83727_BD06-121.d Subtract
x10^5
381.2363 ([C18H34N2O5]+Na)+
382.2391 ([C18H34N2O5]+Na)+
383.2419 ([C18H34N2O5]+Na)+
384.2418 ([C18H34N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-AY-O^{t}Bu **C17**

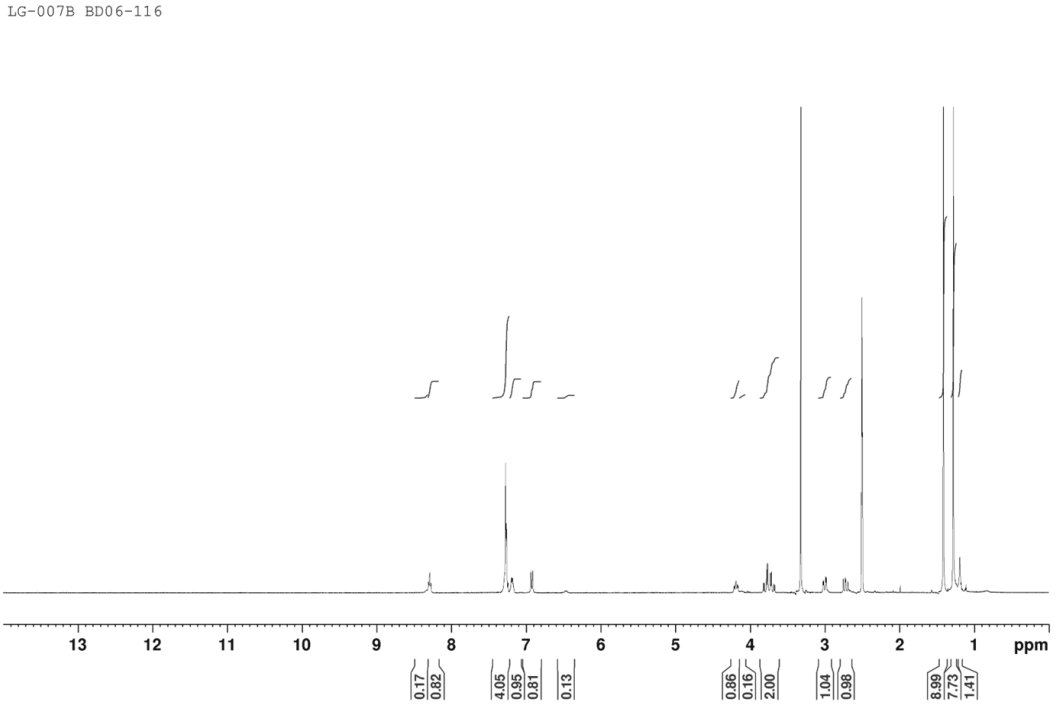


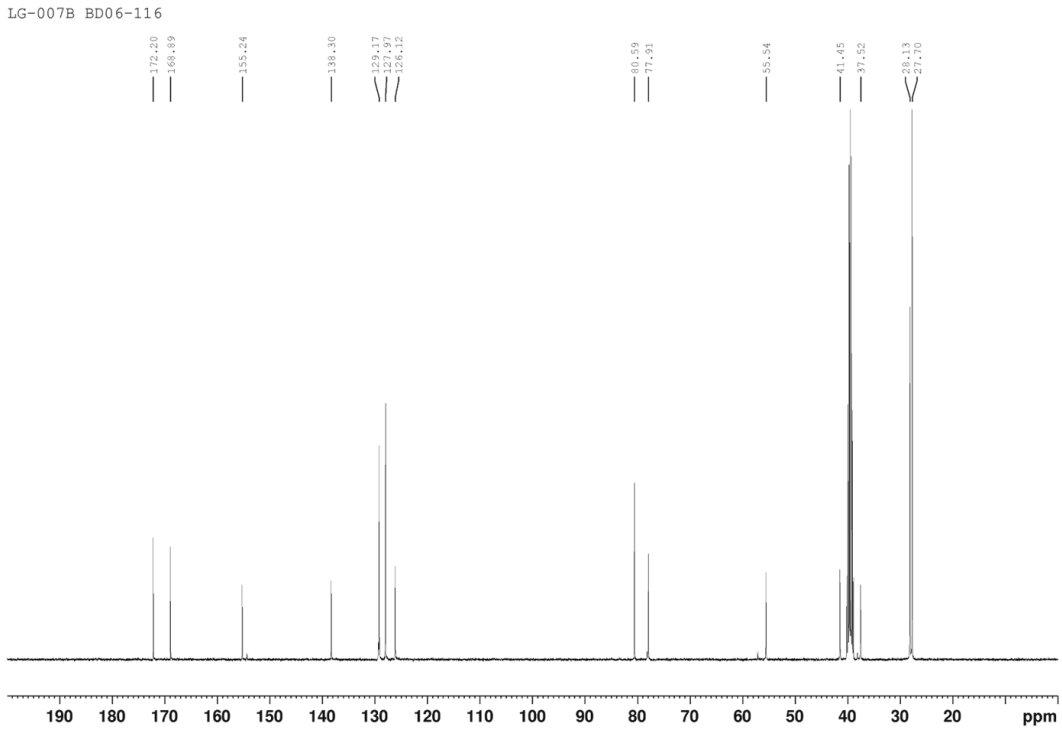


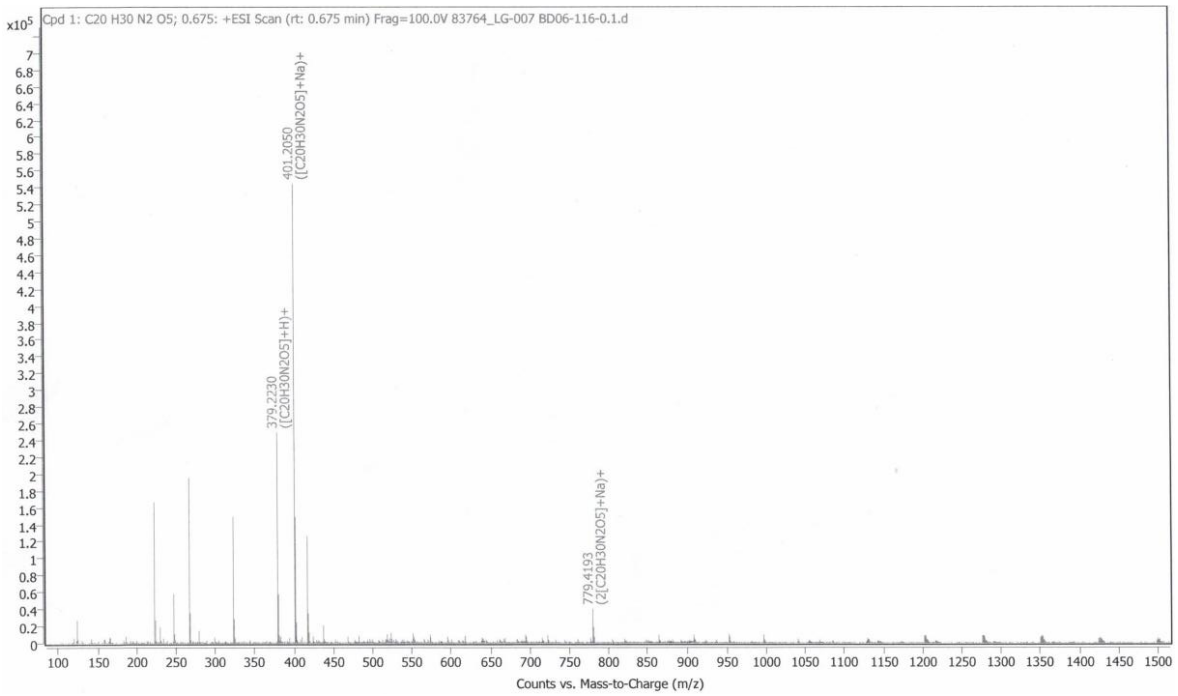

Cpd 1: C25 H40 N2 O6; 0.506: +ESI Scan (rt: 0.506 min) Frag=100.0V 83741_BD06-137.d
x10^6
487.2785 ([C25H40N2O6]+Na)+
503.2519 ([C25H40N2O6]+K)+
951.5676 (2[C25H40N2O6]+Na)+
Counts vs. Mass-to-Charge (m/z)

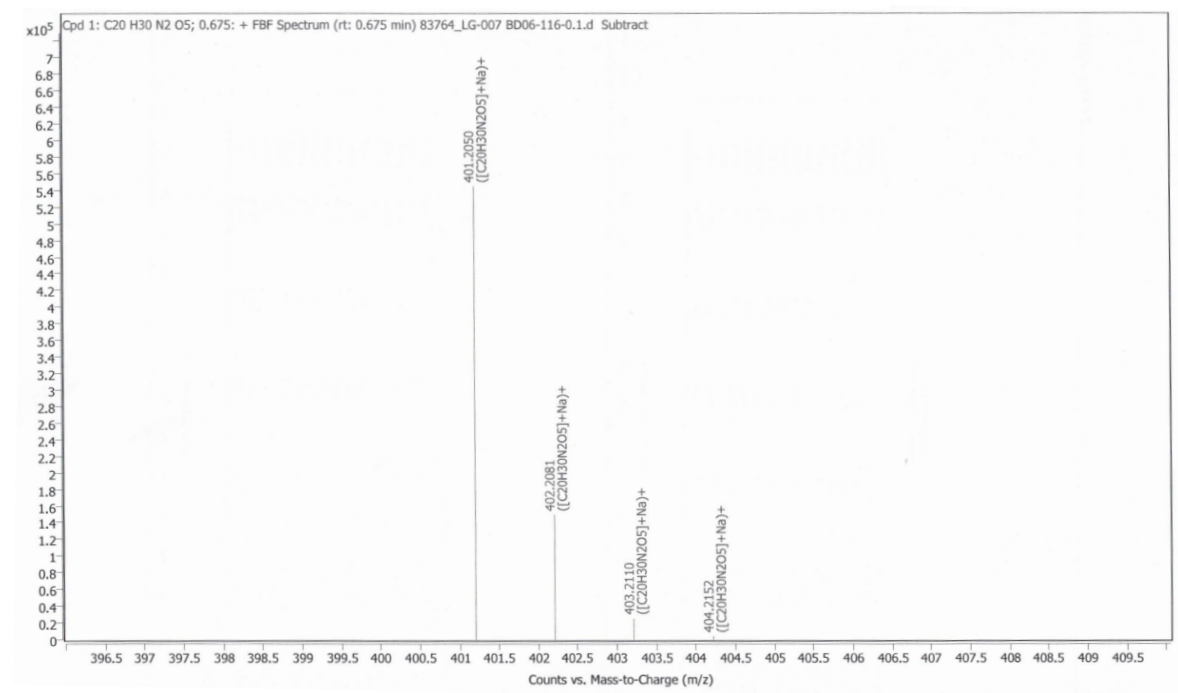

Cpd 1: C25 H40 N2 O6; 0.506: + FBF Spectrum (rt: 0.506 min) 83741_BD06-137.d Subtract
x10^6
487.2785 ([C25H40N2O6]+Na)+
488.2812 ([C25H40N2O6]+Na)+
489.2838 ([C25H40N2O6]+Na)+
490.2863 ([C25H40N2O6]+Na)+
491.2867 ([C25H40N2O6]+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-FG-O^{t}Bu **C18**

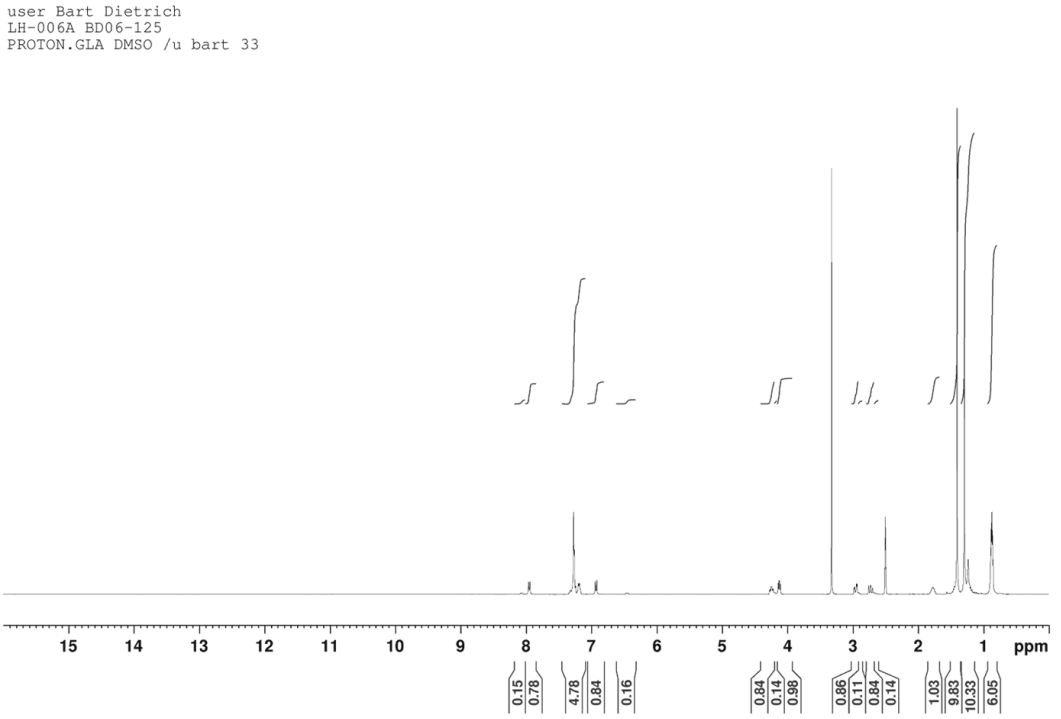


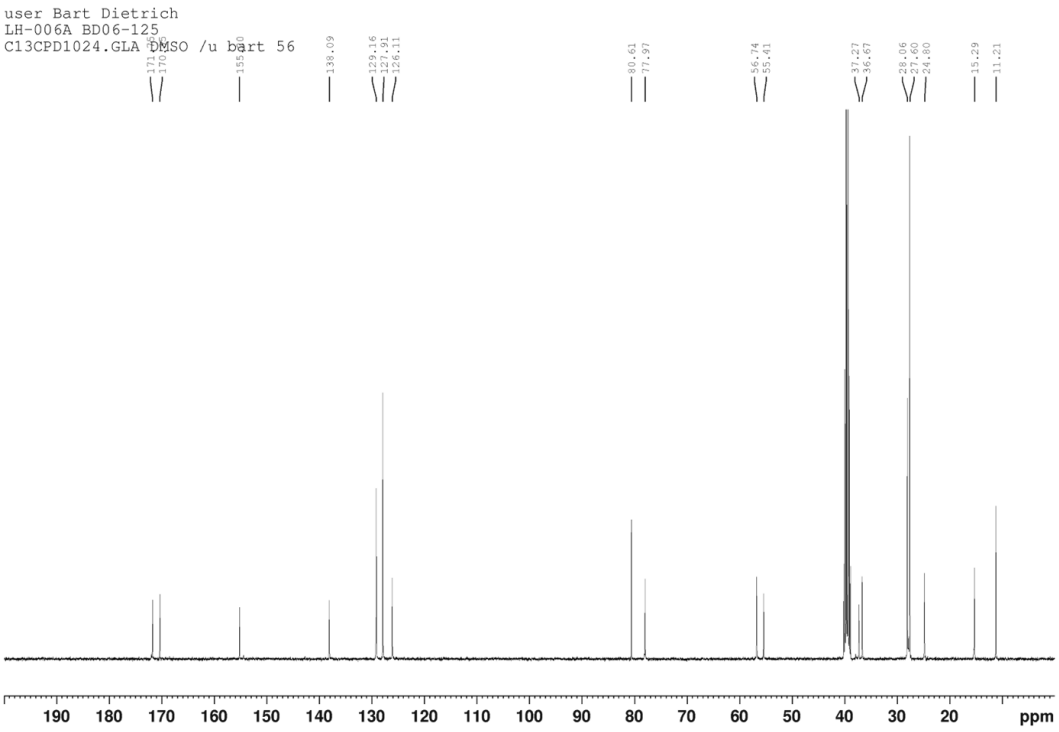


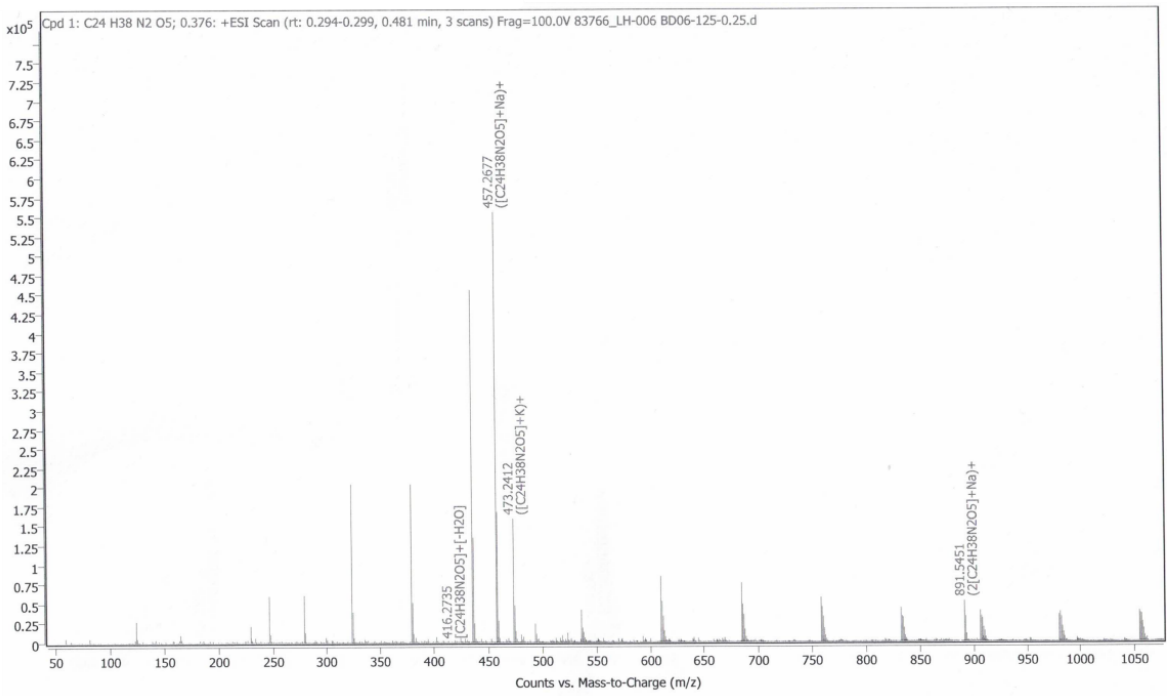

Cpd 1: C20 H30 N2 O5; 0.675: +ESI Scan (rt: 0.675 min) Frag=100.0V 83764_LG-007 BD06-116-0.1.d
x10^5
401.2050
([C20H30N2O5]+Na)+
379.2230
([C20H30N2O5]+H)+
779.4193
(2[C20H30N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

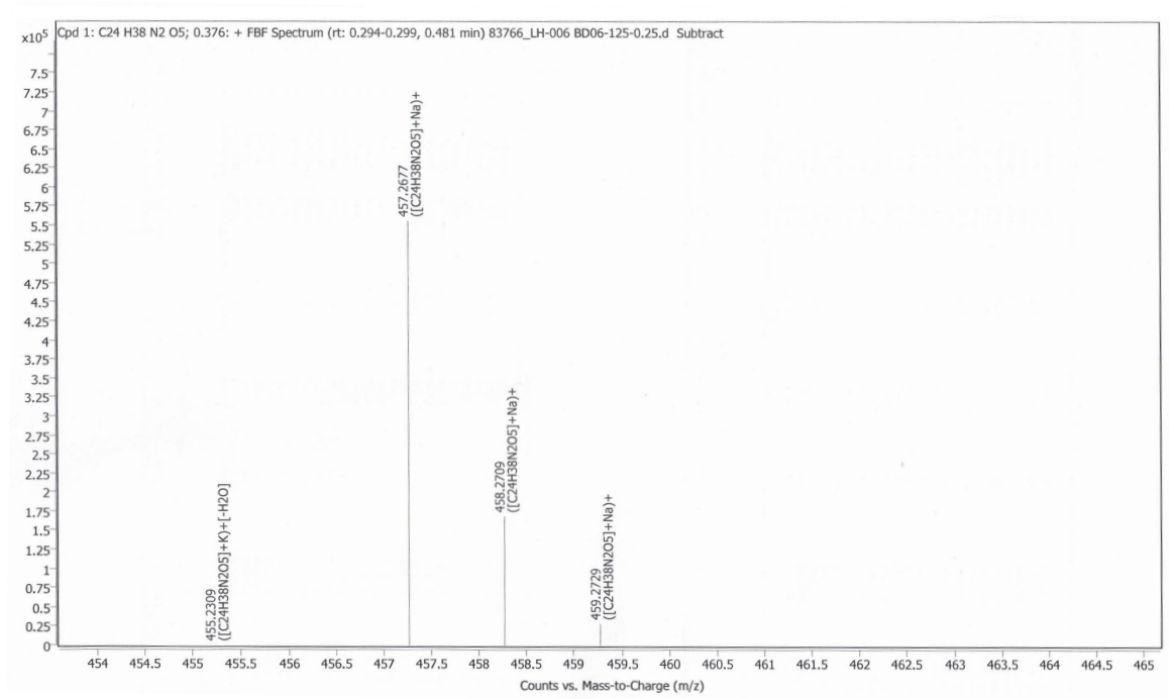

Cpd 1: C20 H30 N2 O5; 0.675: + FBF Spectrum (rt: 0.675 min) 83764_LG-007 BD06-116-0.1.d Subtract
x10^5
401.2050
([C20H30N2O5]+Na)+
402.2081
([C20H30N2O5]+Na)+
403.2110
([C20H30N2O5]+Na)+
404.2152
([C20H30N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-FI-O^{t}Bu **C19**

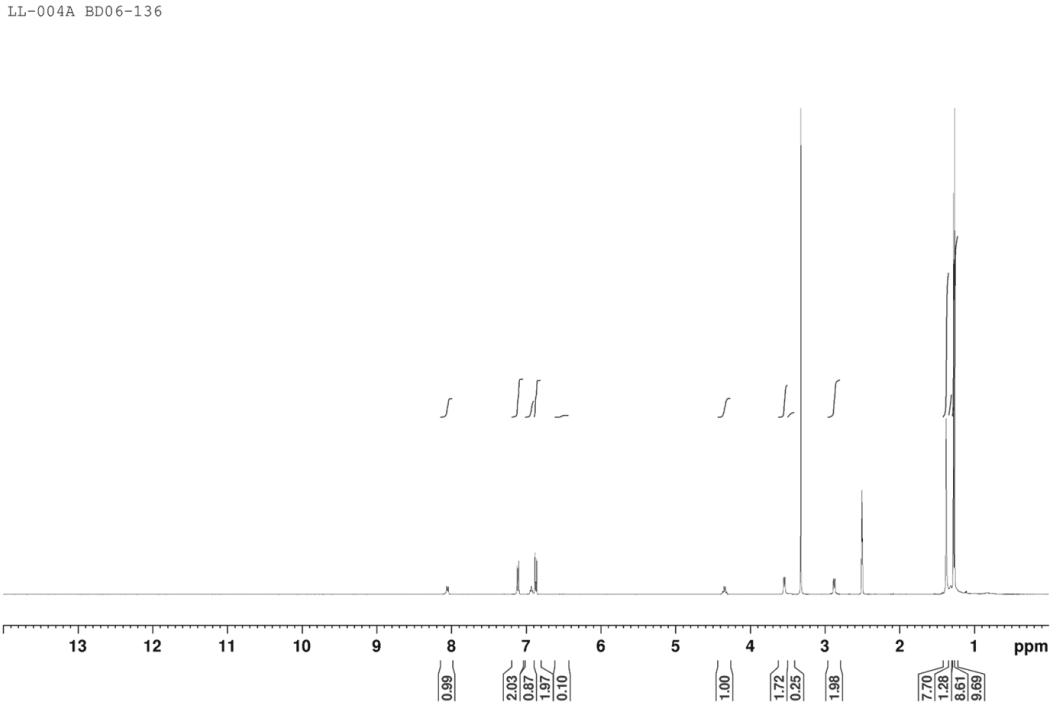


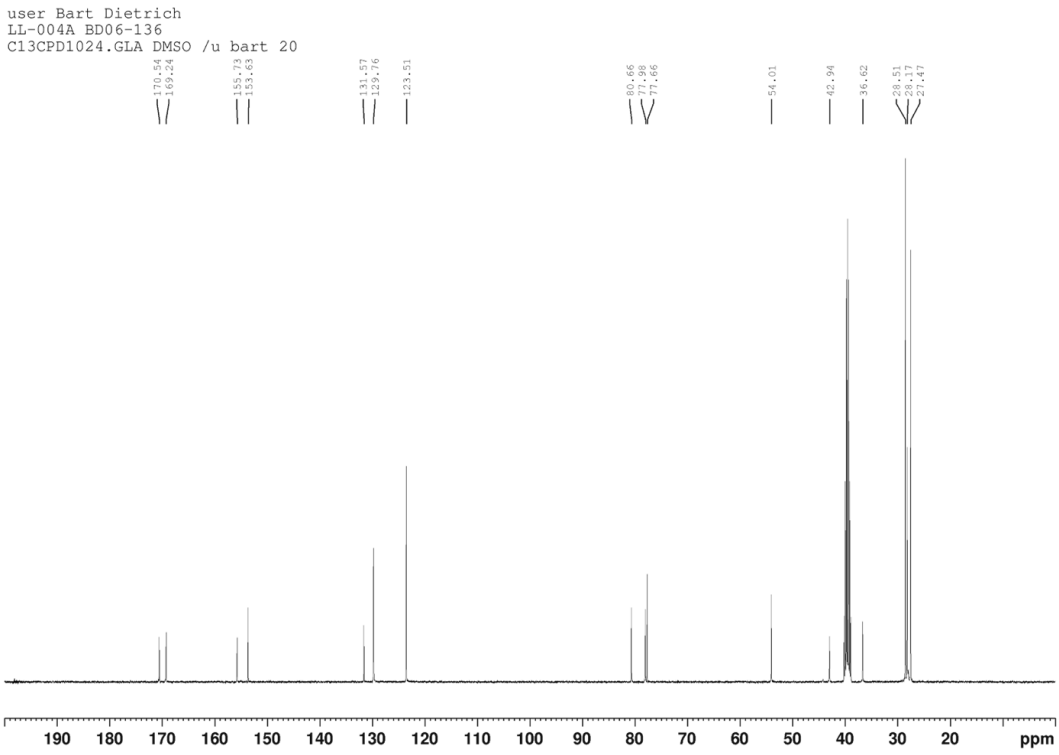


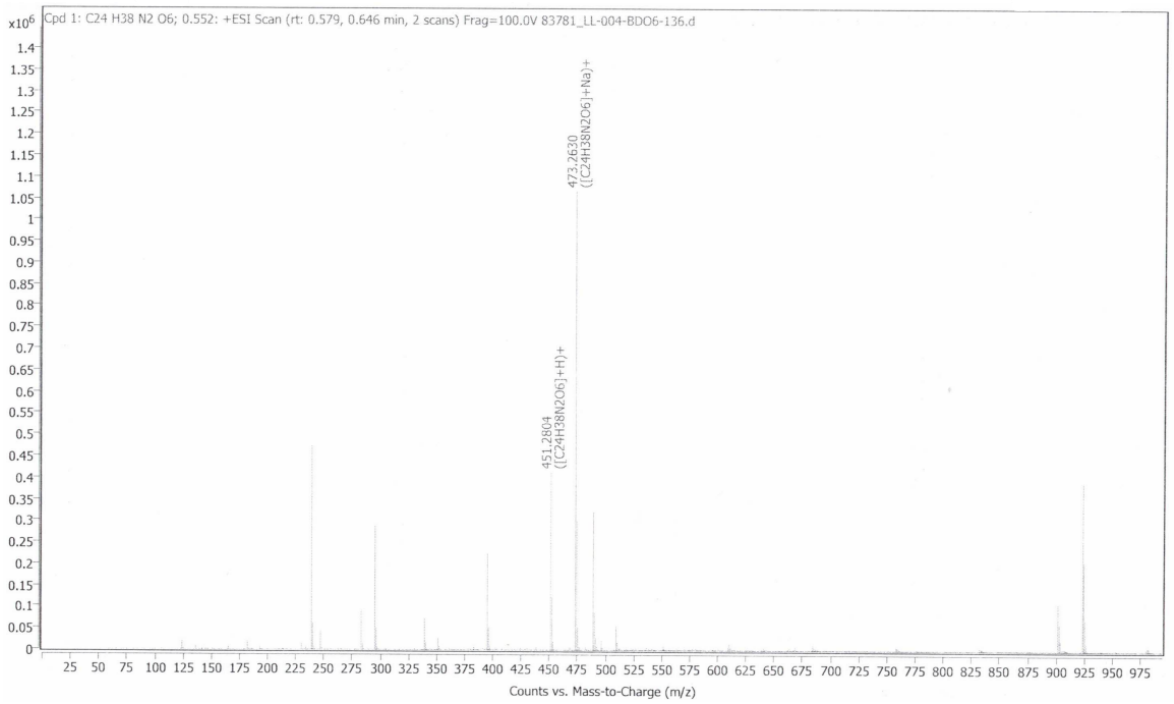

Cpd 1: C24 H38 N2 O5; 0.376: +ESI Scan (rt: 0.294-0.299, 0.481 min, 3 scans) Frag=100.0V 83766_LH-006 BD06-125-0.25.d
x10^5
457.2677 ([C24H38N2O5]+Na)+
473.2412 ([C24H38N2O5]+K)+
416.2735 [C24H38N2O5]+[-H2O]
891.5451 (2[C24H38N2O5]+Na)+
Counts vs. Mass-to-Charge (m/z)

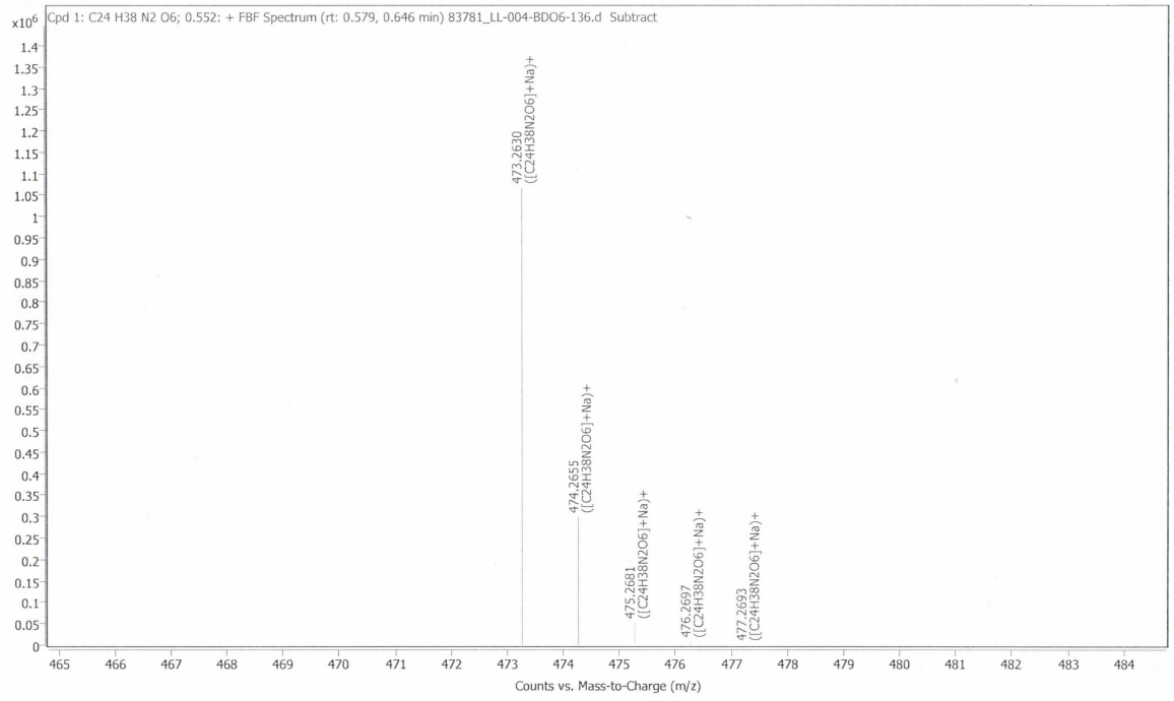

Cpd 1: C24 H38 N2 O5; 0.376: + FBF Spectrum (rt: 0.294-0.299, 0.481 min) 83766_LH-006 BD06-125-0.25.d Subtract
x10^5
457.2677 ([C24H38N2O5]+Na)+
458.2709 ([C24H38N2O5]+Na)+
459.2729 ([C24H38N2O5]+Na)+
455.2309 [C24H38N2O5]+K+[-H2O]
Counts vs. Mass-to-Charge (m/z)

Boc-GY-O^{t}Bu **C20**

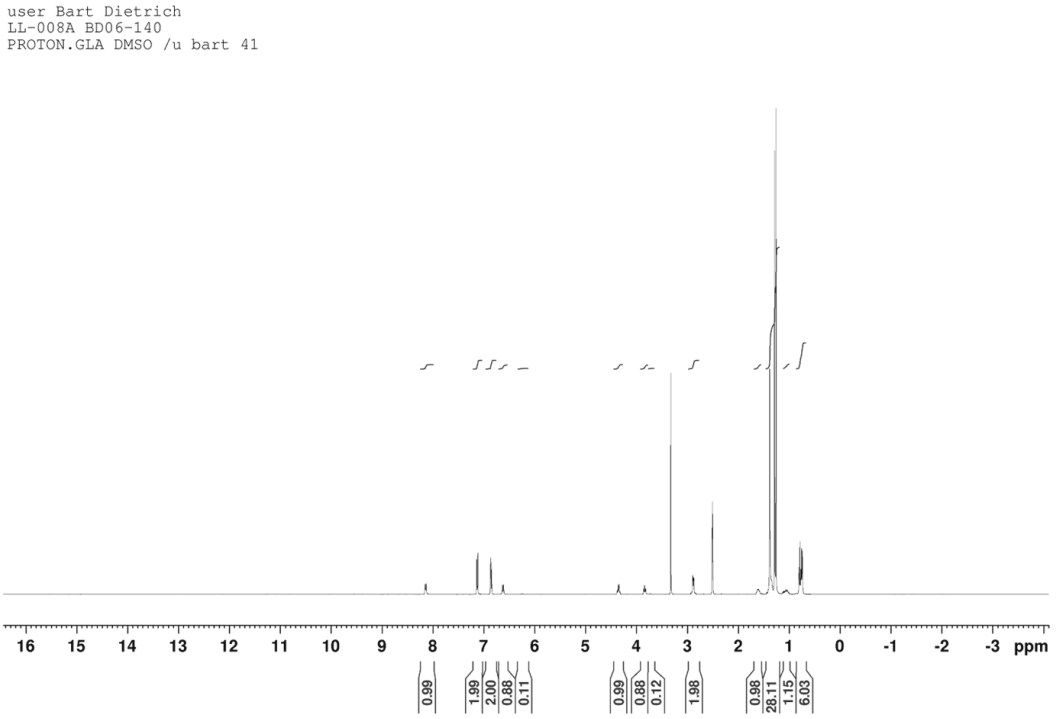


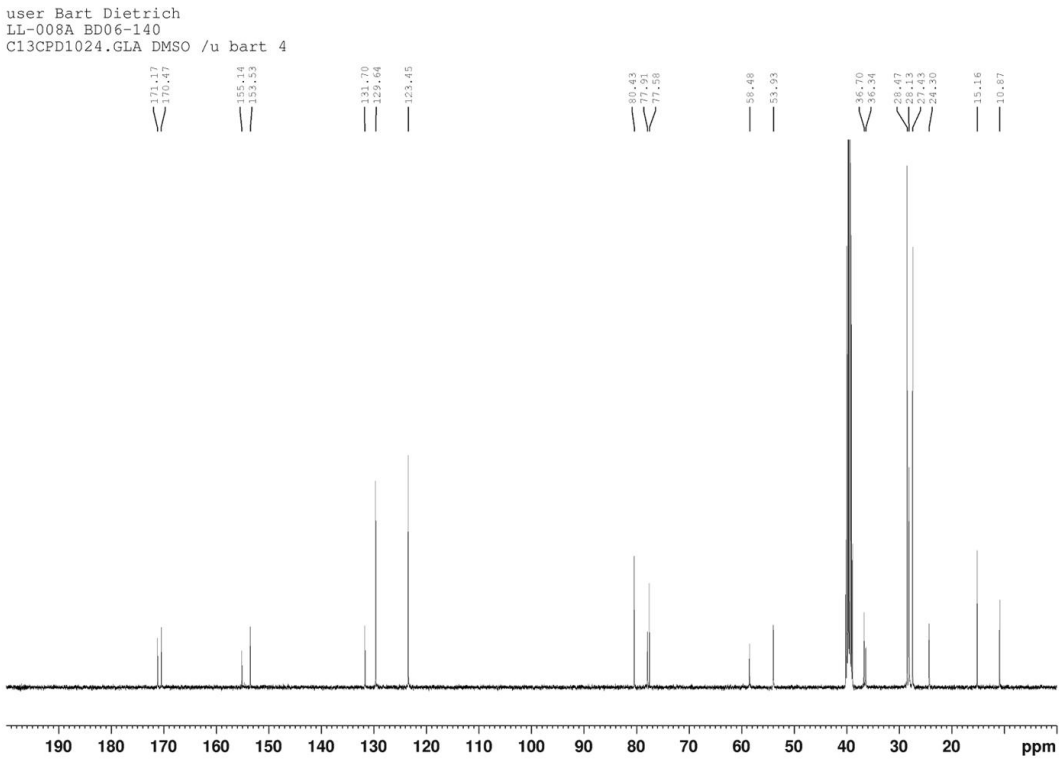


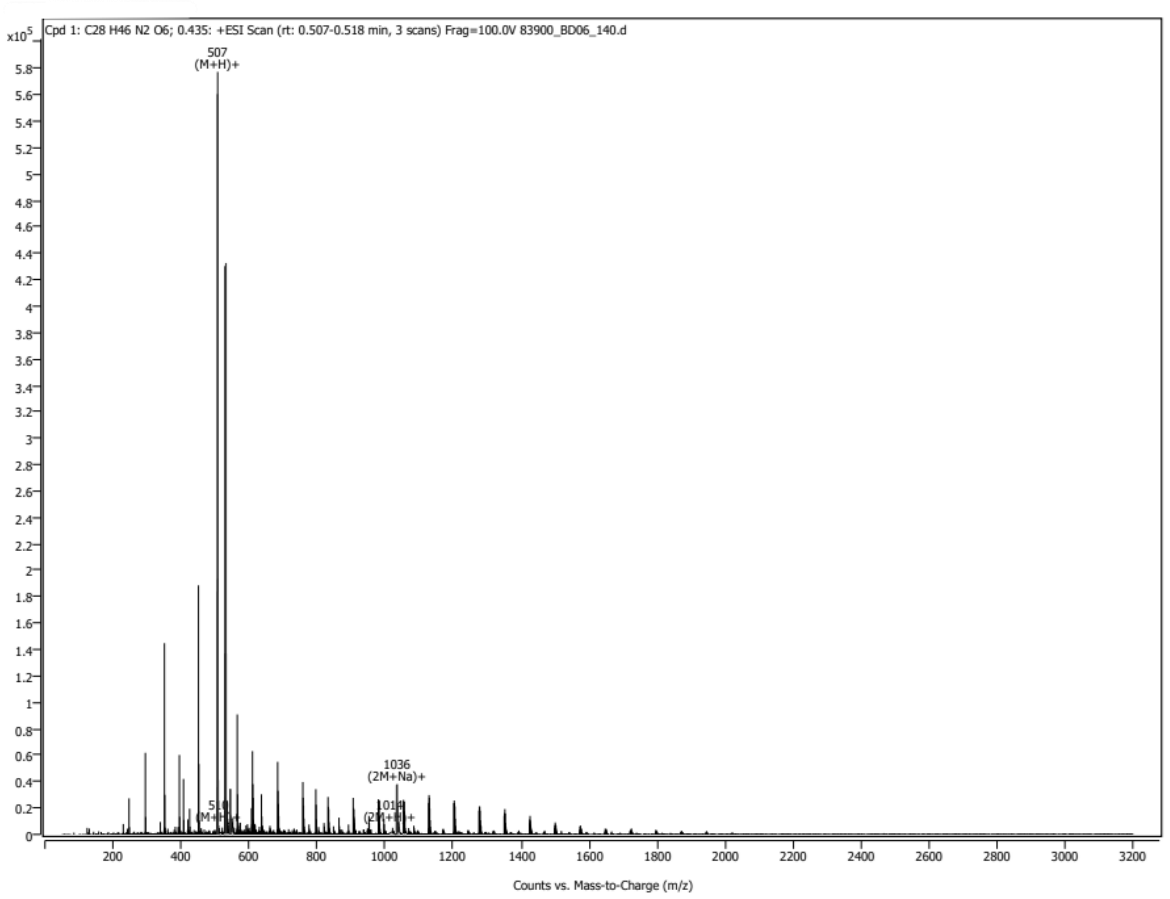

Cpd 1: C24 H38 N2 O6; 0.552: +ESI Scan (rt: 0.579, 0.646 min, 2 scans) Frag=100.0V 83781_LL-004-BDO6-136.d
x10^6
473.2630
([C24H38N2O6]+Na)+
451.2804
([C24H38N2O6]+H)+
Counts vs. Mass-to-Charge (m/z)

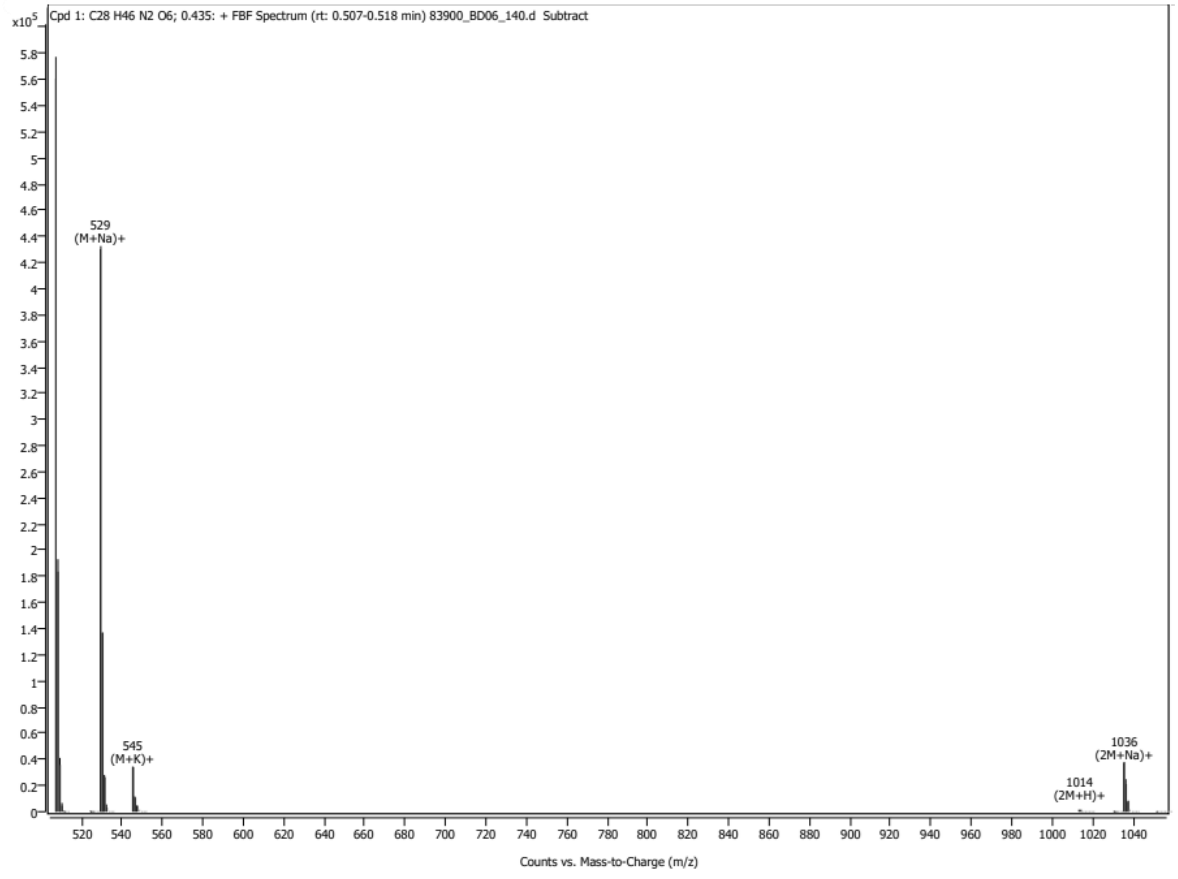

Cpd 1: C24 H38 N2 O6; 0.552: + FBF Spectrum (rt: 0.579, 0.646 min) 83781_LL-004-BDO6-136.d Subtract
x10^6
473.2630
([C24H38N2O6]+Na)+
474.2655
([C24H38N2O6]+Na)+
475.2681
([C24H38N2O6]+Na)+
476.2697
([C24H38N2O6]+Na)+
477.2693
([C24H38N2O6]+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-IY-O^{t}Bu **C21**

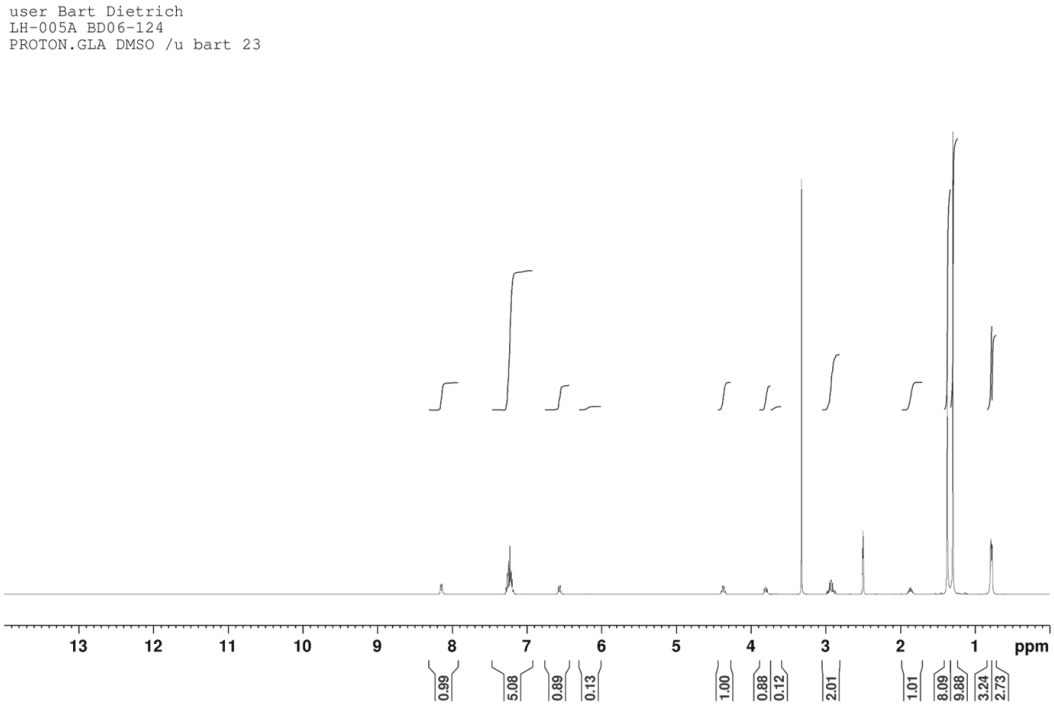


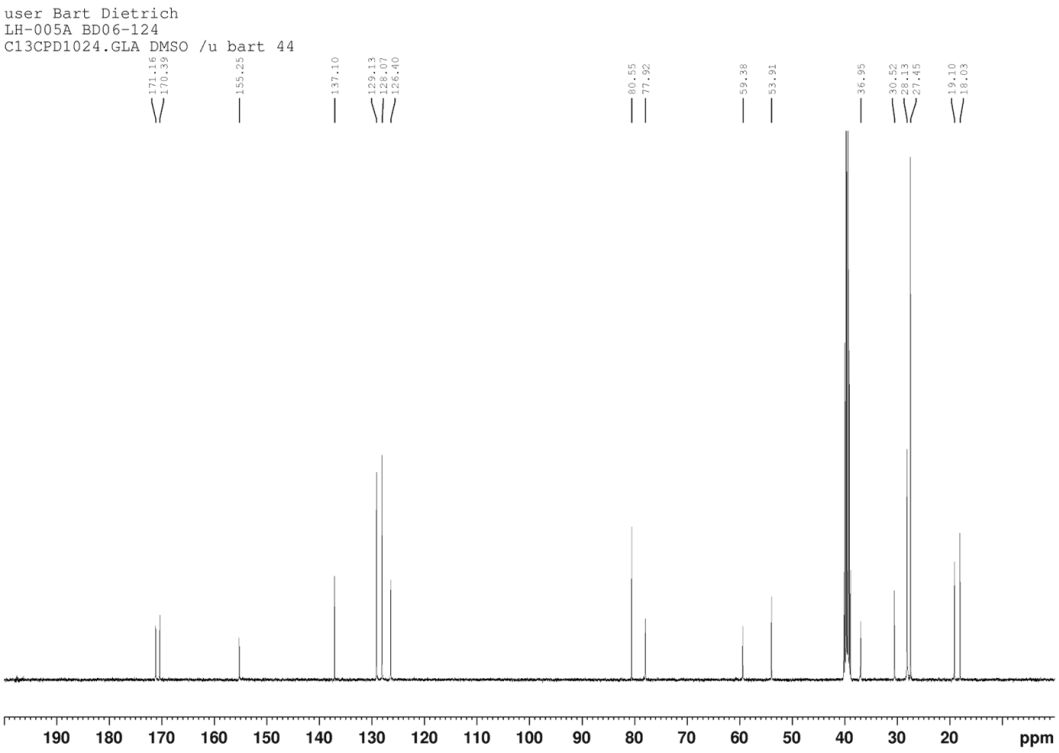


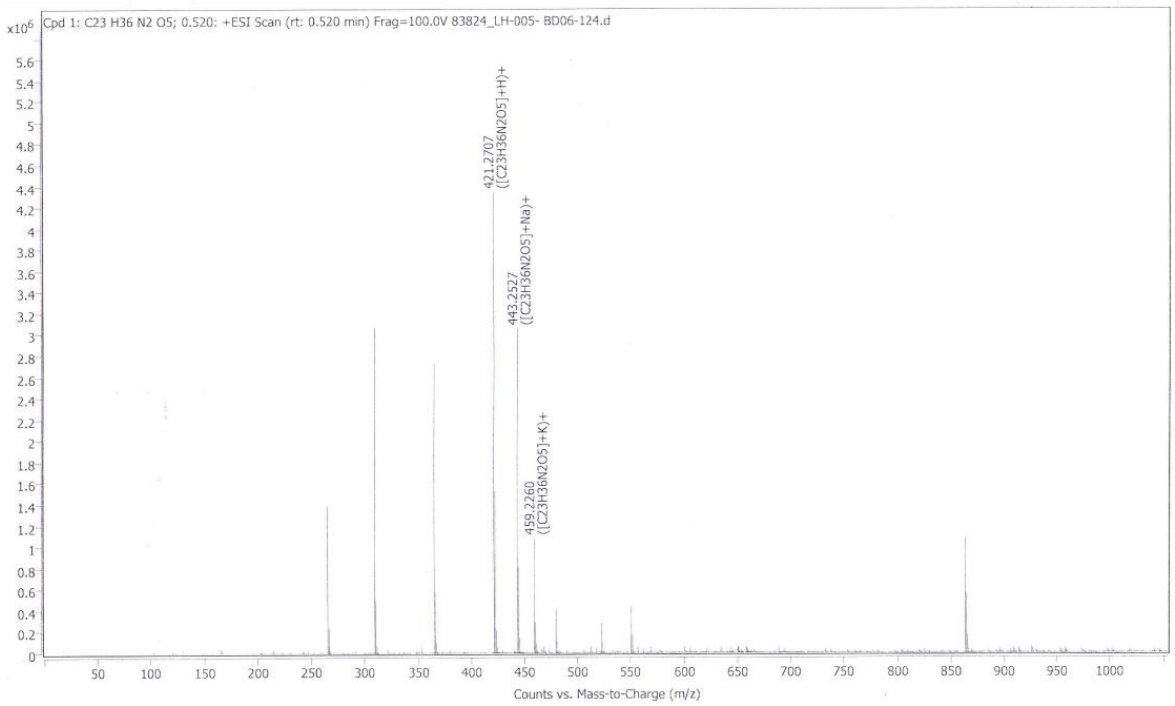

Cpd 1: C28 H46 N2 O6; 0.435: +ESI Scan (rt: 0.507-0.518 min, 3 scans) Frag=100.0V 83900_BD06_140.d
x10^5
507
(M+H)+
1036
(2M+Na)+
Counts vs. Mass-to-Charge (m/z)

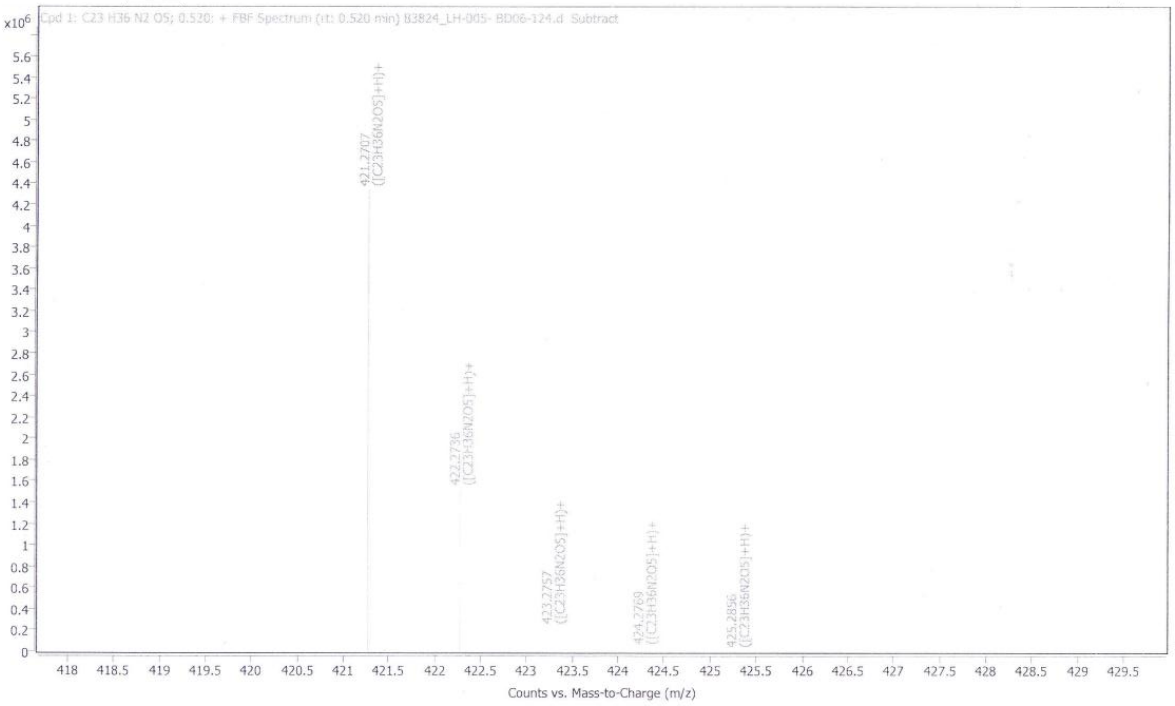

Cpd 1: C28 H46 N2 O6; 0.435: + FBF Spectrum (rt: 0.507-0.518 min) 83900_BD06_140.d Subtract
x10^5
529
(M+Na)+
545
(M+K)+
1014
(2M+H)+
1036
(2M+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-VF-O^{t}Bu **C22**

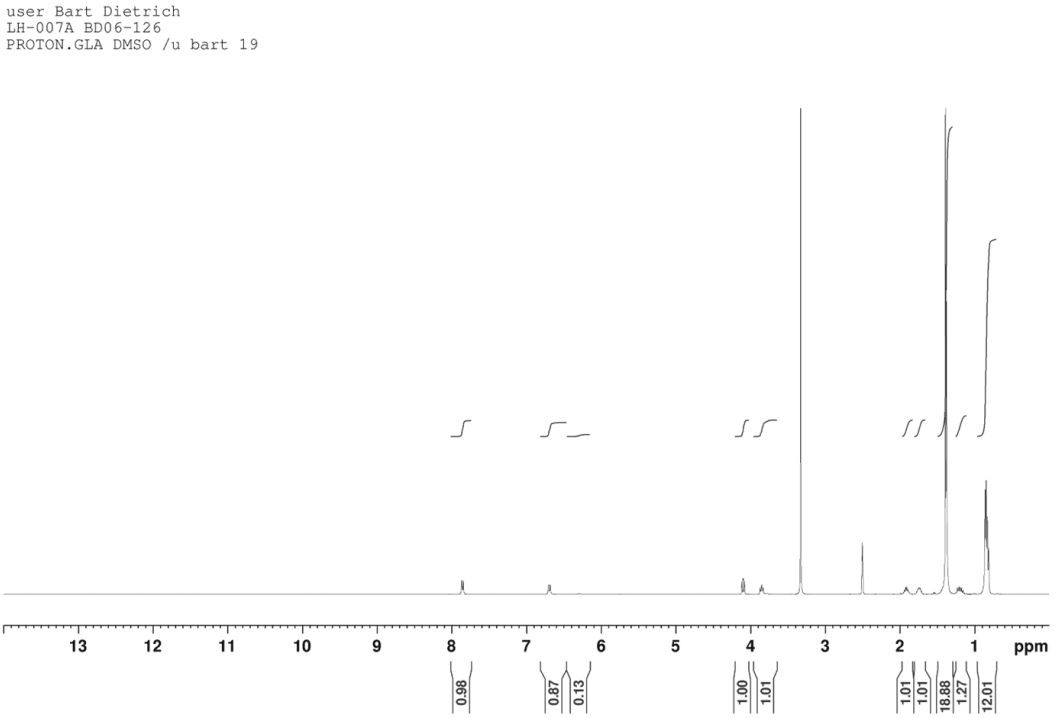


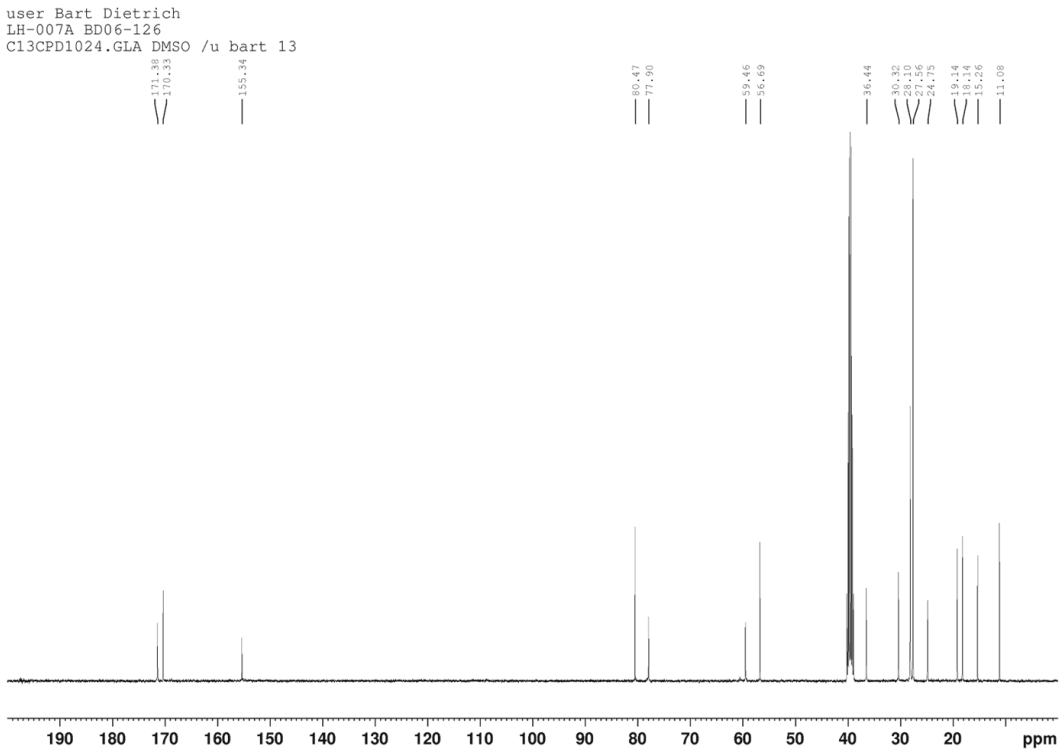


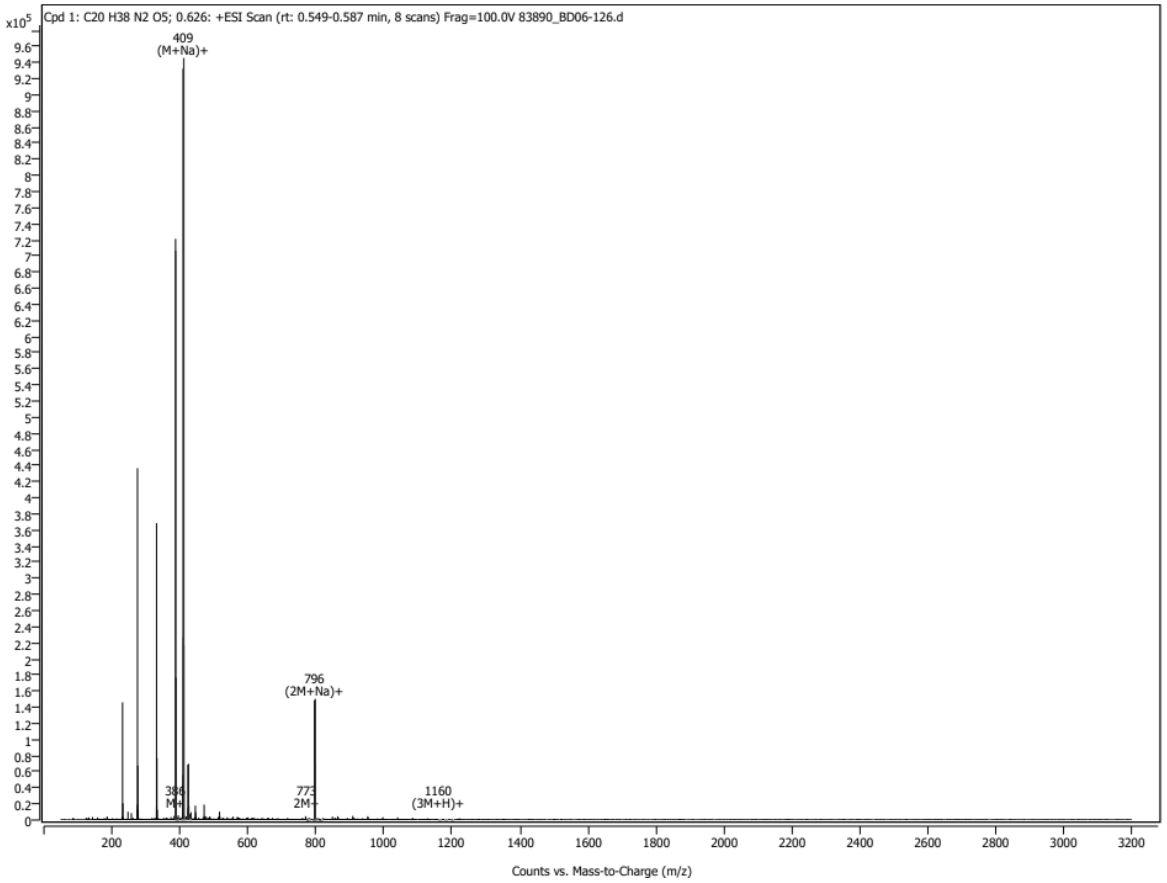

Cpd 1: C23 H36 N2 O5; 0.520: +ESI Scan (rt: 0.520 min) Frag=100.0V 83824_LH-005- BD06-124.d
x10^6
421.2707 ([C23H36N2O5]+H)+
443.2527 ([C23H36N2O5]+Na)+
459.2260 ([C23H36N2O5]+K)+
Counts vs. Mass-to-Charge (m/z)

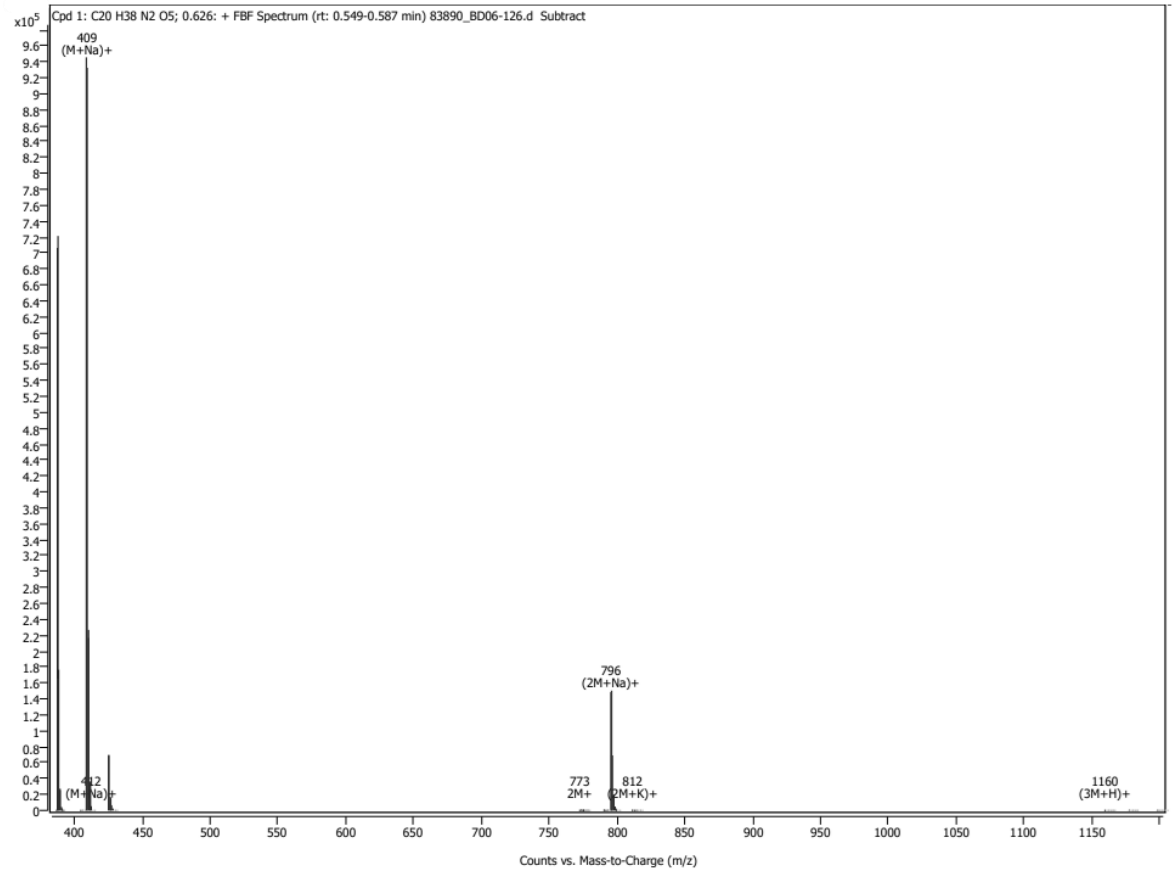

Cpd 1: C23 H36 N2 O5; 0.520: + FBF Spectrum (rt: 0.520 min) 83824_LH-005- BD06-124.d Subtract
x10^6
421.2707 ([C23H36N2O5]+H)+
422.2736 ([C23H36N2O5]+H)+
423.2757 ([C23H36N2O5]+H)+
424.2769 ([C23H36N2O5]+H)+
425.2856 ([C23H36N2O5]+H)+
Counts vs. Mass-to-Charge (m/z)

Boc-VI-O^{t}Bu **C23**

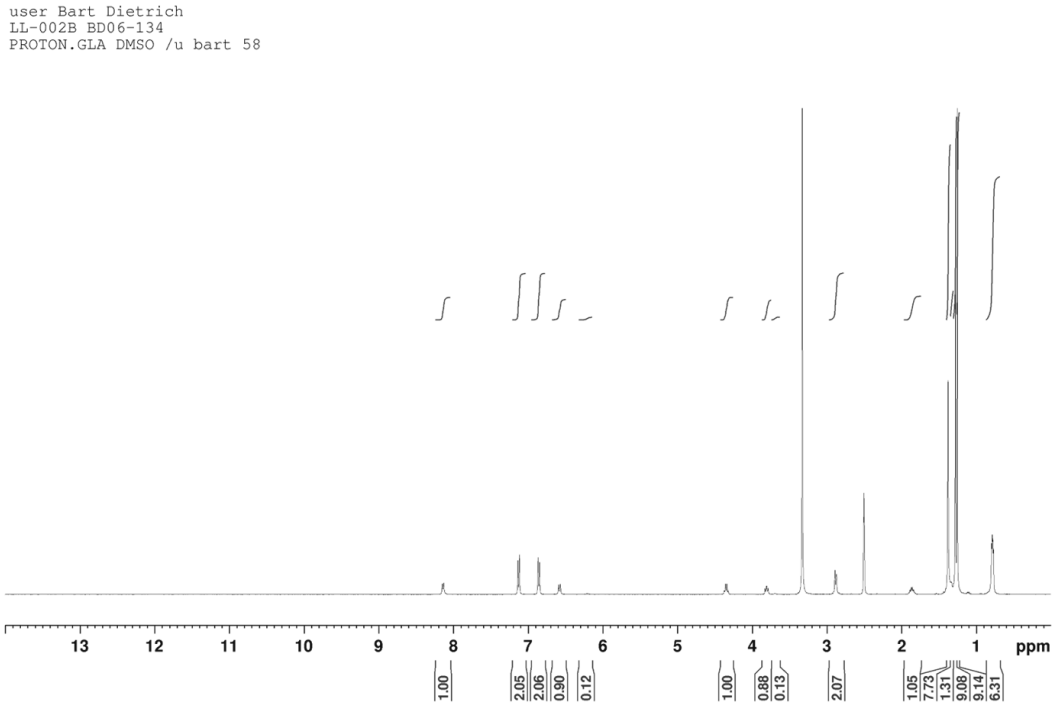


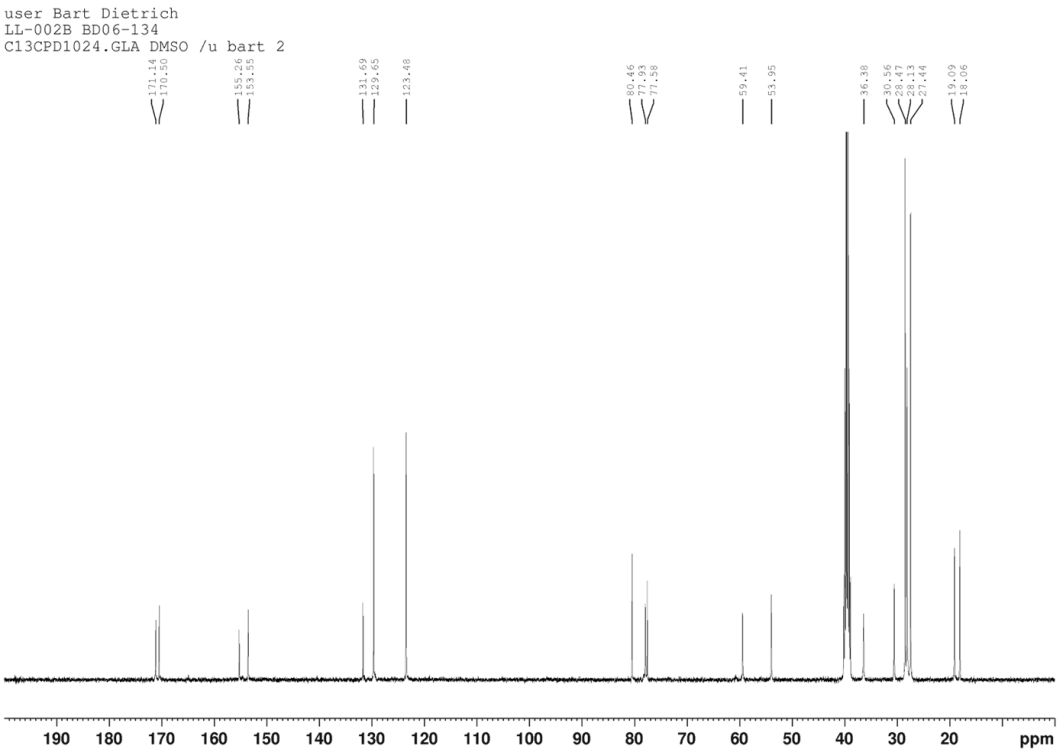


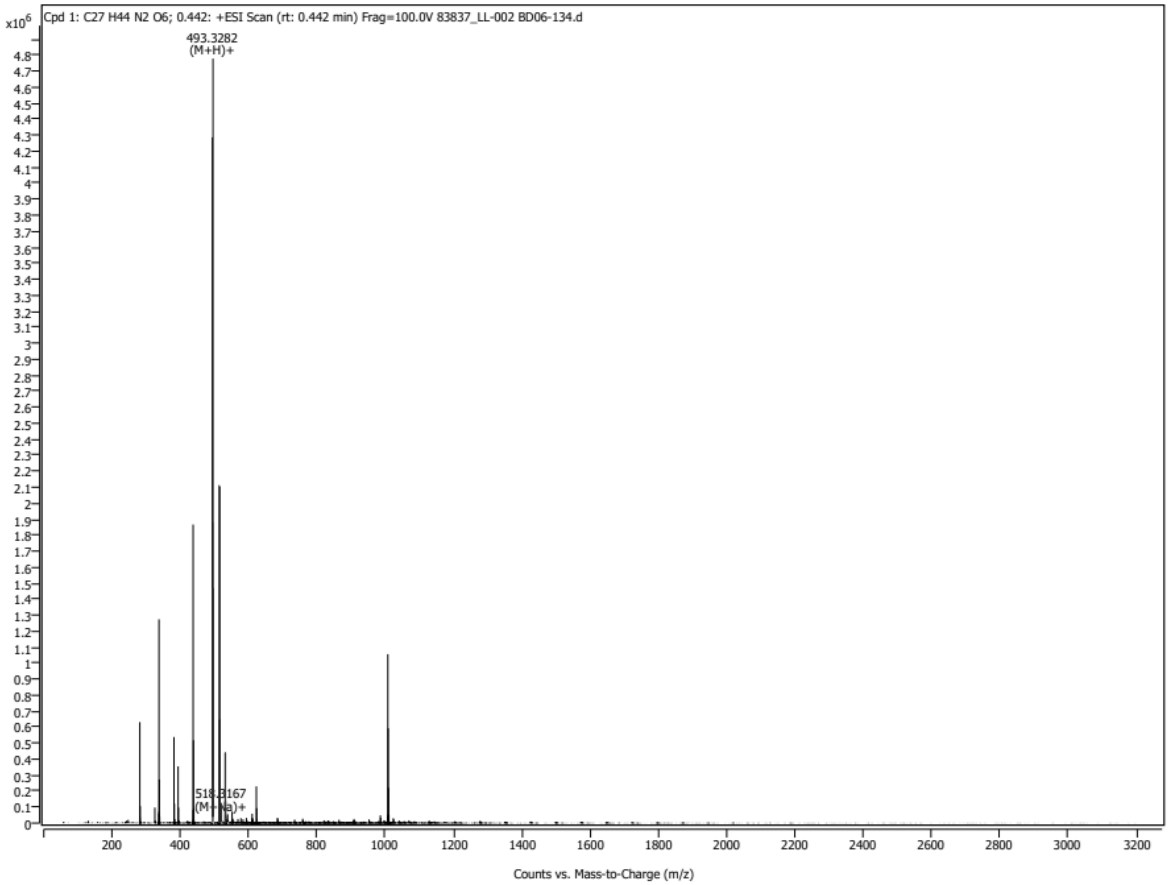

Cpd 1: C20 H38 N2 O5; 0.626: +ESI Scan (rt: 0.549-0.587 min, 8 scans) Frag=100.0V 83890_BD06-126.d
x10^5
409
(M+Na)+
796
(2M+Na)+
386
M+
773
2M+
1160
(3M+H)+
Counts vs. Mass-to-Charge (m/z)

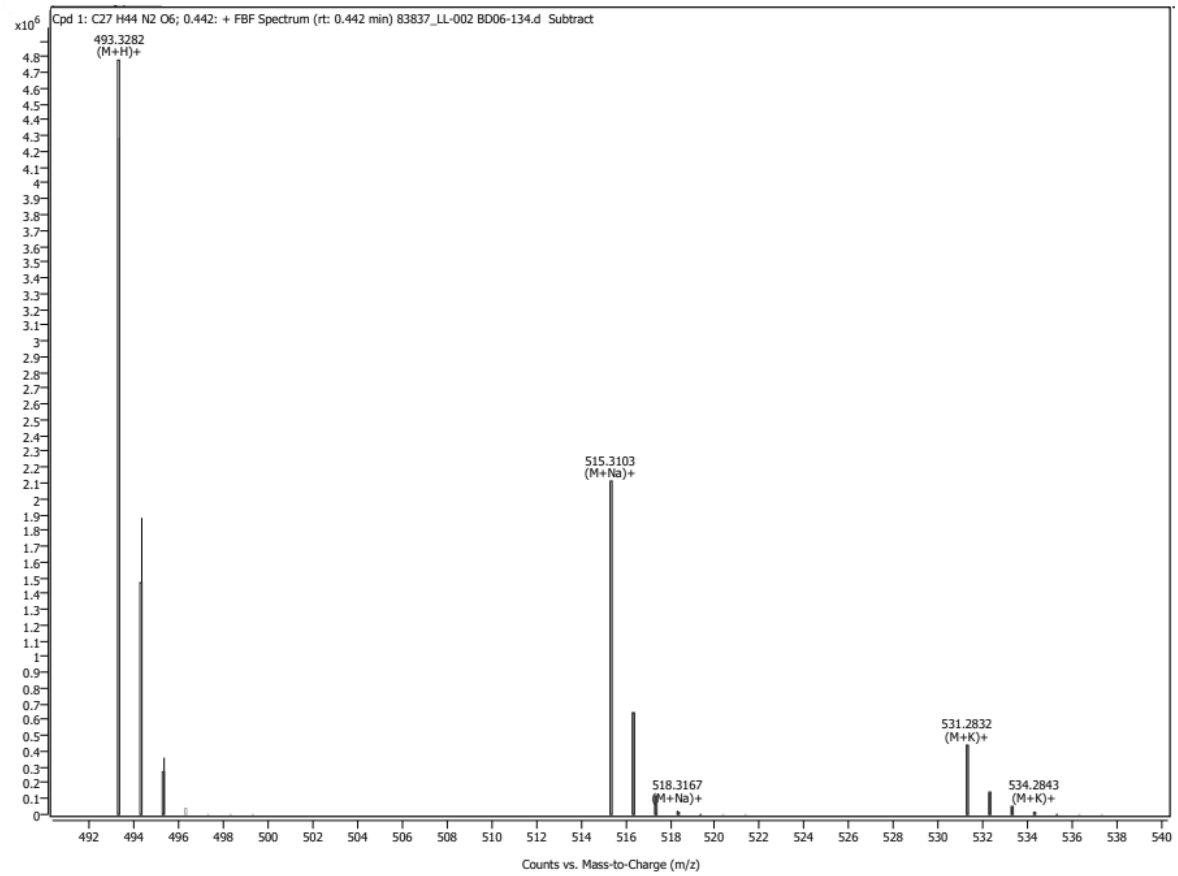

Cpd 1: C20 H38 N2 O5; 0.626: + FBF Spectrum (rt: 0.549-0.587 min) 83890_BD06-126.d Subtract
x10^5
409
(M+Na)+
796
(2M+Na)+
412
(M+Na)+
773
2M+
812
(2M+K)+
1160
(3M+H)+
Counts vs. Mass-to-Charge (m/z)

Boc-VY-O^{t}Bu **C24**

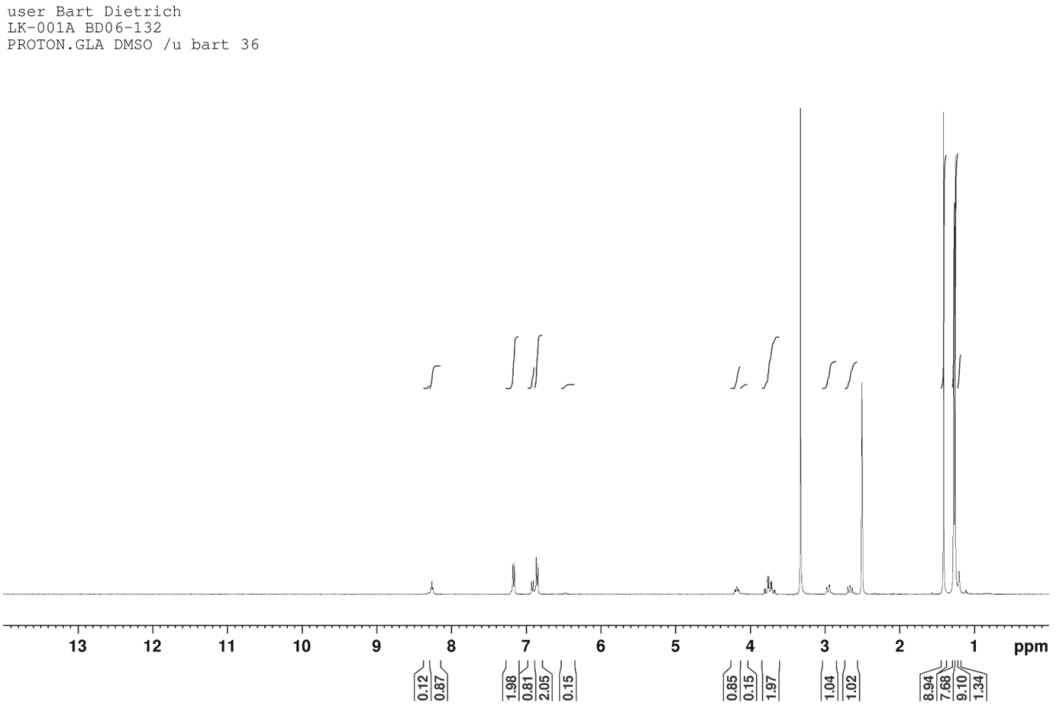


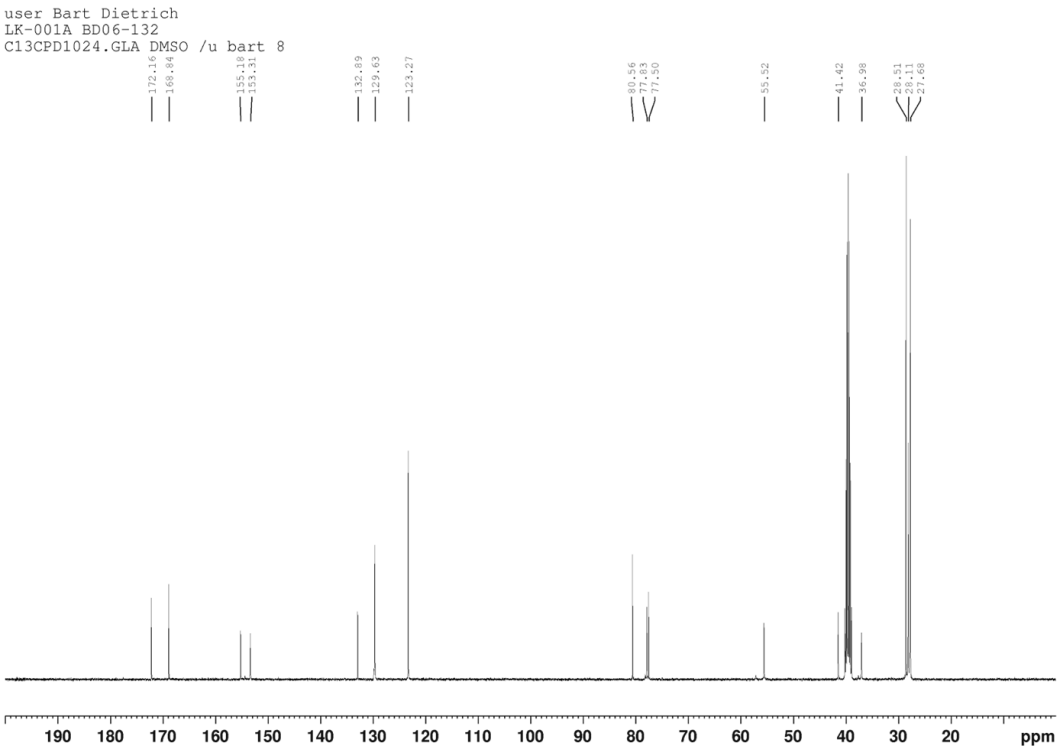


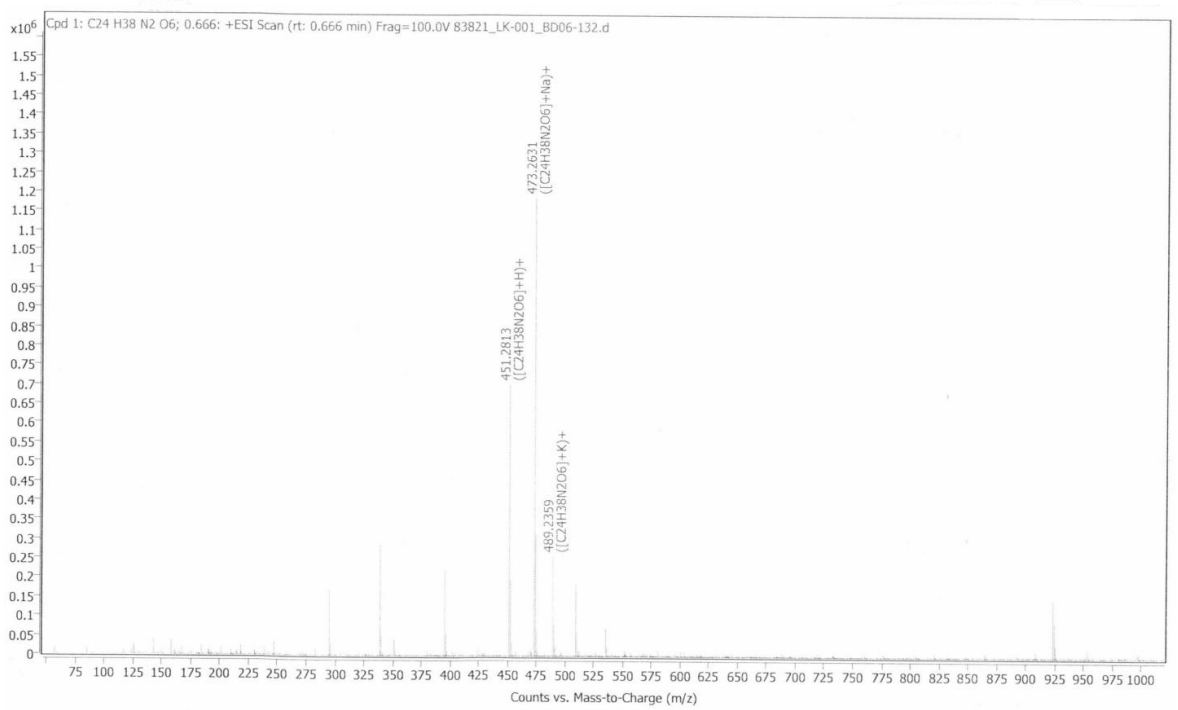

Cpd 1: C27 H44 N2 O6; 0.442: +ESI Scan (rt: 0.442 min) Frag=100.0V 83837_LL-002 BD06-134.d
x10^6
493.3282
(M+H)+
518.3167
(M+Na)+
Counts vs. Mass-to-Charge (m/z)

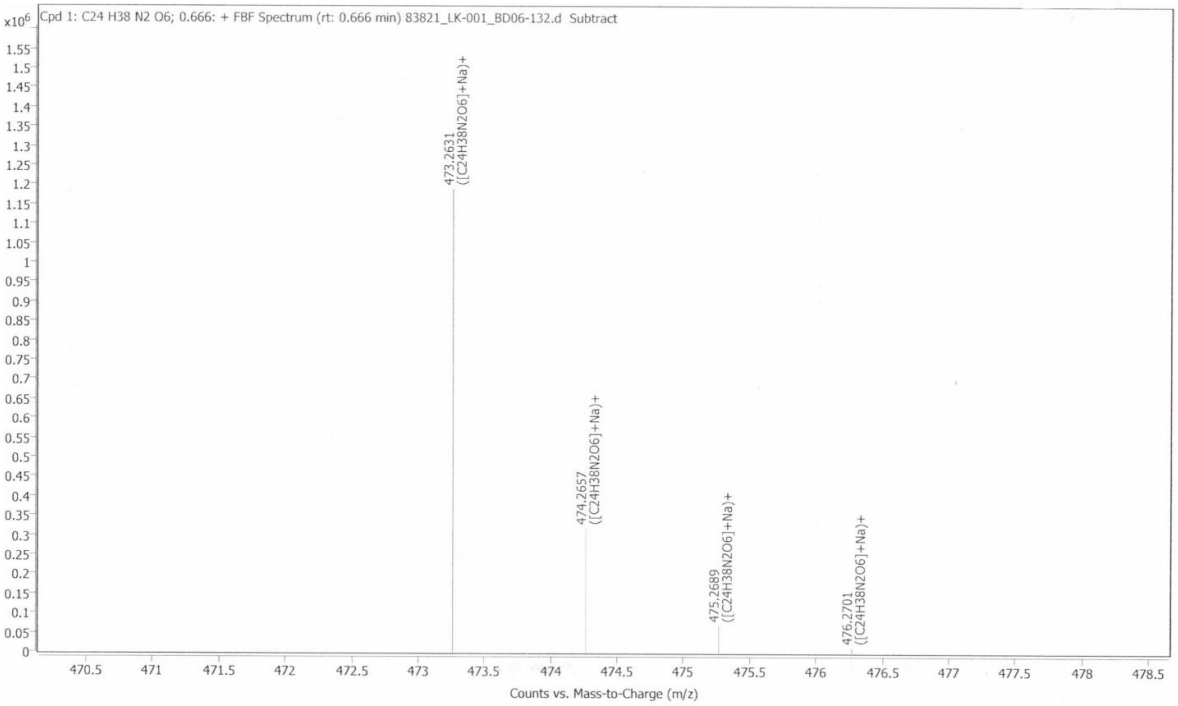

Cpd 1: C27 H44 N2 O6; 0.442: + FBF Spectrum (rt: 0.442 min) 83837_LL-002 BD06-134.d Subtract
x10^6
493.3282
(M+H)+
515.3103
(M+Na)+
518.3167
(M+Na)+
531.2832
(M+K)+
534.2843
(M+K)+
Counts vs. Mass-to-Charge (m/z)

Boc-YG-O^{t}Bu **C25**

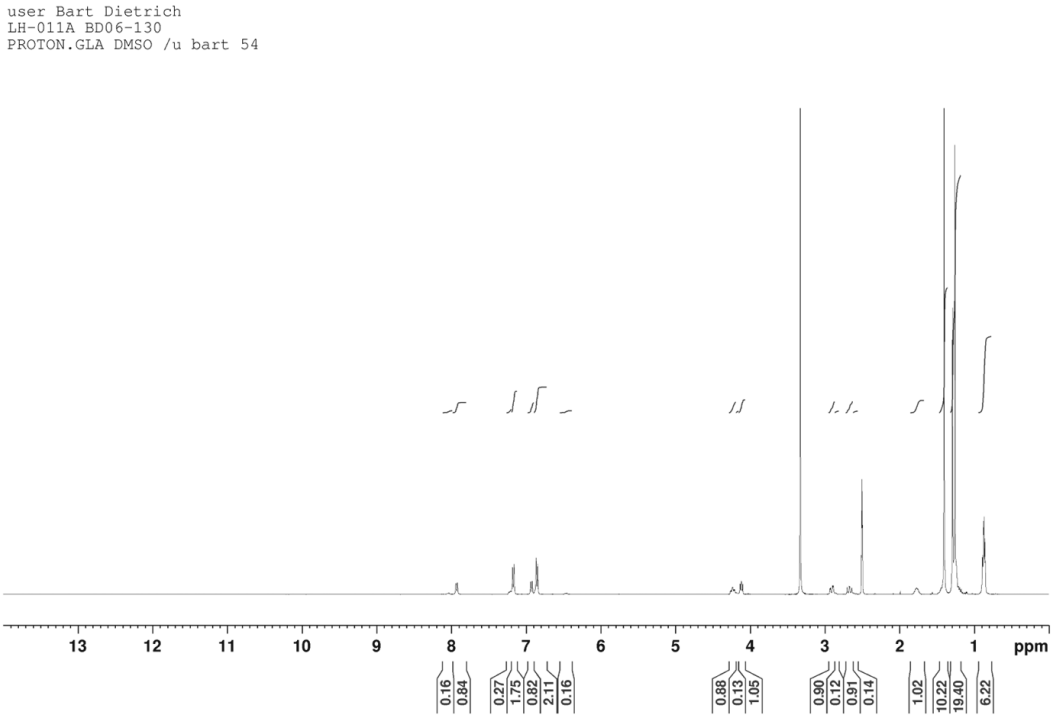


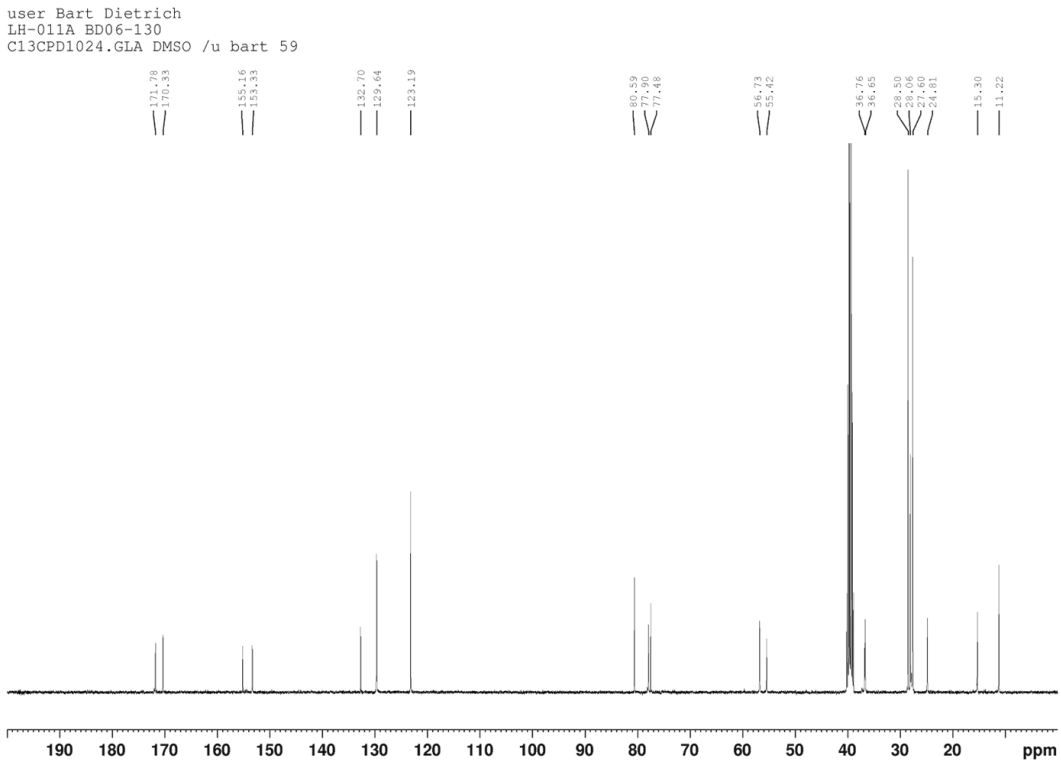


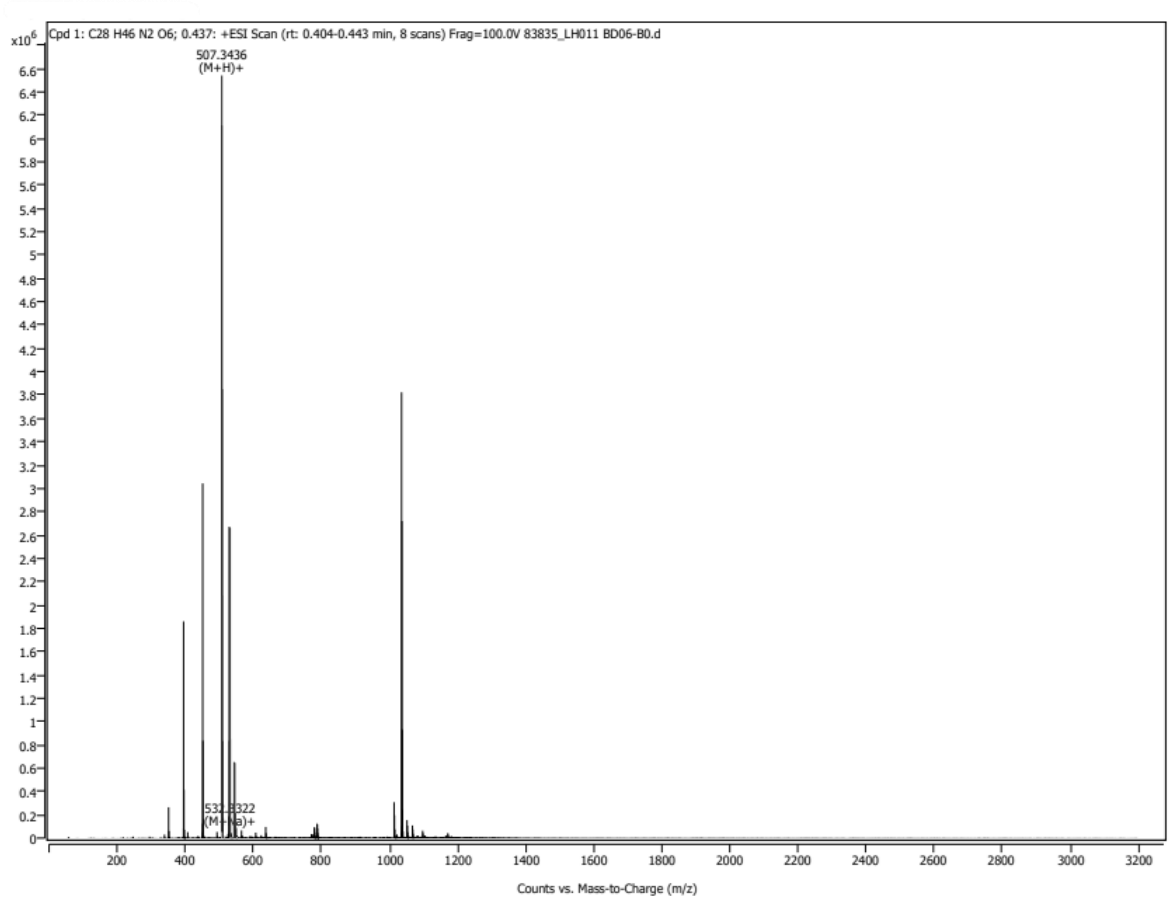

Cpd 1: C24 H38 N2 O6; 0.666: +ESI Scan (rt: 0.666 min) Frag=100.0V 83821_LK-001_BD06-132.d
x10^6
473.2631
([C24H38N2O6]+Na)+
451.2813
([C24H38N2O6]+H)+
489.2359
([C24H38N2O6]+K)+
Counts vs. Mass-to-Charge (m/z)

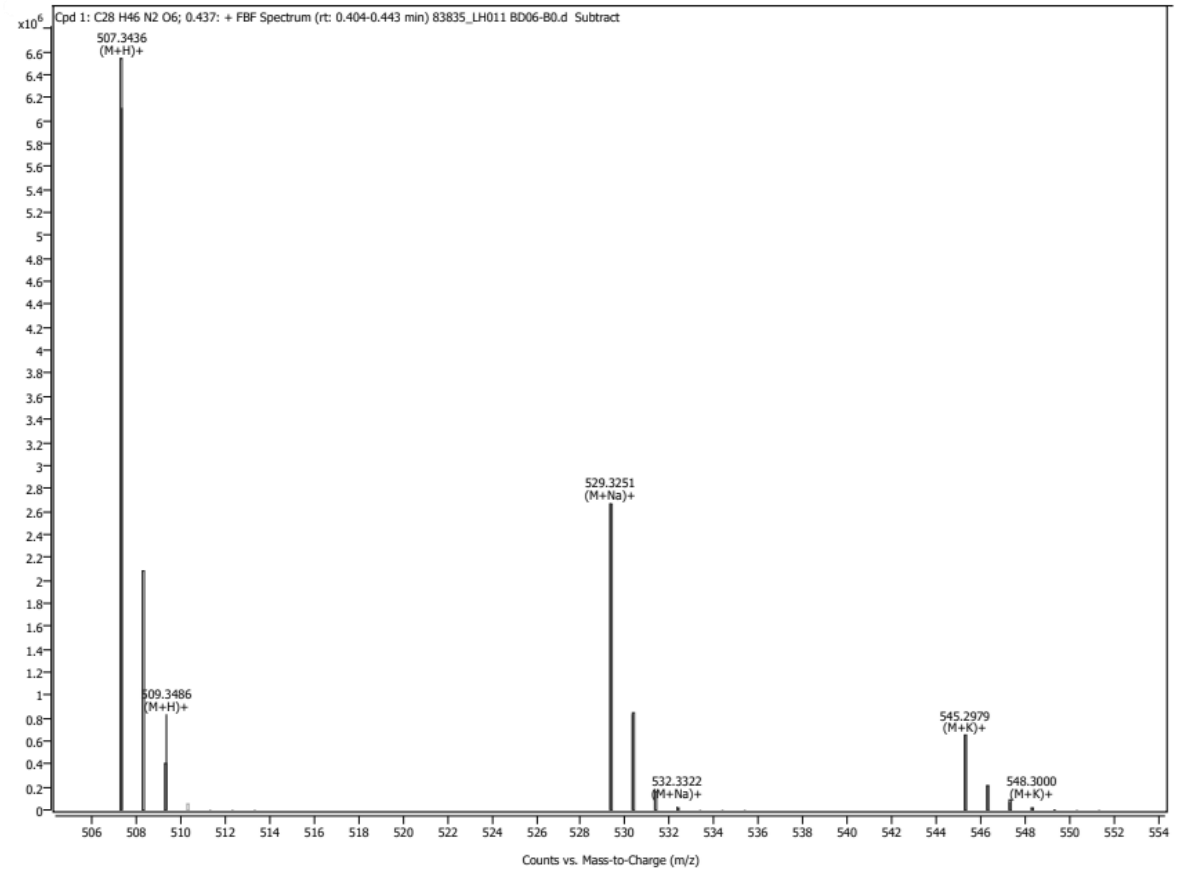

Cpd 1: C24 H38 N2 O6; 0.666: + FBF Spectrum (rt: 0.666 min) 83821_LK-001_BD06-132.d Subtract
x10^6
473.2631
([C24H38N2O6]+Na)+
474.2657
([C24H38N2O6]+Na)+
475.2689
([C24H38N2O6]+Na)+
476.2701
([C24H38N2O6]+Na)+
Counts vs. Mass-to-Charge (m/z)

Boc-YI-O^{t}Bu **C26**



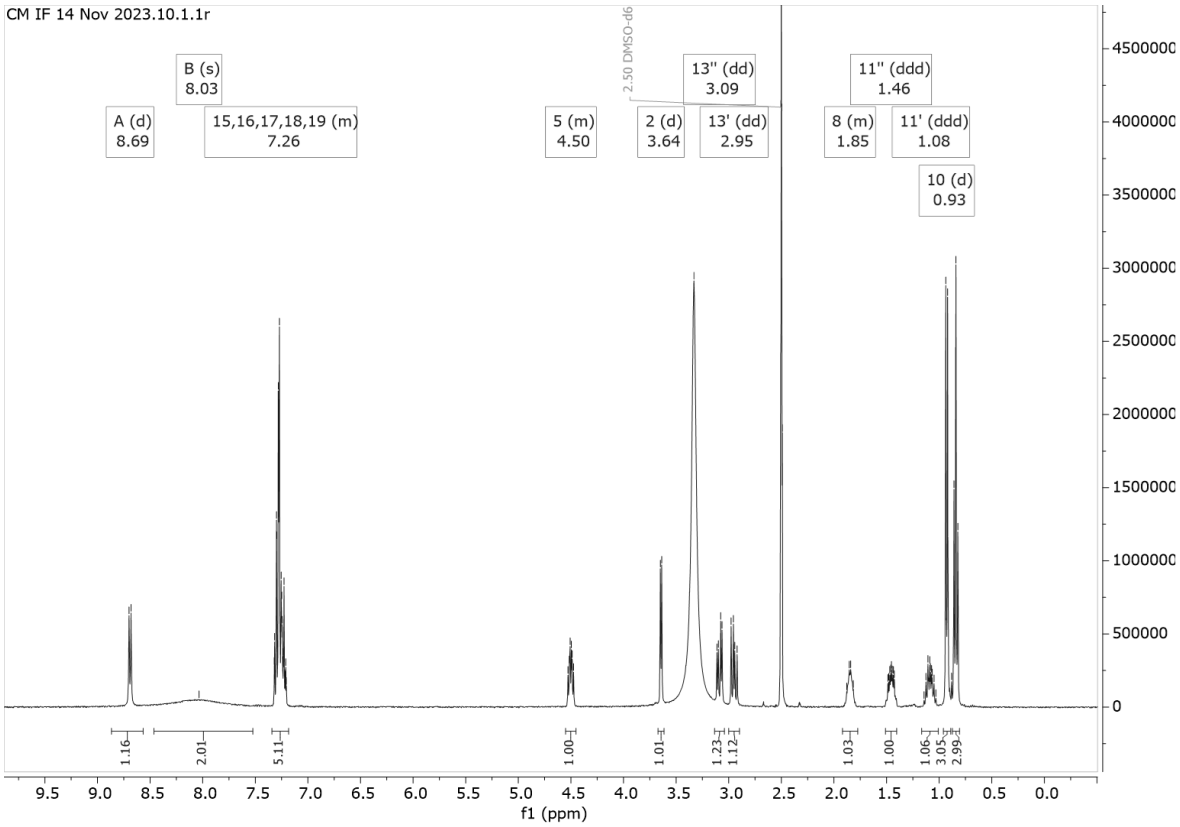


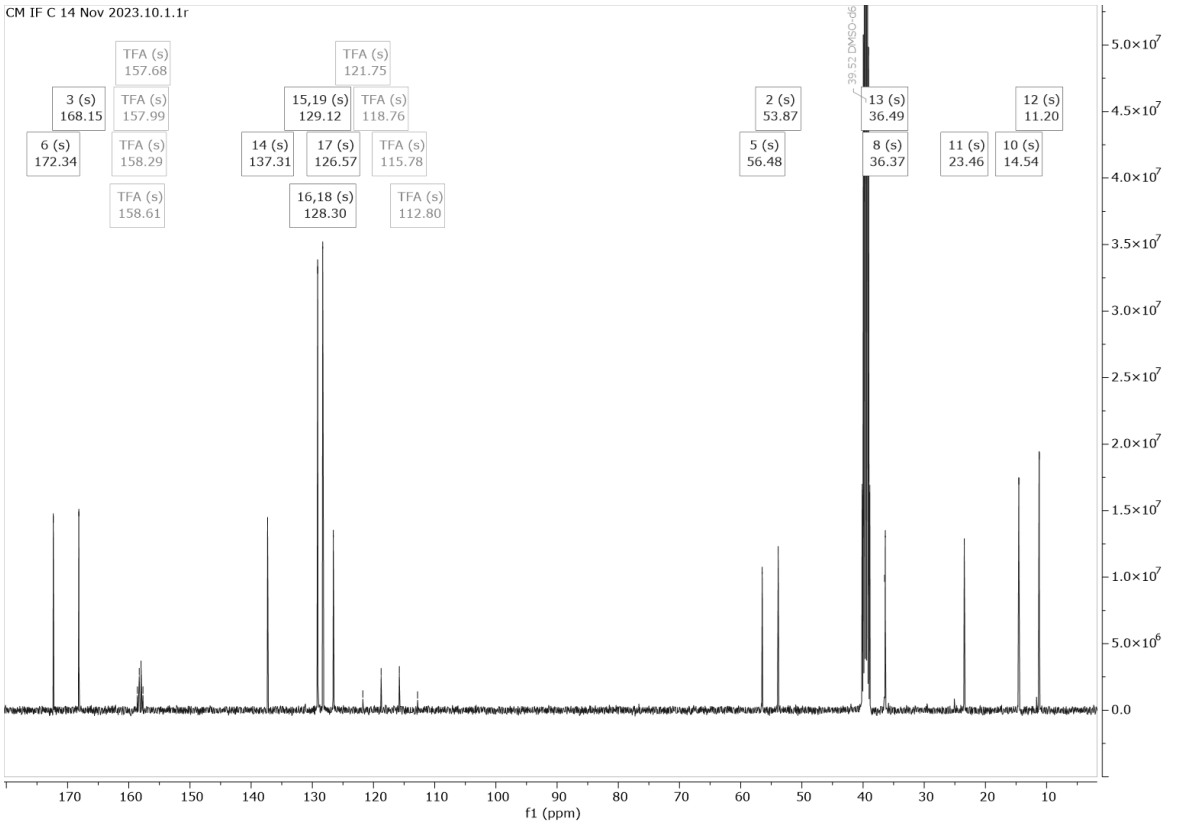

Cpd 1: C28 H46 N2 O6; 0.437: +ESI Scan (rt: 0.404-0.443 min, 8 scans) Frag=100.0V 83835_LH011 BD06-B0.d
x10^6
507.3436
(M+H)+
532.3322
(M+Na)+
Counts vs. Mass-to-Charge (m/z)

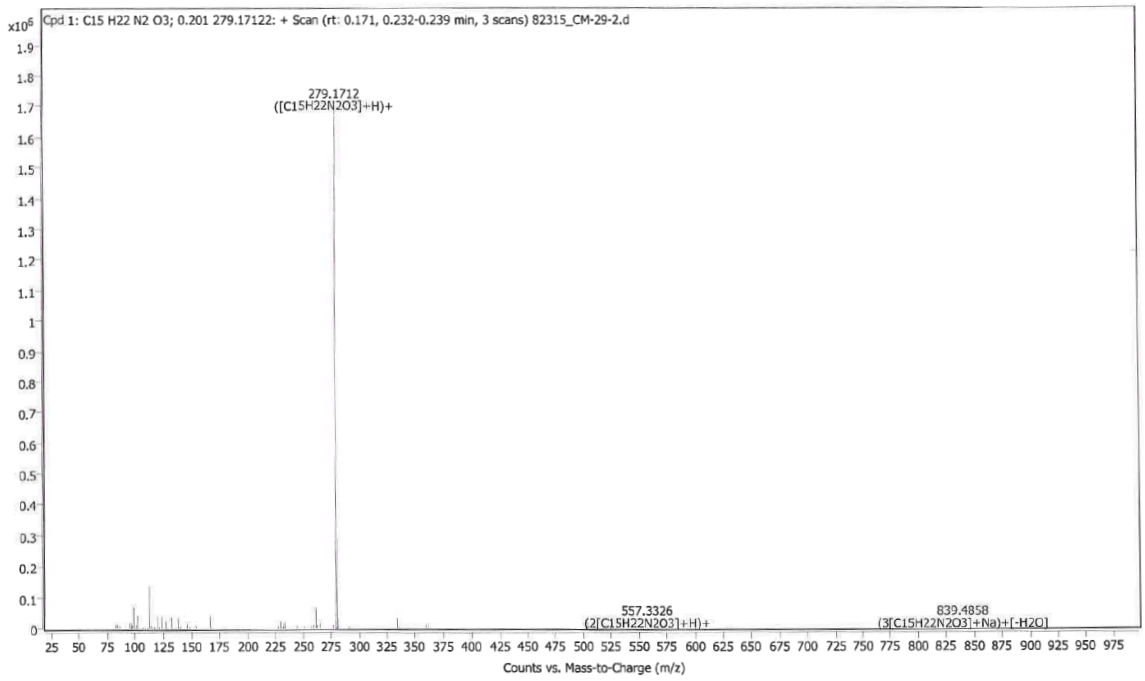

Cpd 1: C28 H46 N2 O6; 0.437: + FBF Spectrum (rt: 0.404-0.443 min) 83835_LH011 BD06-B0.d Subtract
x10^6
507.3436
(M+H)+
509.3486
(M+H)+
529.3251
(M+Na)+
532.3322
(M+Na)+
545.2979
(M+K)+
548.3000
(M+K)+
Counts vs. Mass-to-Charge (m/z)

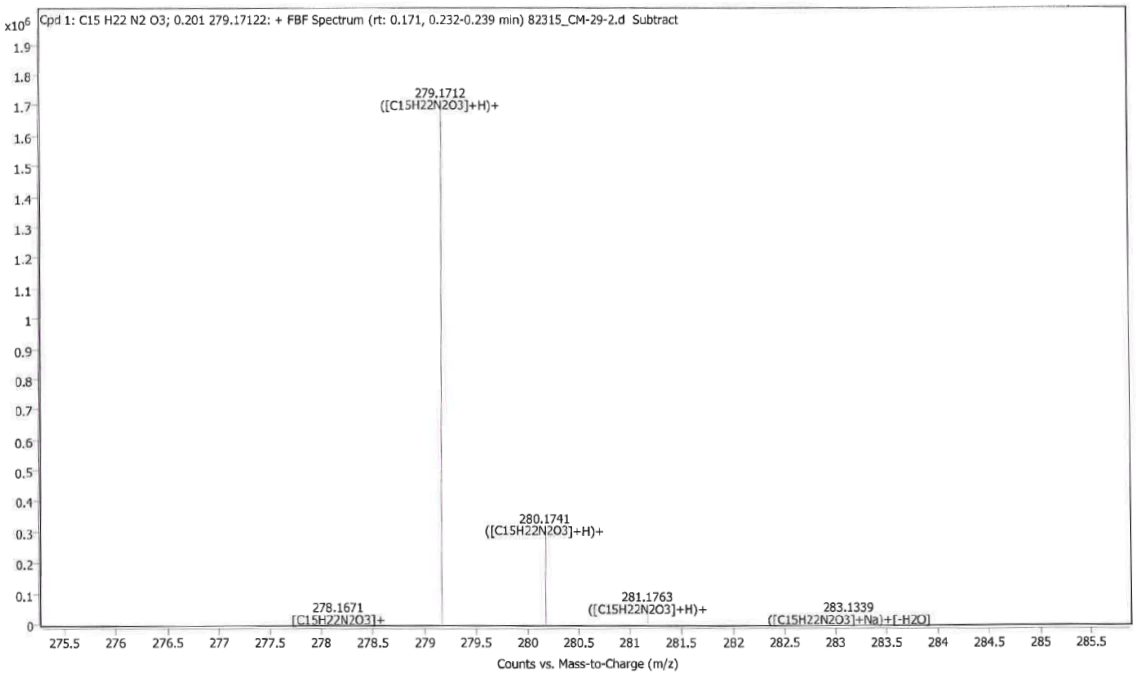
A 3.2.2 Deprotected Dipeptide Products

IF·TFA **D1**

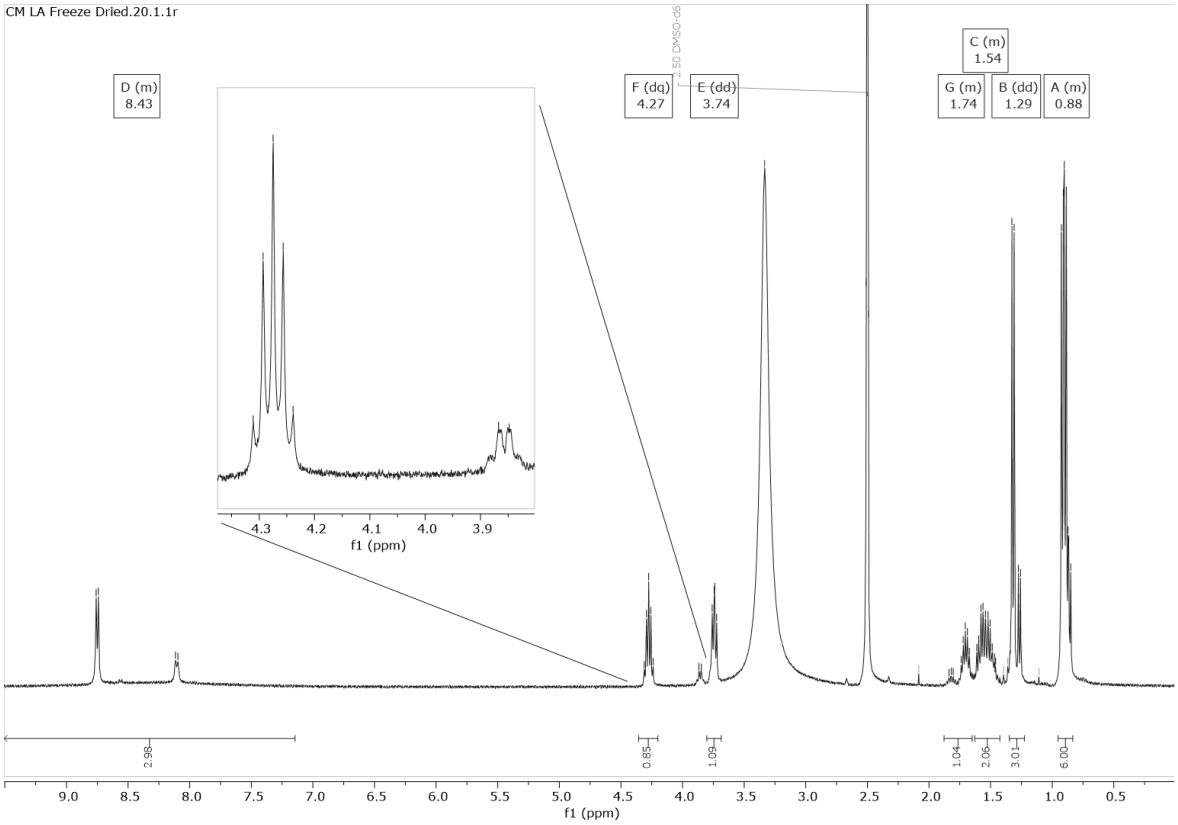


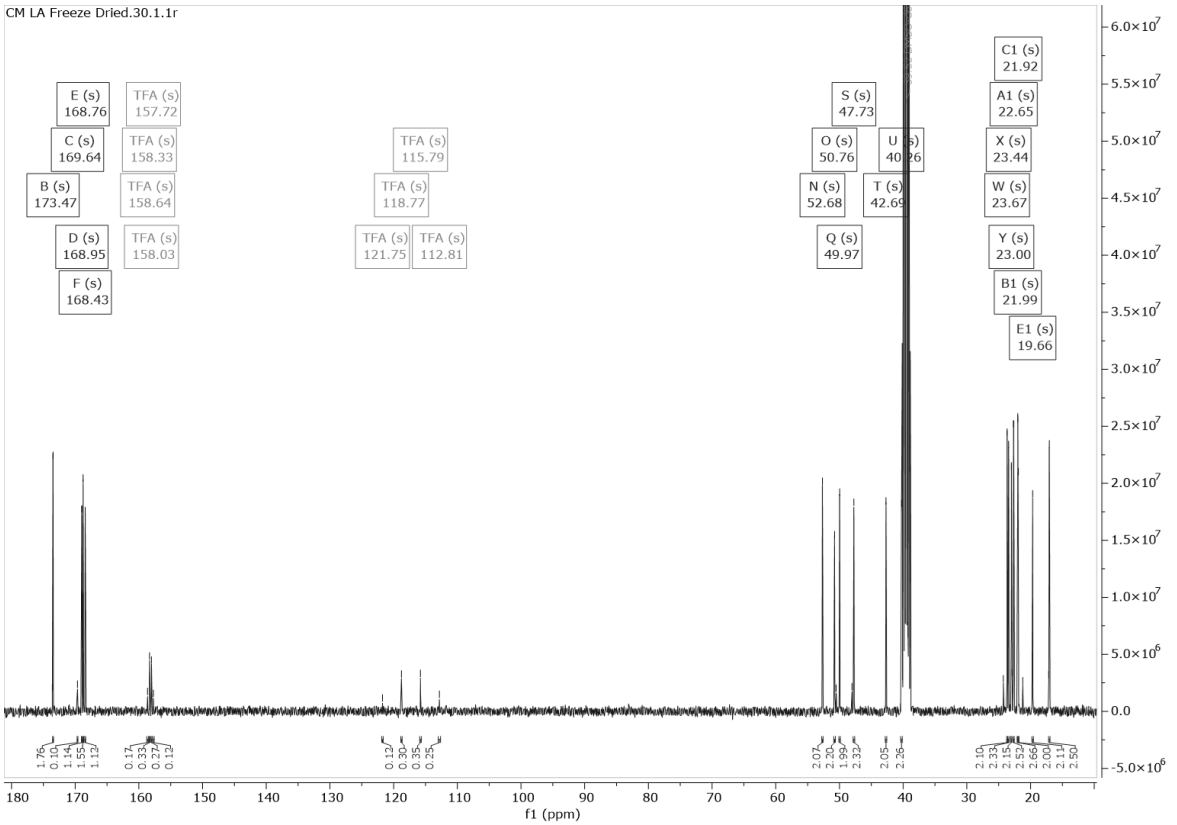


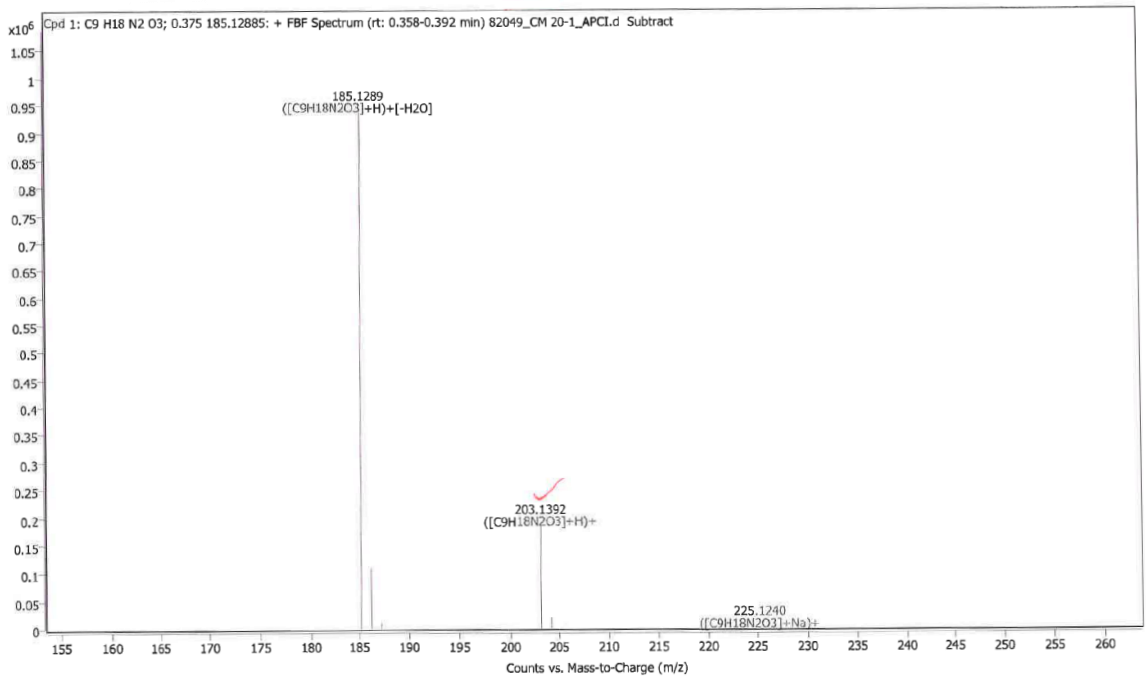

Cpd 1: C15 H22 N2 O3; 0.201 279.17122: + Scan (rt: 0.171, 0.232-0.239 min, 3 scans) 82315_CM-29-2.d
x10^6
279.1712
([C15H22N2O3]+H)+
557.3326
(2[C15H22N2O3]+H)+
839.4858
(3[C15H22N2O3]+Na)+[-H2O]
Counts vs. Mass-to-Charge (m/z)

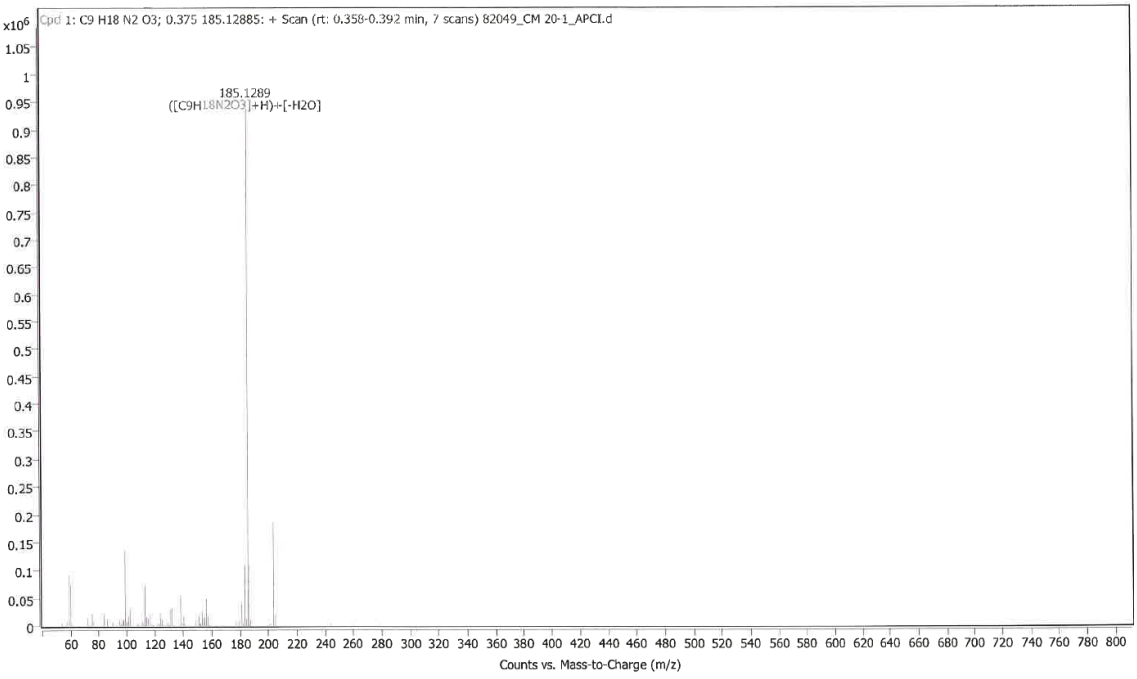

Cpd 1: C15 H22 N2 O3; 0.201 279.17122: + FBF Spectrum (rt: 0.171, 0.232-0.239 min) 82315_CM-29-2.d Subtract
x10^6
279.1712
([C15H22N2O3]+H)+
280.1741
([C15H22N2O3]+H)+
278.1671
[C15H22N2O3]+
281.1763
([C15H22N2O3]+H)+
283.1339
([C15H22N2O3]+Na)+[-H2O]
Counts vs. Mass-to-Charge (m/z)

LA·TFA **D2** (Cyclisation impurity present. Not fully characterised)

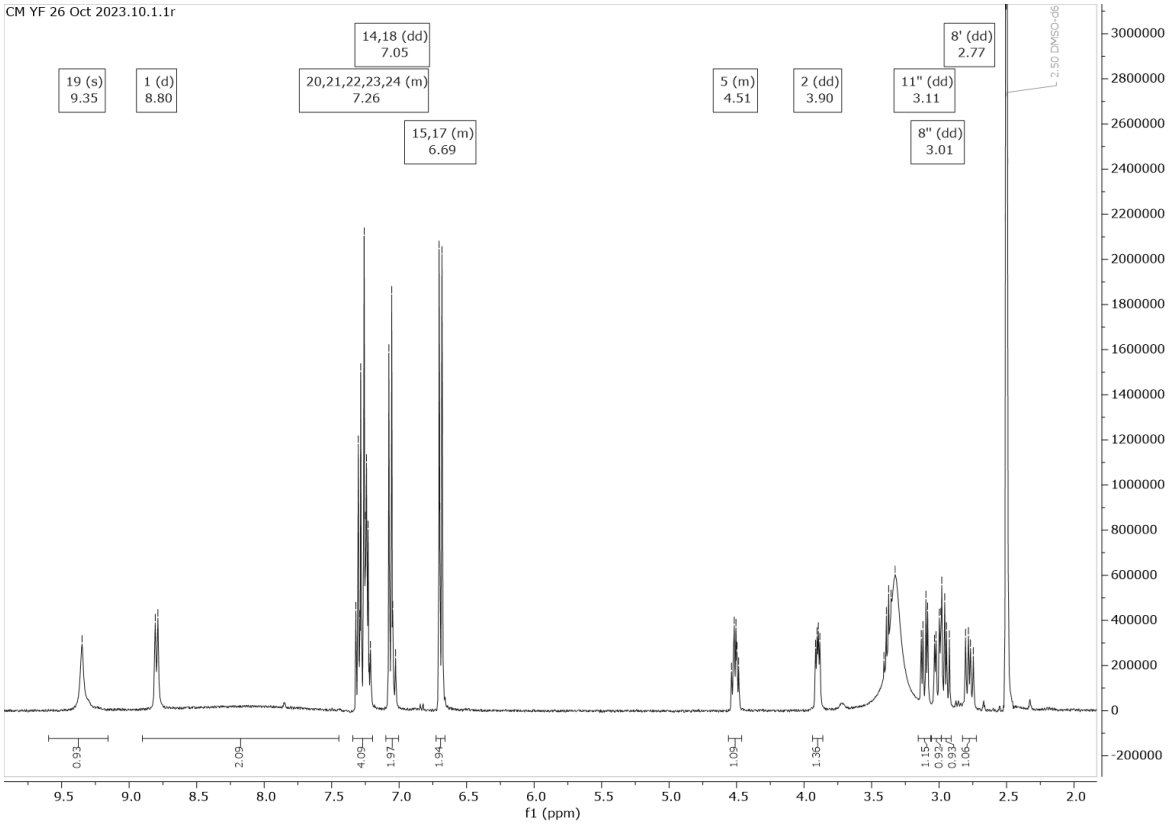


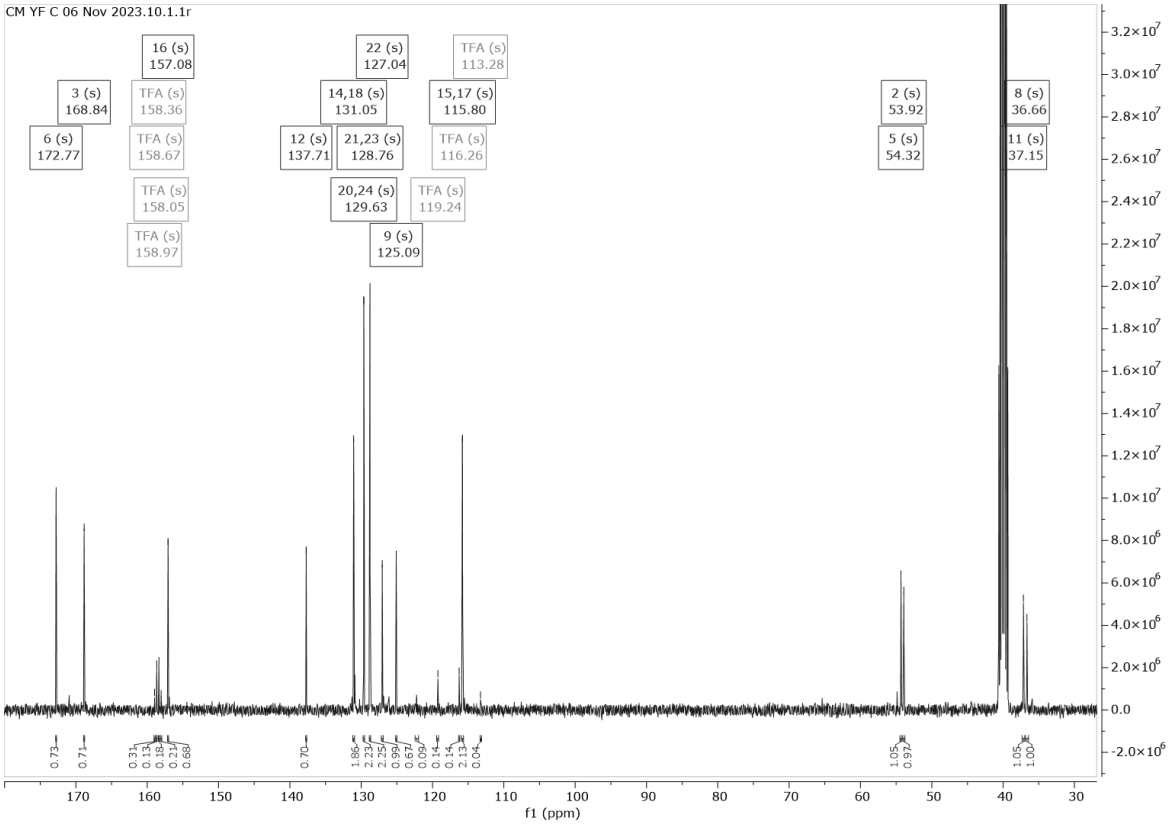


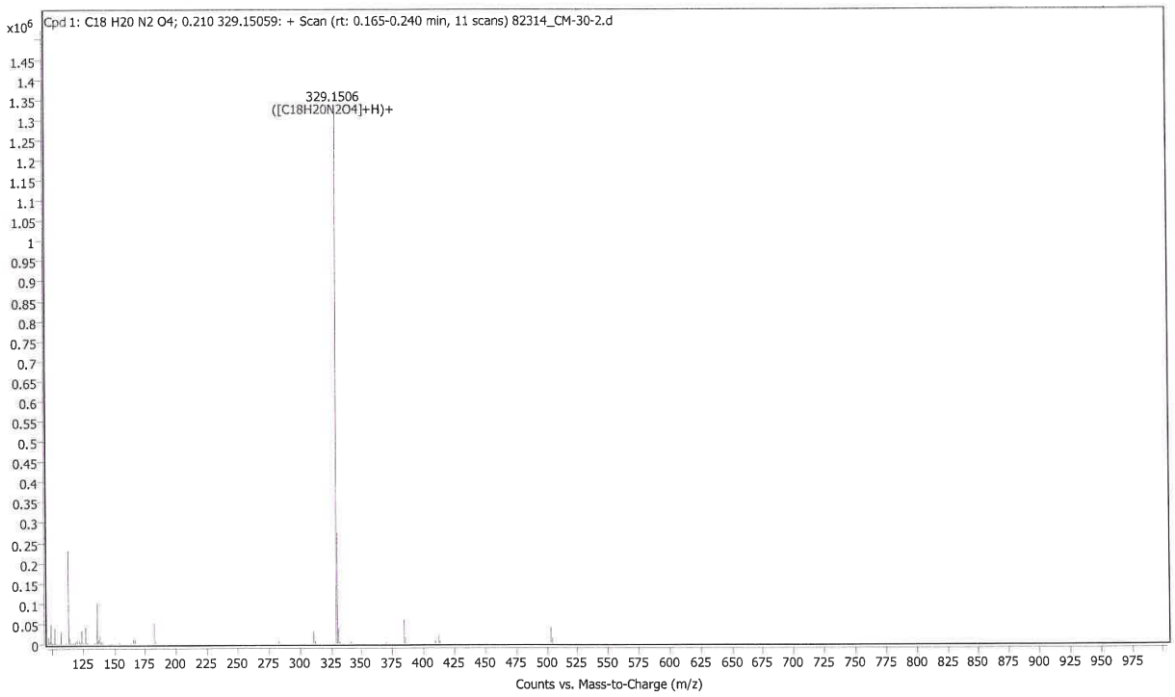

Cpd 1: C9 H18 N2 O3; 0.375 185.12885: + FBF Spectrum (rt: 0.358-0.392 min) 82049_CM 20-1_APCI.d Subtract
x10^6
185.1289
([C9H18N2O3]+H)+[-H2O]
203.1392
([C9H18N2O3]+H)+
225.1240
([C9H18N2O3]+Na)+
Counts vs. Mass-to-Charge (m/z)

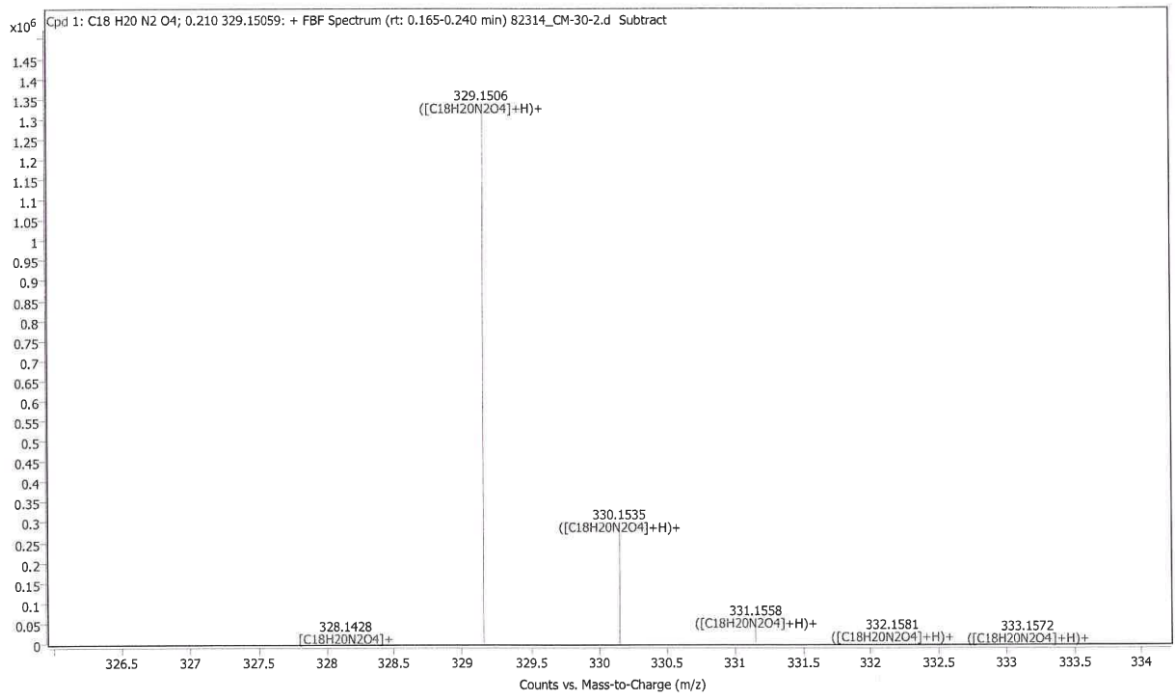

Cpd 1: C9 H18 N2 O3; 0.375 185.12885: + Scan (rt: 0.358-0.392 min, 7 scans) 82049_CM 20-1_APCI.d
x10^6
185.1289
([C9H18N2O3]+H)+[-H2O]
Counts vs. Mass-to-Charge (m/z)

YF·TFA **D3**

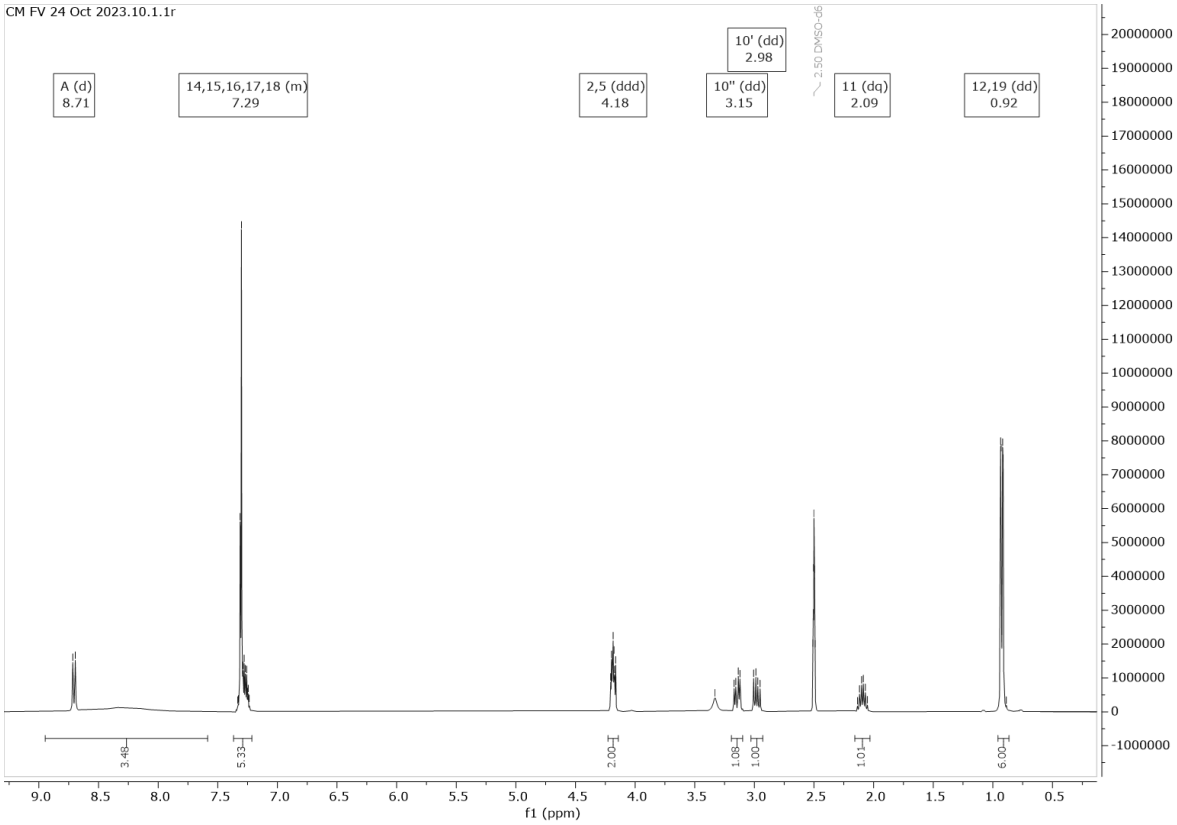

CM YF 26 Oct 2023.10.1.1r
19 (s) 9.35
1 (d) 8.80
14,18 (dd) 7.05
20,21,22,23,24 (m) 7.26
15,17 (m) 6.69
5 (m) 4.51
2 (dd) 3.90
11" (dd) 3.11
8" (dd) 3.01
8' (dd) 2.77
2.50 DMSO-d6
f1 (ppm)

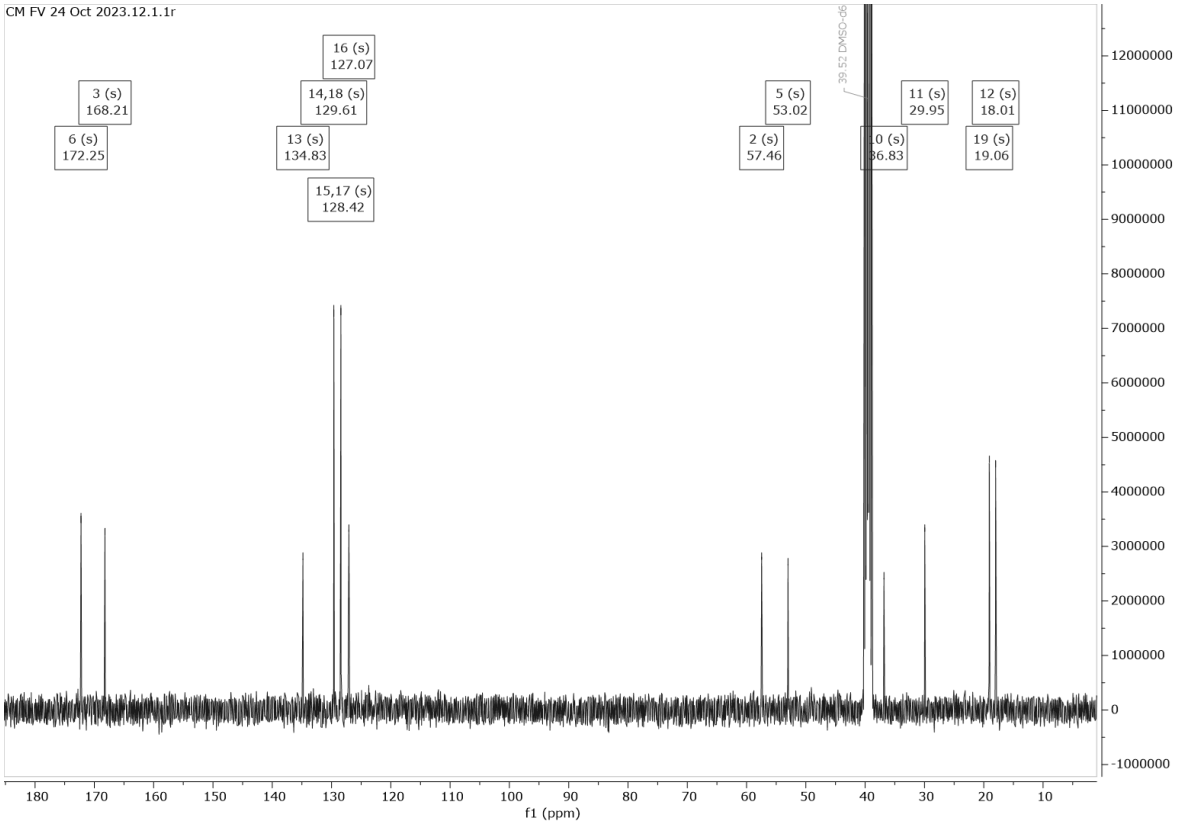

CM YF C 06 Nov 2023.10.1.1r
16 (s) 157.08
3 (s) 168.84
6 (s) 172.77
TFA (s) 158.36
TFA (s) 158.67
TFA (s) 158.05
TFA (s) 158.97
22 (s) 127.04
14,18 (s) 131.05
12 (s) 137.71
21,23 (s) 128.76
20,24 (s) 129.63
9 (s) 125.09
TFA (s) 113.28
15,17 (s) 115.80
TFA (s) 116.26
TFA (s) 119.24
2 (s) 53.92
5 (s) 54.32
8 (s) 36.66
11 (s) 37.15
f1 (ppm)

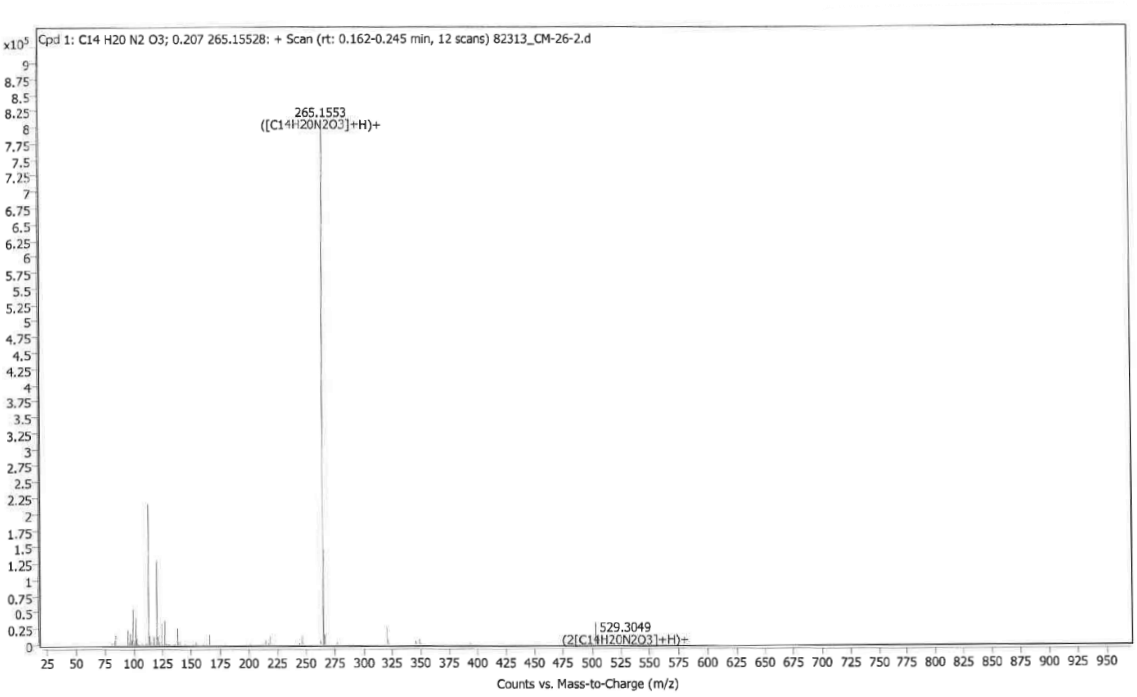

Cpd 1: C18 H20 N2 O4; 0.210 329.15059: + Scan (rt: 0.165-0.240 min, 11 scans) 82314_CM-30-2.d
x10^6
329.1506
([C18H20N2O4]+H)+
Counts vs. Mass-to-Charge (m/z)

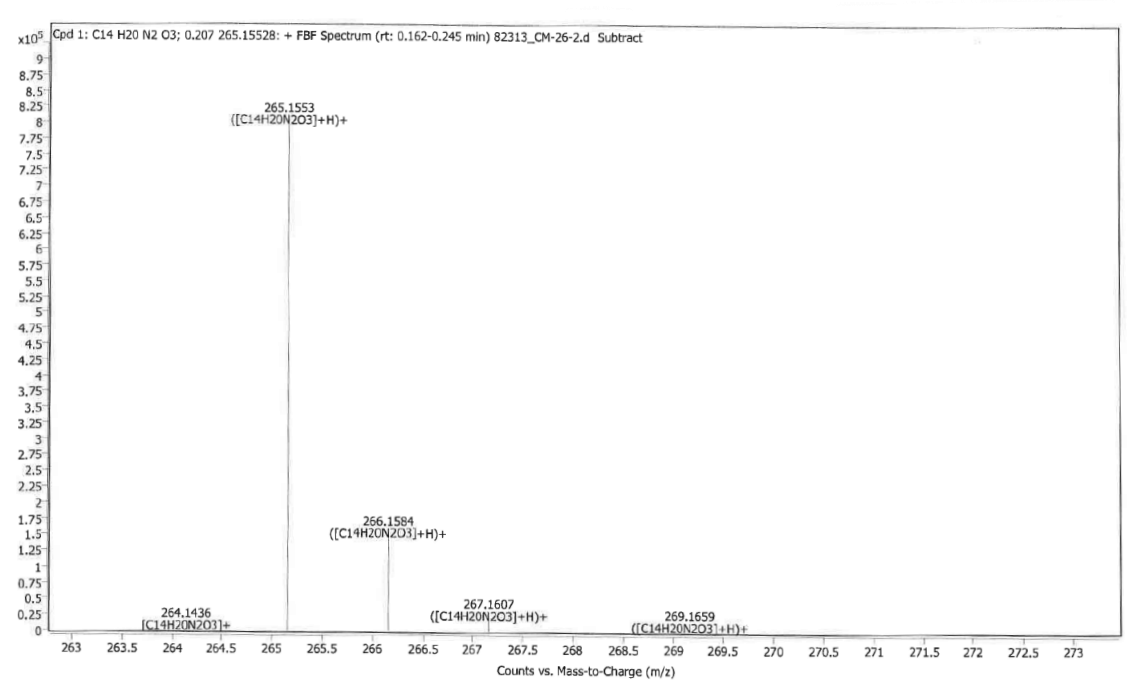

Cpd 1: C18 H20 N2 O4; 0.210 329.15059: + FBF Spectrum (rt: 0.165-0.240 min) 82314_CM-30-2.d Subtract
x10^6
329.1506
([C18H20N2O4]+H)+
330.1535
([C18H20N2O4]+H)+
328.1428
[C18H20N2O4]+
331.1558
([C18H20N2O4]+H)+
332.1581
([C18H20N2O4]+H)+
333.1572
([C18H20N2O4]+H)+
Counts vs. Mass-to-Charge (m/z)

FV·HCl **D4**

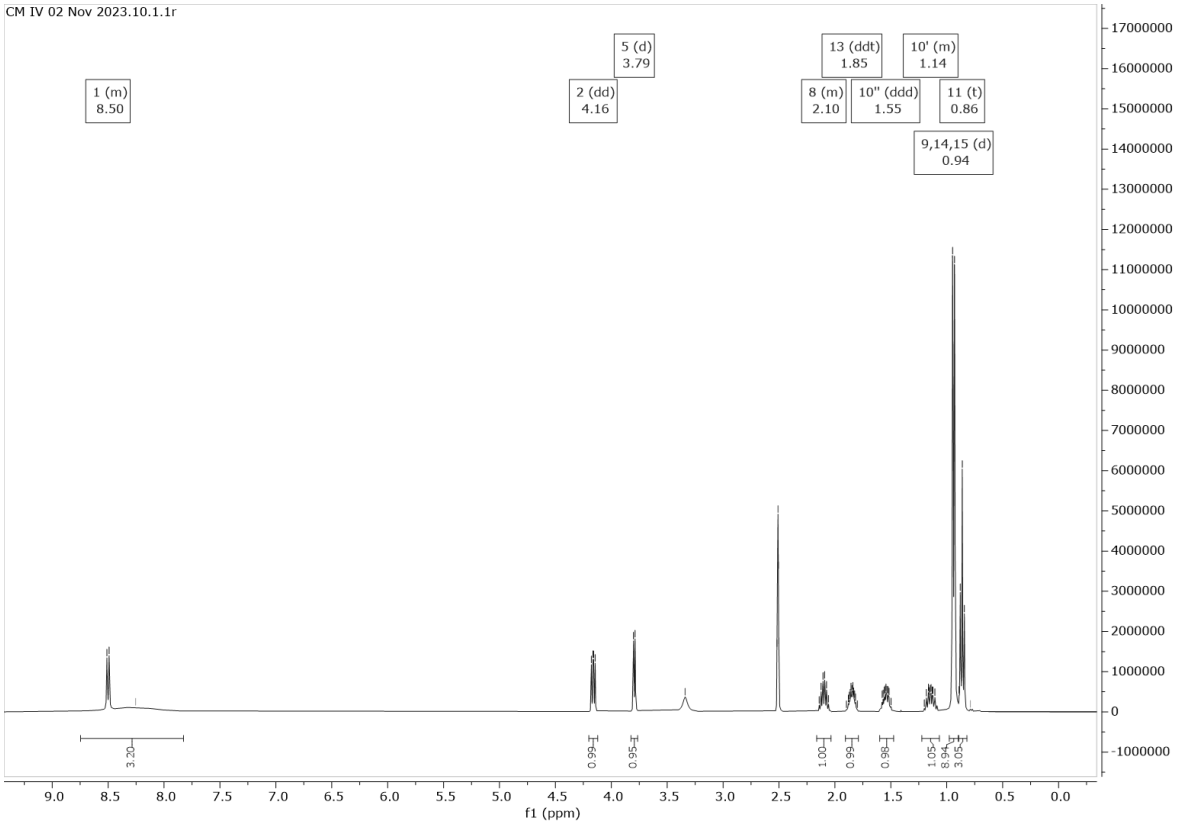

CM FV 24 Oct 2023.10.1.1r
A (d)
8.71
14,15,16,17,18 (m)
7.29
2,5 (ddd)
4.18
10' (dd)
2.98
10'' (dd)
3.15
11 (dq)
2.09
12,19 (dd)
0.92
2.50 DMSO-d6
3.48
5.33
2.00
1.08
1.00
1.01
6.00
f1 (ppm)

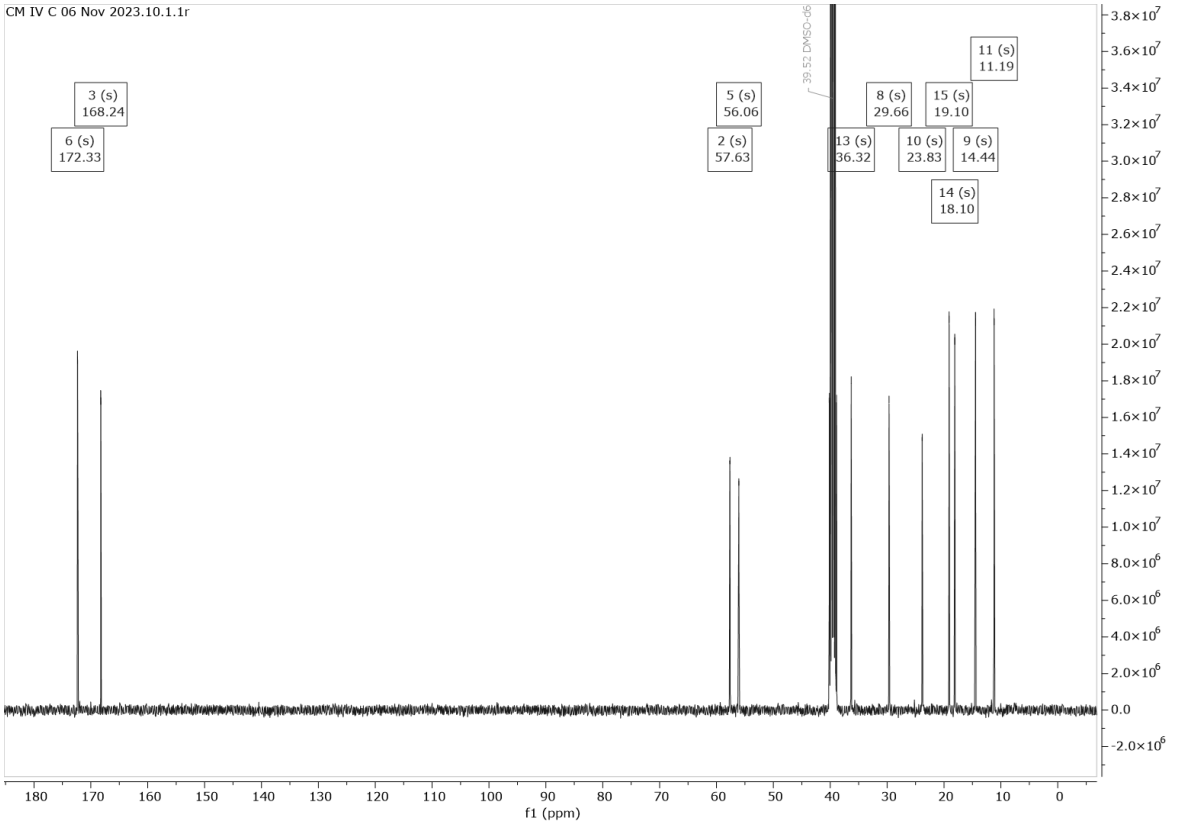

CM FV 24 Oct 2023.12.1.1r
16 (s)
127.07
3 (s)
168.21
14,18 (s)
129.61
6 (s)
172.25
13 (s)
134.83
15,17 (s)
128.42
5 (s)
53.02
11 (s)
29.95
12 (s)
18.01
2 (s)
57.46
10 (s)
36.83
19 (s)
19.06
39.52 DMSO-d6
f1 (ppm)

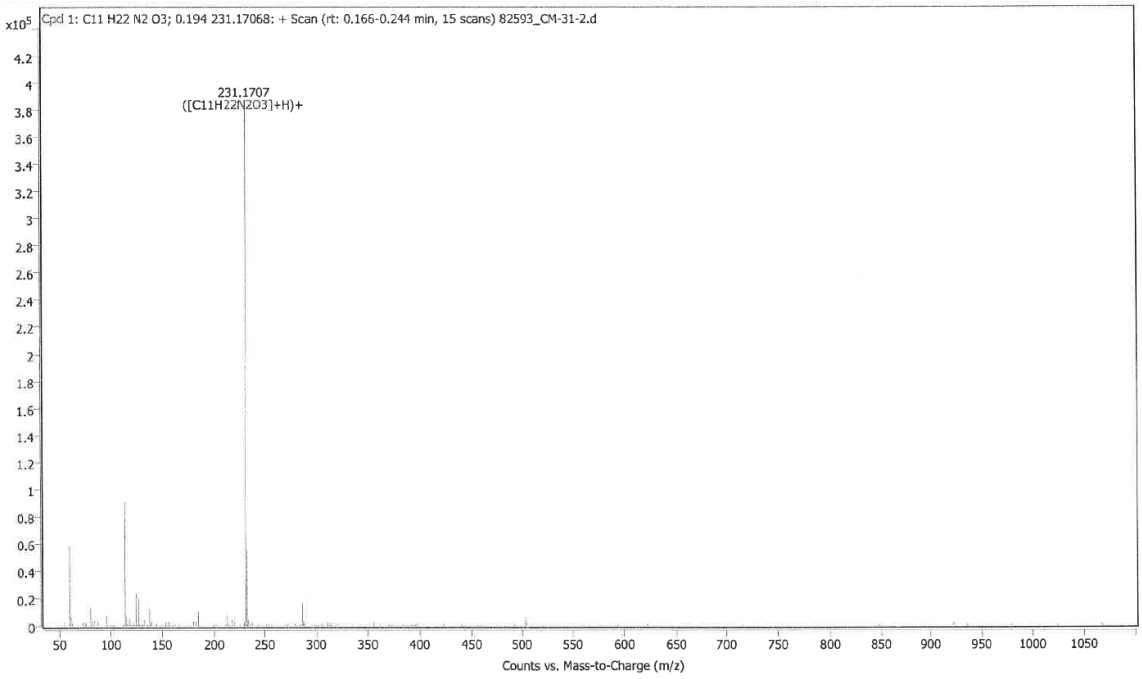

Cpd 1: C14 H20 N2 O3; 0.207 265.15528: + Scan (rt: 0.162-0.245 min, 12 scans) 82313_CM-26-2.d
x10^5
265.1553
([C14H20N2O3]+H)+
529.3049
(2[C14H20N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

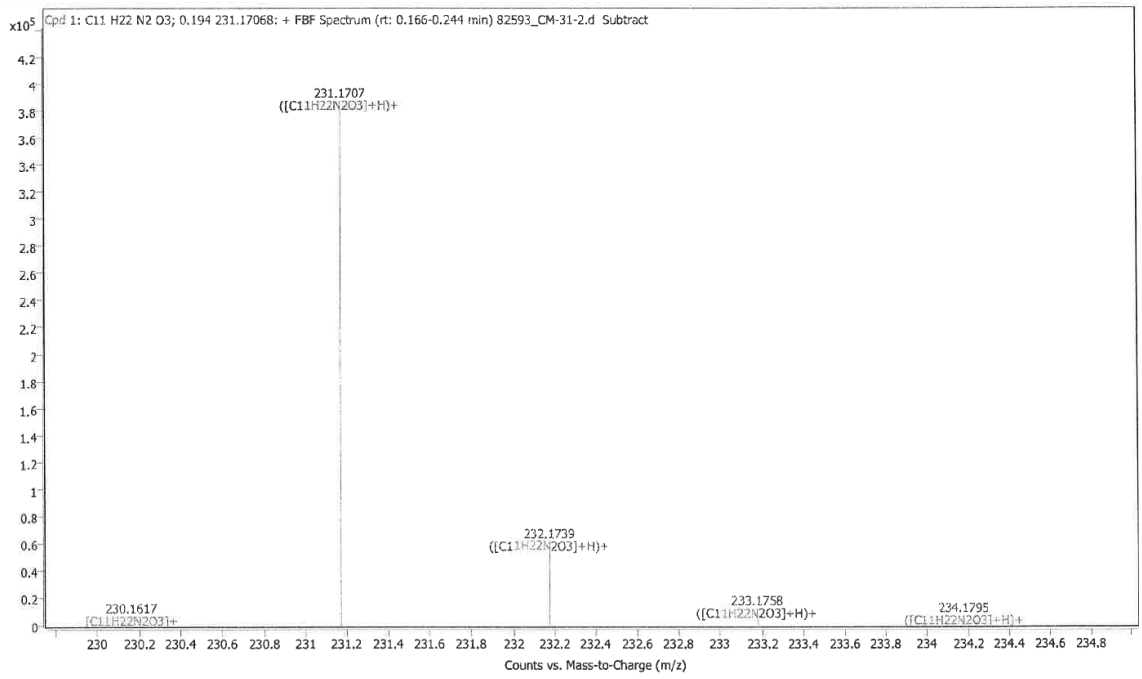

Cpd 1: C14 H20 N2 O3; 0.207 265.15528: + FBF Spectrum (rt: 0.162-0.245 min) 82313_CM-26-2.d Subtract
x10^5
265.1553
([C14H20N2O3]+H)+
266.1584
([C14H20N2O3]+H)+
264.1436
[C14H20N2O3]+
267.1607
([C14H20N2O3]+H)+
269.1659
([C14H20N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

IV·HCl **D5**

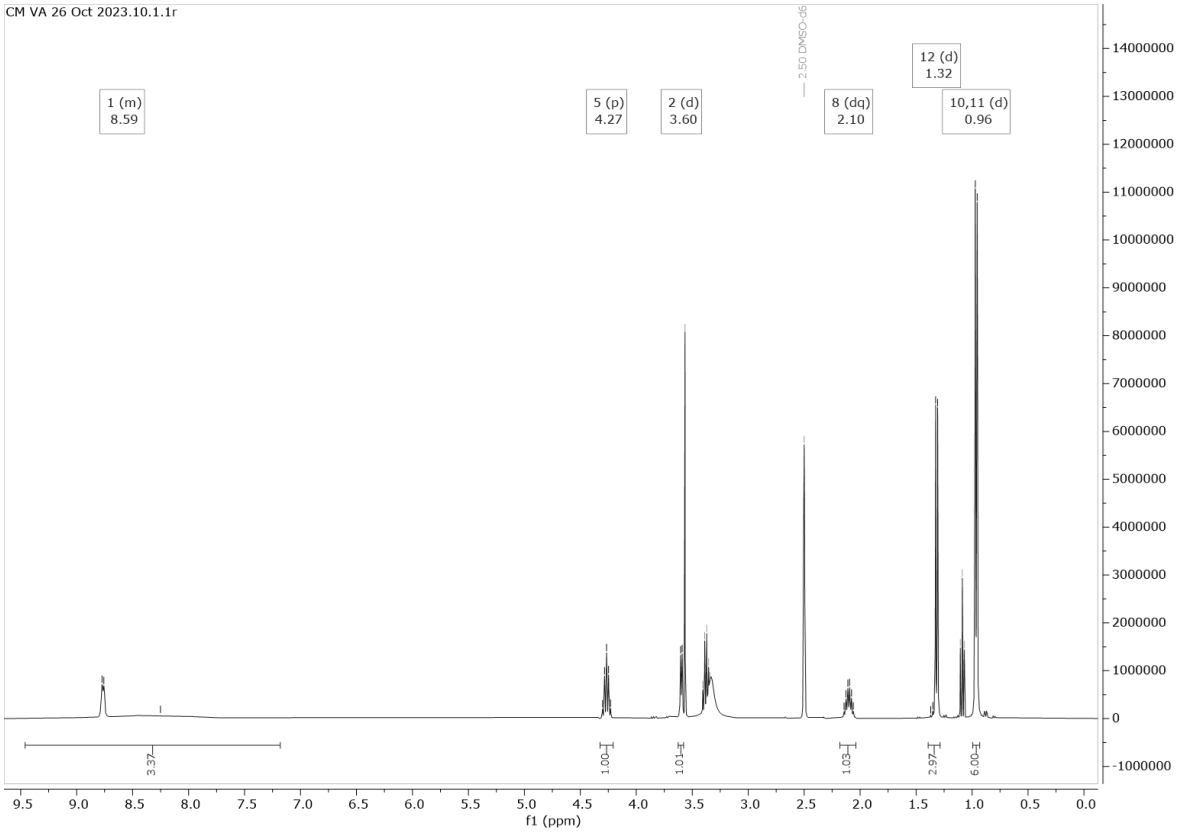

CM IV 02 Nov 2023.10.1.1r
1 (m)
8.50
2 (dd)
4.16
5 (d)
3.79
8 (m)
2.10
10" (ddd)
1.55
13 (ddt)
1.85
11 (t)
0.86
10' (m)
1.14
9,14,15 (d)
0.94
3.20
0.99
0.95
1.00
0.99
0.98
1.05
8.94
3.05
f1 (ppm)

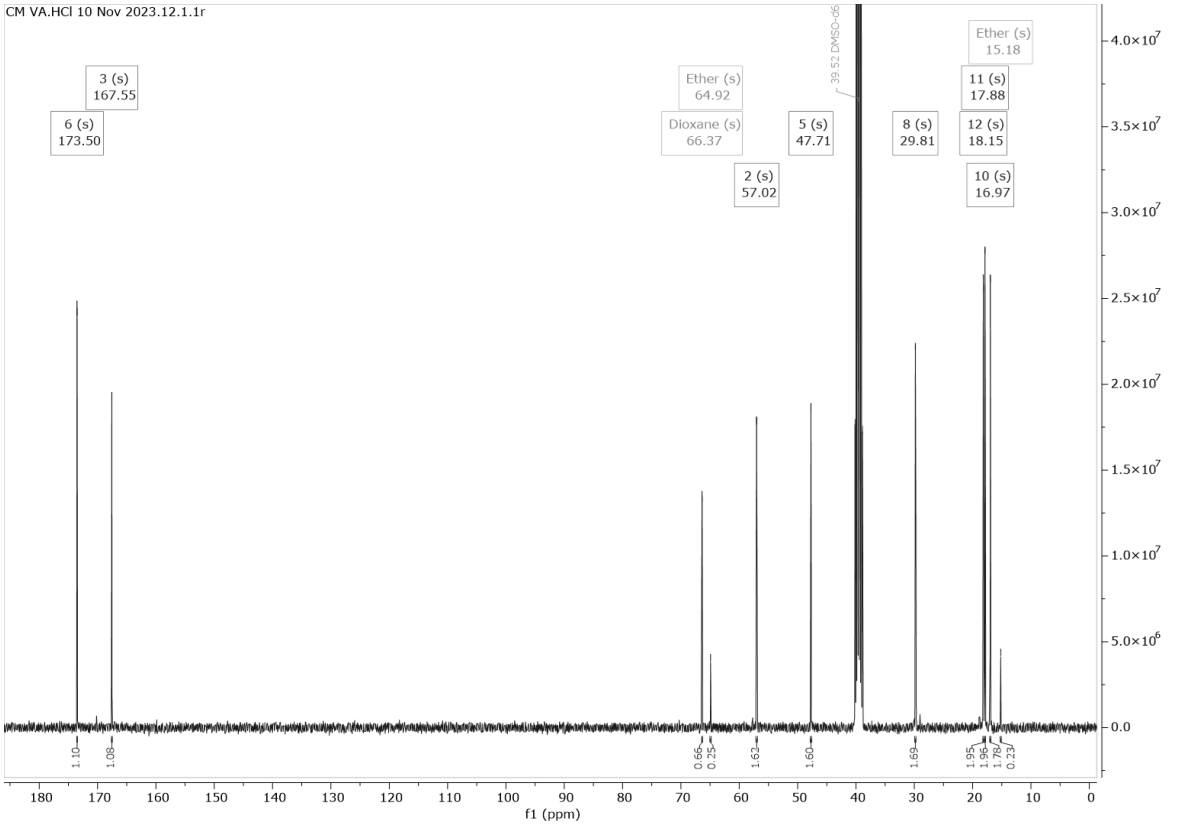

CM IV C 06 Nov 2023.10.1.1r
39.52 DMSO-d6
3 (s)
168.24
6 (s)
172.33
5 (s)
56.06
2 (s)
57.63
13 (s)
36.32
8 (s)
29.66
10 (s)
23.83
15 (s)
19.10
9 (s)
14.44
11 (s)
11.19
14 (s)
18.10
f1 (ppm)

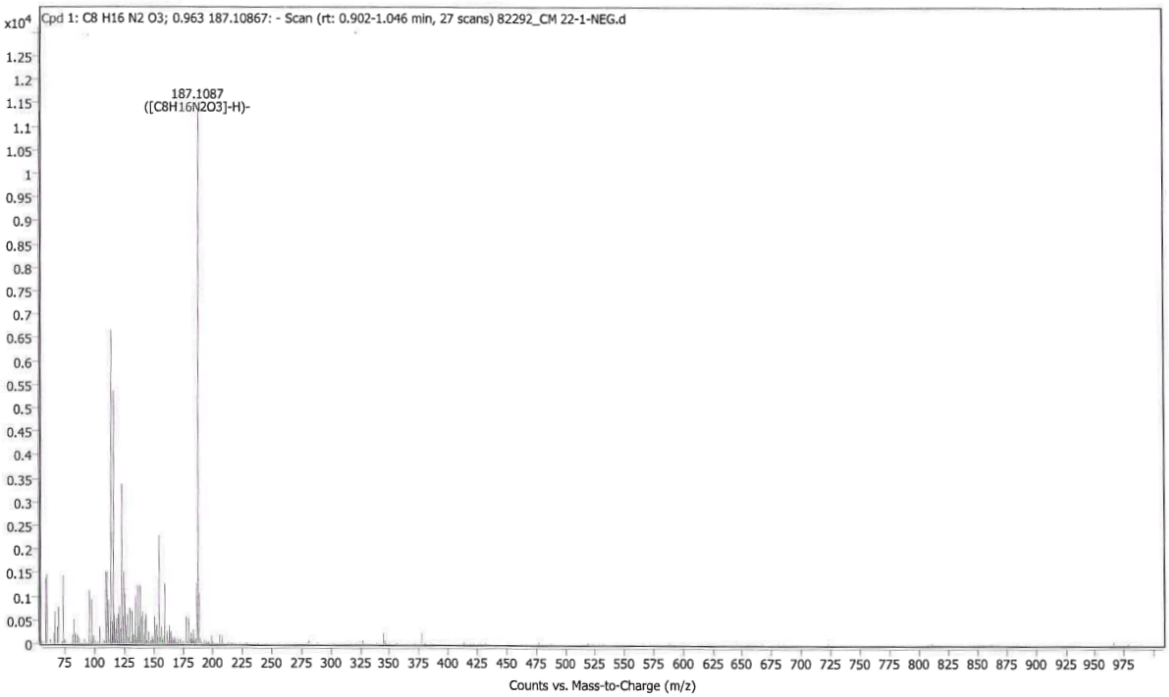

Cpd 1: C11 H22 N2 O3; 0.194 231.17068: + Scan (rt: 0.166-0.244 min, 15 scans) 82593_CM-31-2.d
x10^5
231.1707
([C11H22N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

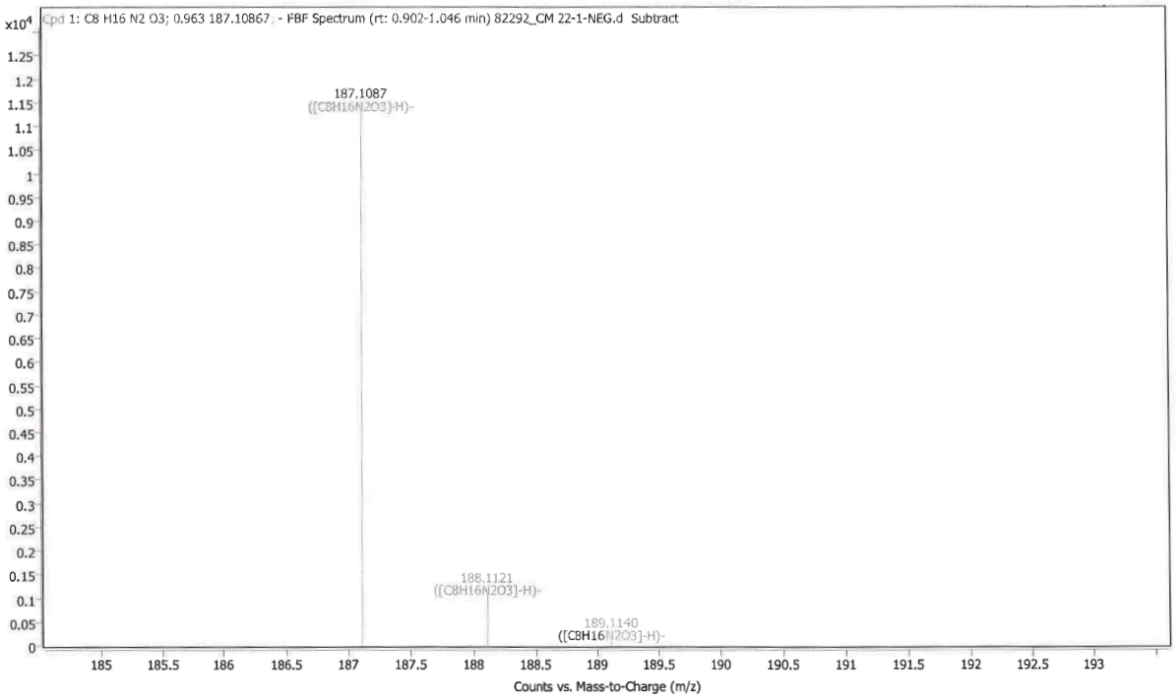

Cpd 1: C11 H22 N2 O3; 0.194 231.17068: + FBF Spectrum (rt: 0.166-0.244 min) 82593_CM-31-2.d Subtract
x10^5
231.1707
([C11H22N2O3]+H)+
232.1739
([C11H22N2O3]+H)+
230.1617
233.1758
([C11H22N2O3]+H)+
234.1795
([C11H22N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

VA·HCl **D6**

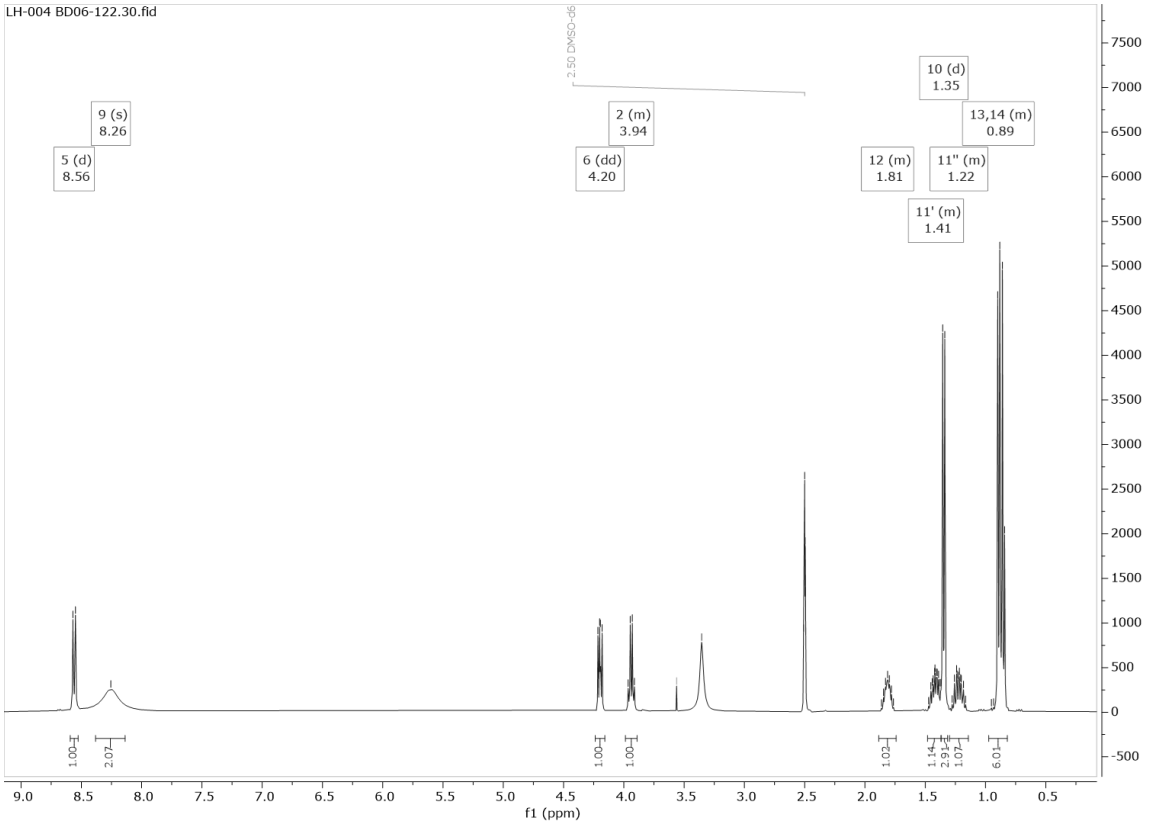

CM VA 26 Oct 2023.10.1.1r
2.50 DMSO-d6
1 (m) 8.59
5 (p) 4.27
2 (d) 3.60
8 (dq) 2.10
12 (d) 1.32
10,11 (d) 0.96
3.37
1.00
1.01
1.03
2.97
6.00
f1 (ppm)

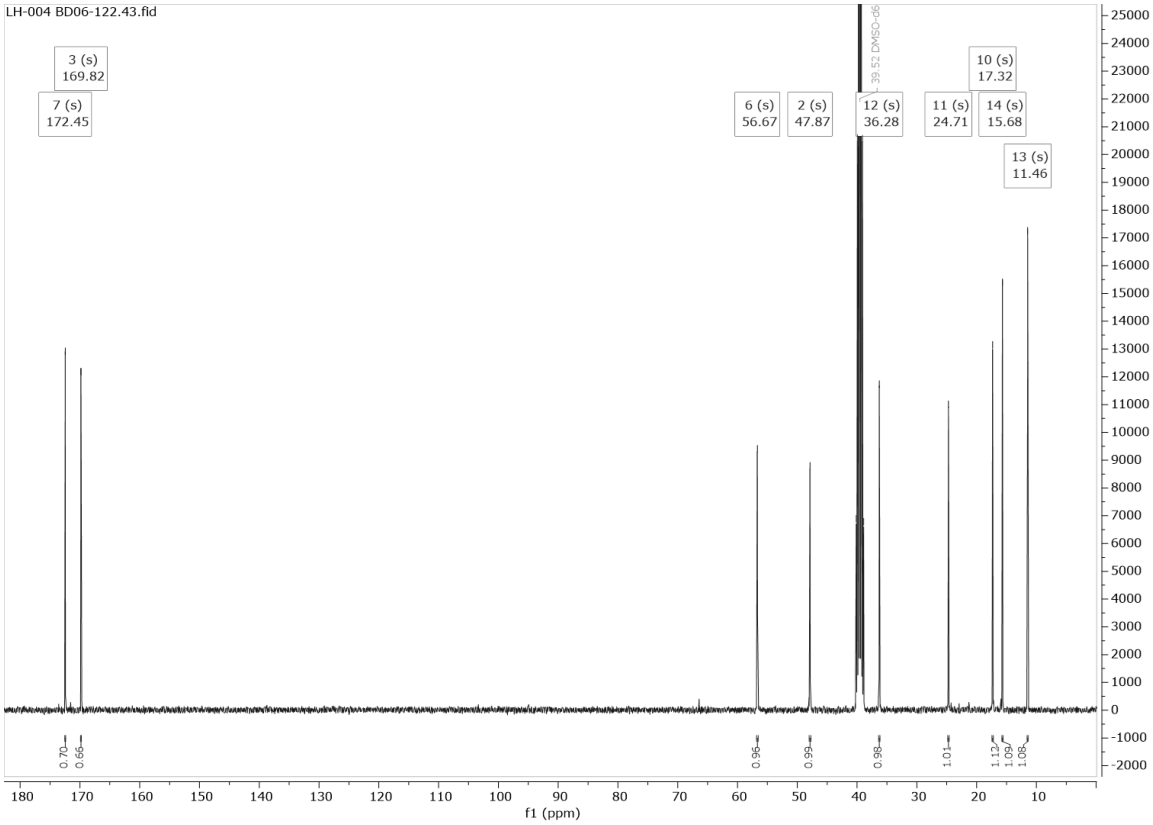

CM VA.HCl 10 Nov 2023.12.1.1r
39.52 DMSO-d6
3 (s) 167.55
6 (s) 173.50
Ether (s) 64.92
Dioxane (s) 66.37
2 (s) 57.02
5 (s) 47.71
8 (s) 29.81
Ether (s) 15.18
11 (s) 17.88
12 (s) 18.15
10 (s) 16.97
1.10
1.08
0.66
0.25
1.62
1.60
1.69
1.95
1.96
1.78
0.23
f1 (ppm)

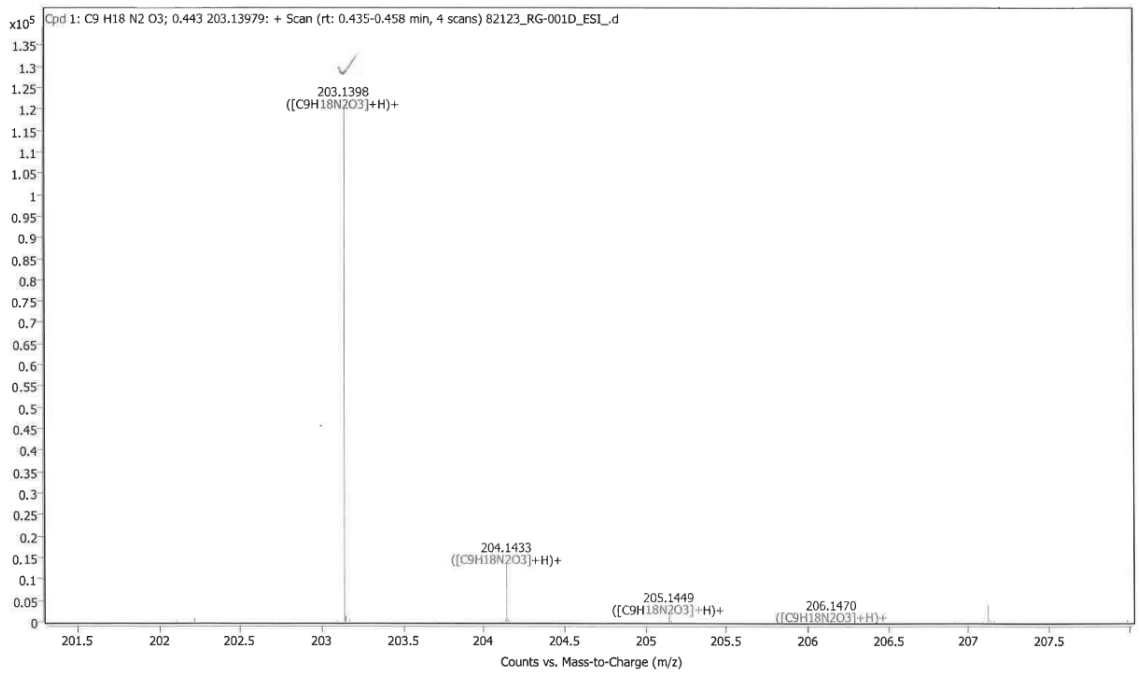

Cpd 1: C8 H16 N2 O3; 0.963 187.10867: - Scan (rt: 0.902-1.046 min, 27 scans) 82292_CM 22-1-NEG.d
x10^4
187.1087
([C8H16N2O3]-H)-
Counts vs. Mass-to-Charge (m/z)

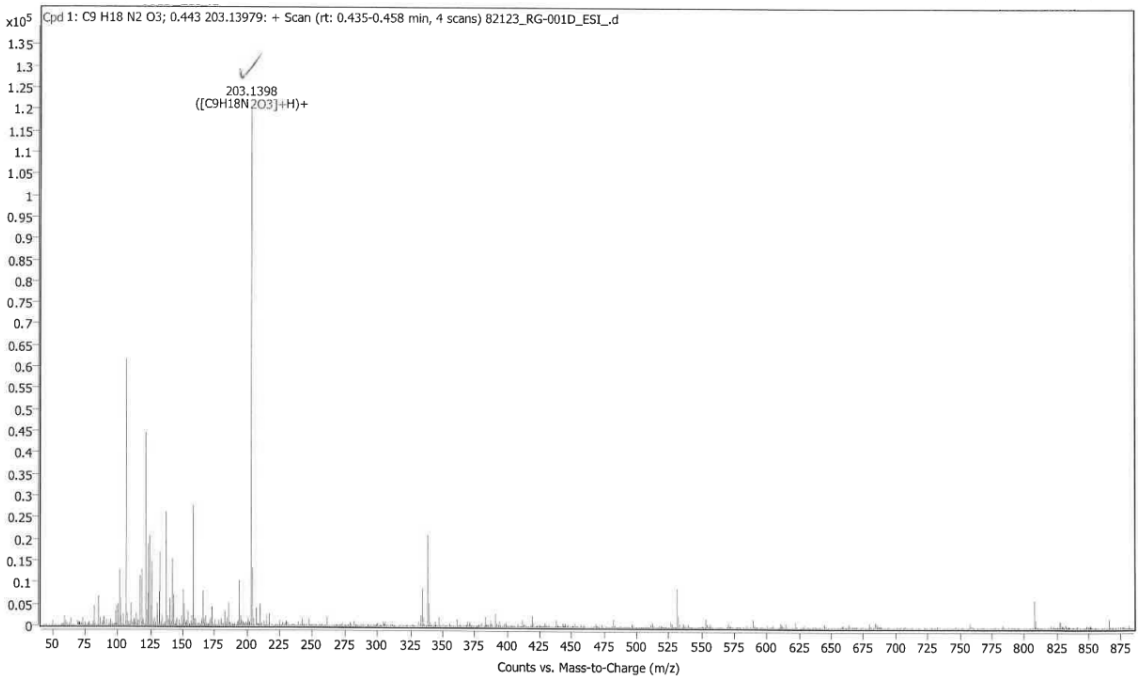

Cpd 1: C8 H16 N2 O3; 0.963 187.10867 - FBF Spectrum (rt: 0.902-1.046 min) 82292_CM 22-1-NEG.d Subtract
x10^4
187.1087
([C8H16N2O3]-H)-
188.1121
([C8H16N2O3]-H)-
189.1140
([C8H16N2O3]-H)-
Counts vs. Mass-to-Charge (m/z)

AL·TFA **D7**

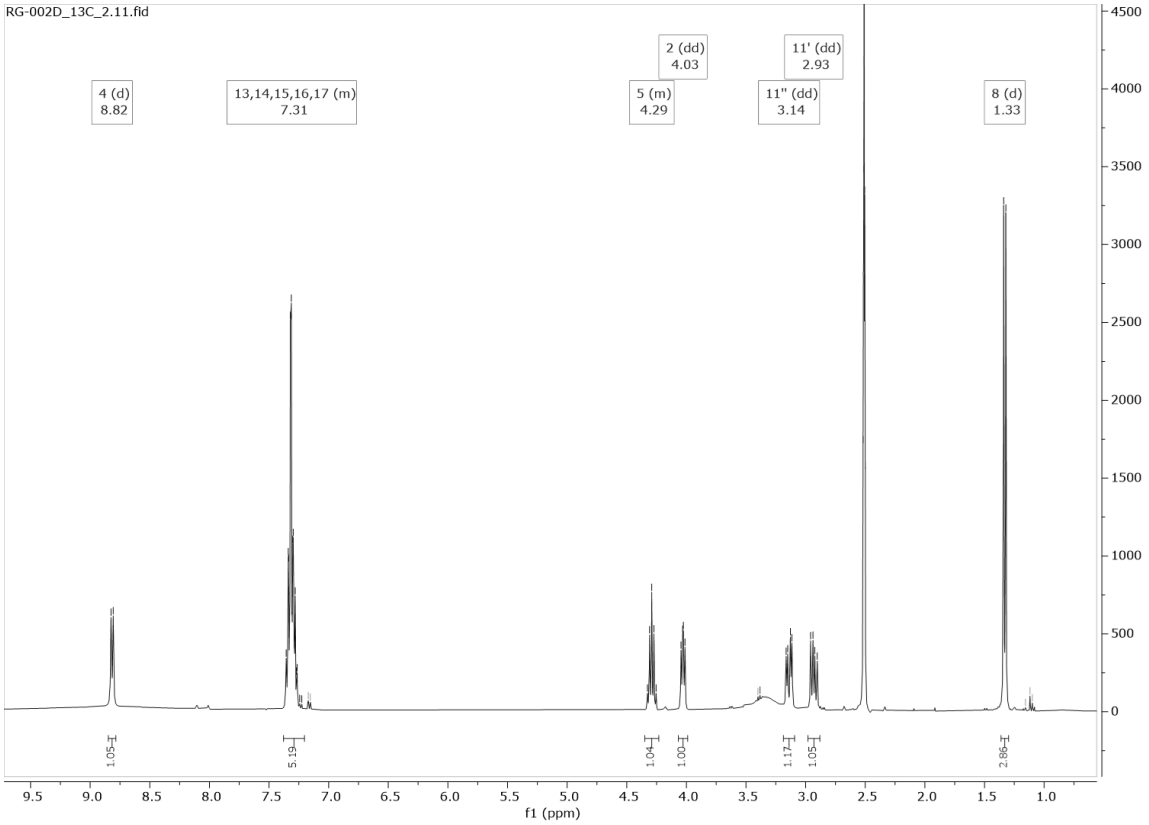

LH-004 BD06-122.30.fid
2.50 DMSO-d6
9 (s)
8.26
5 (d)
8.56
2 (m)
3.94
6 (dd)
4.20
10 (d)
1.35
13,14 (m)
0.89
12 (m)
1.81
11" (m)
1.22
11' (m)
1.41
1.00
2.07
1.00
1.00
1.02
1.14
2.91
1.07
6.01
f1 (ppm)

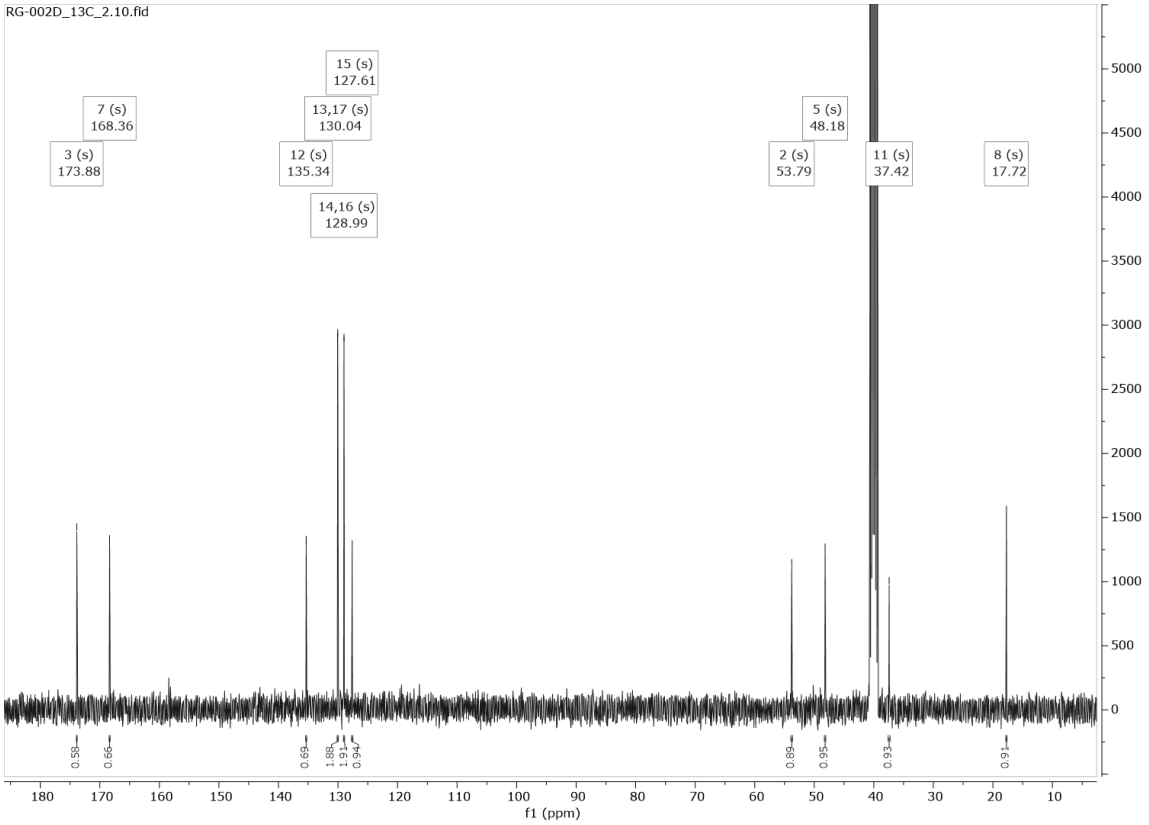

LH-004 BD06-122.43.fid
39.52 DMSO-d6
3 (s)
169.82
7 (s)
172.45
6 (s)
56.67
2 (s)
47.87
12 (s)
36.28
11 (s)
24.71
10 (s)
17.32
14 (s)
15.68
13 (s)
11.46
0.70
0.66
0.96
0.99
0.98
1.01
1.13
1.09
1.08
f1 (ppm)


Cpd 1: C9 H18 N2 O3; 0.443 203.13979: + Scan (rt: 0.435-0.458 min, 4 scans) 82123_RG-001D_ESI_.d
x10^5
203.1398
([C9H18N2O3]+H)+
204.1433
([C9H18N2O3]+H)+
205.1449
([C9H18N2O3]+H)+
206.1470
([C9H18N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

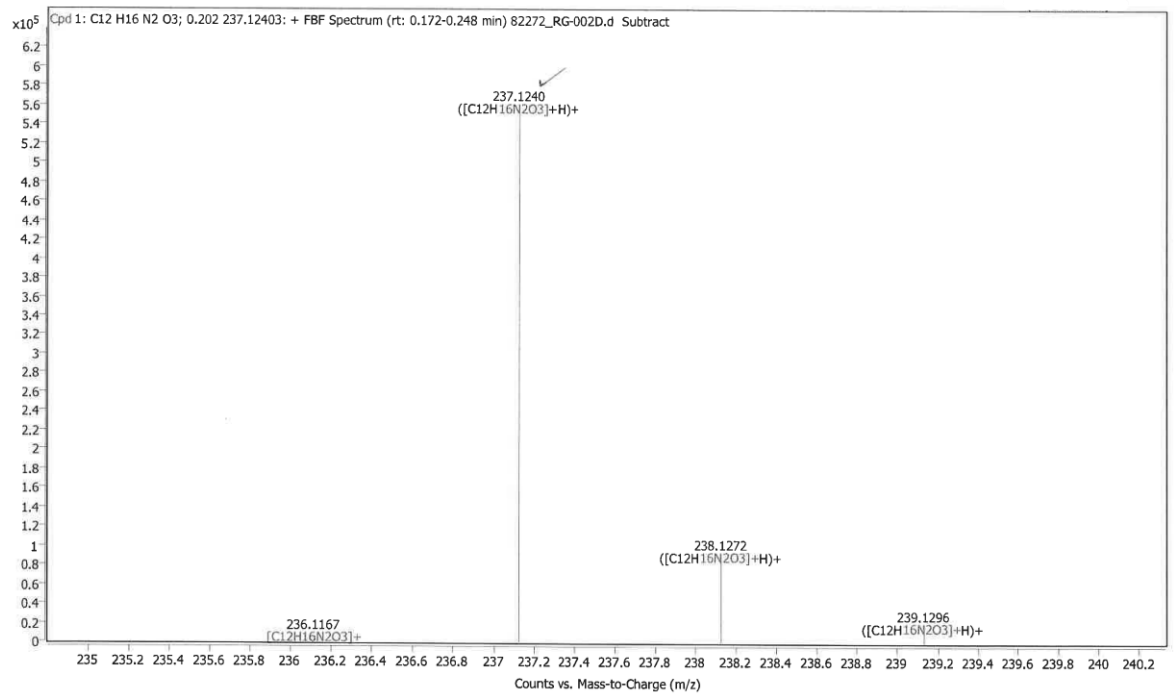

Cpd 1: C9 H18 N2 O3; 0.443 203.13979: + Scan (rt: 0.435-0.458 min, 4 scans) 82123_RG-001D_ESI_.d
x10^5
203.1398
([C9H18N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

FA·TFA **D8**

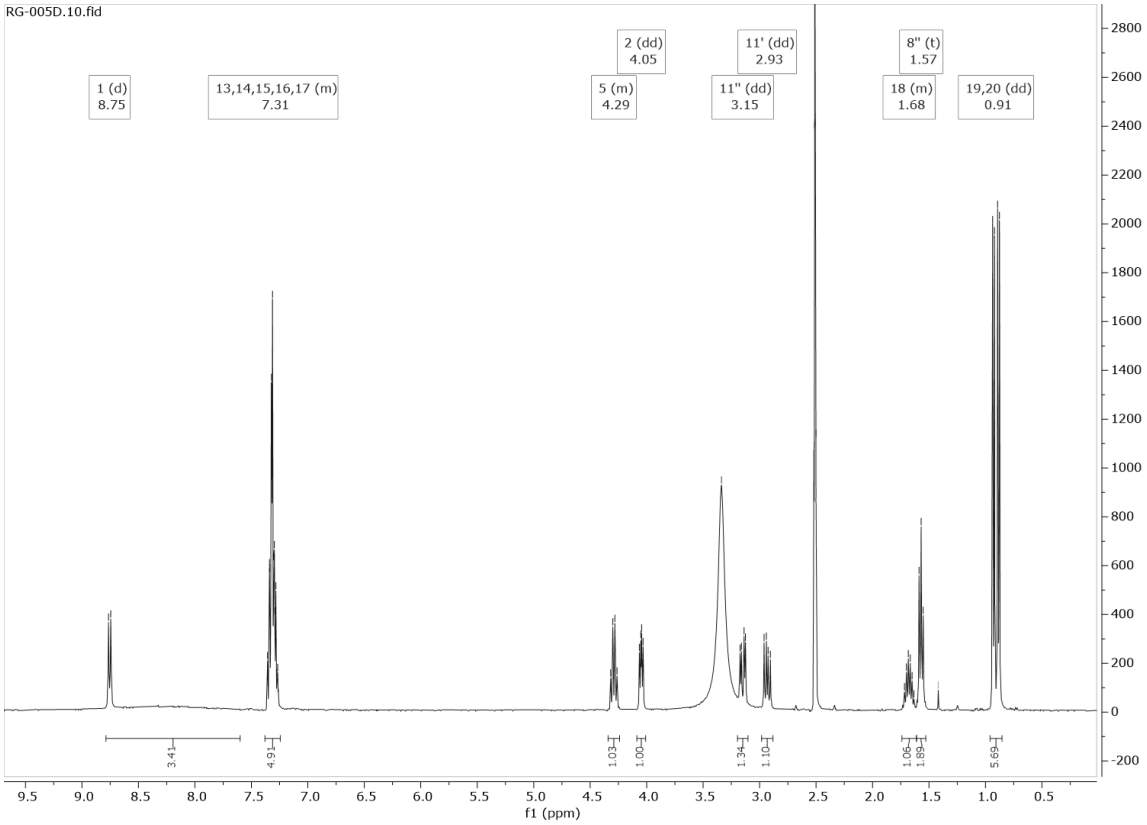

RG-002D_13C_2.11.fid
4 (d)
8.82
13,14,15,16,17 (m)
7.31
2 (dd)
4.03
5 (m)
4.29
11' (dd)
2.93
11" (dd)
3.14
8 (d)
1.33
1.05
5.19
1.04
1.00
1.17
1.05
2.86
f1 (ppm)

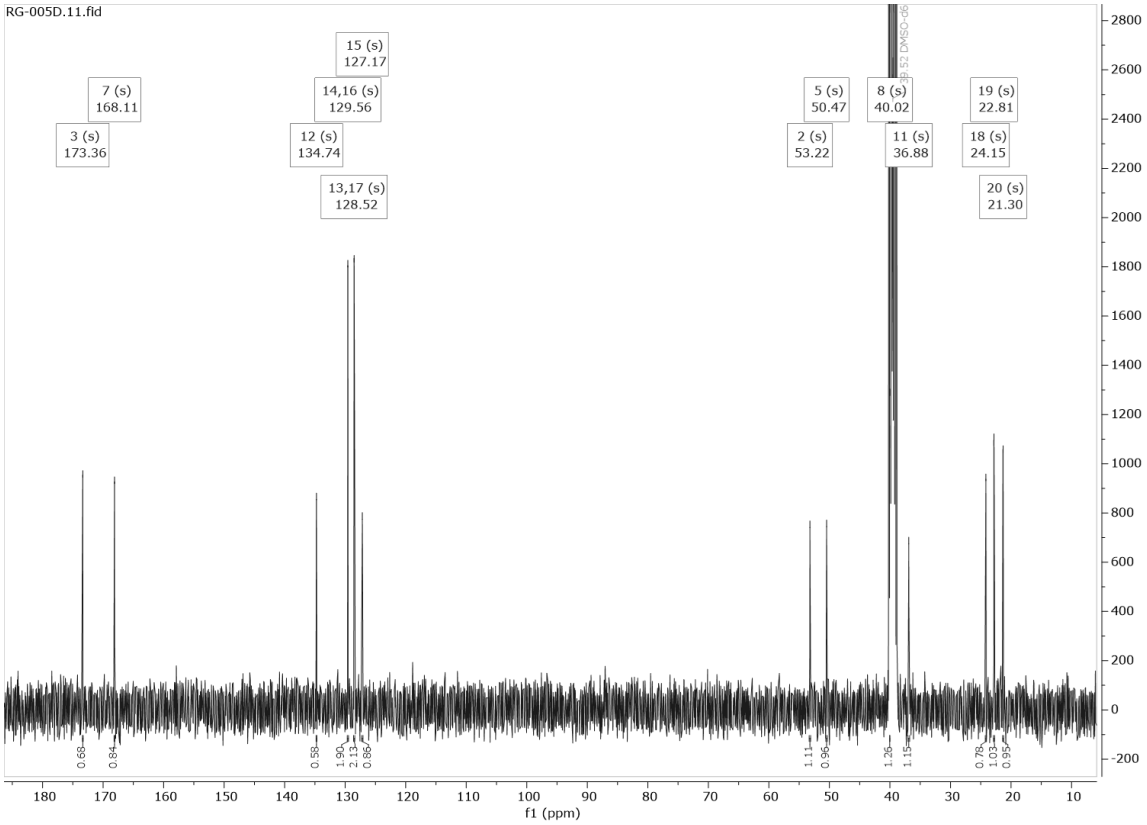

RG-002D_13C_2.10.fid
15 (s)
127.61
7 (s)
168.36
13,17 (s)
130.04
5 (s)
48.18
3 (s)
173.88
12 (s)
135.34
2 (s)
53.79
11 (s)
37.42
8 (s)
17.72
14,16 (s)
128.99
0.58
0.66
0.69
1.88
1.91
0.94
0.89
0.95
0.93
0.91
f1 (ppm)

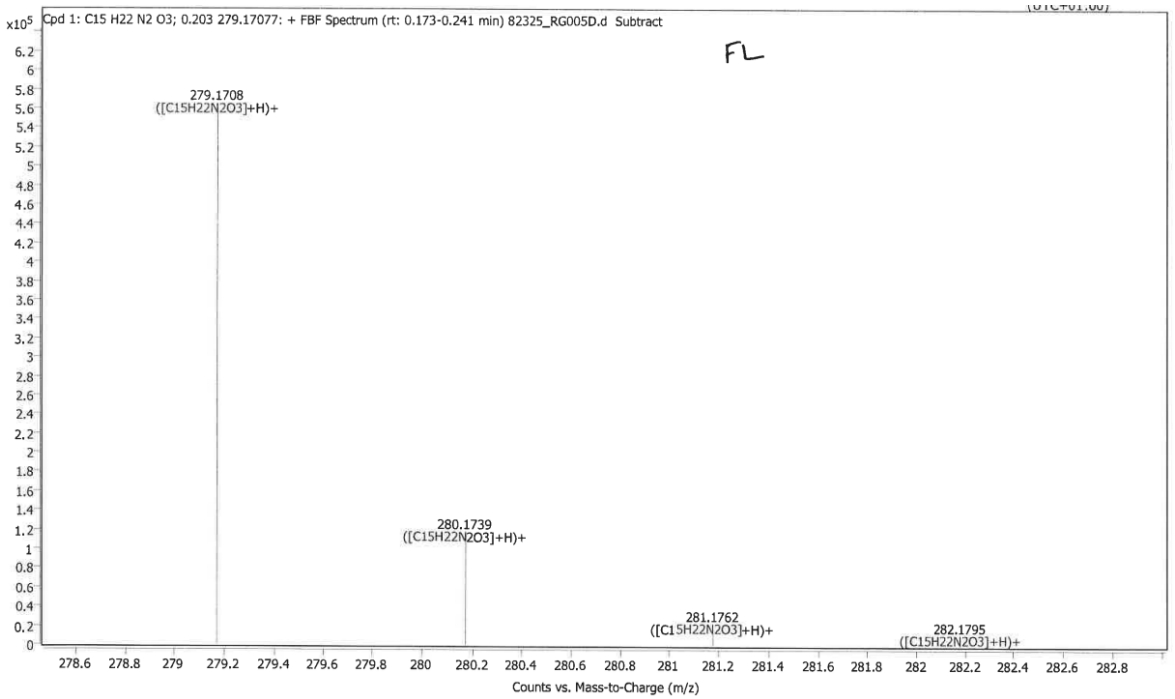

Cpd 1: C12 H16 N2 O3; 0.202 237.12403: + Scan (rt: 0.172-0.248 min, 11 scans) 82272_RG-002D.d
x10^5
237.1240
([C12H16N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

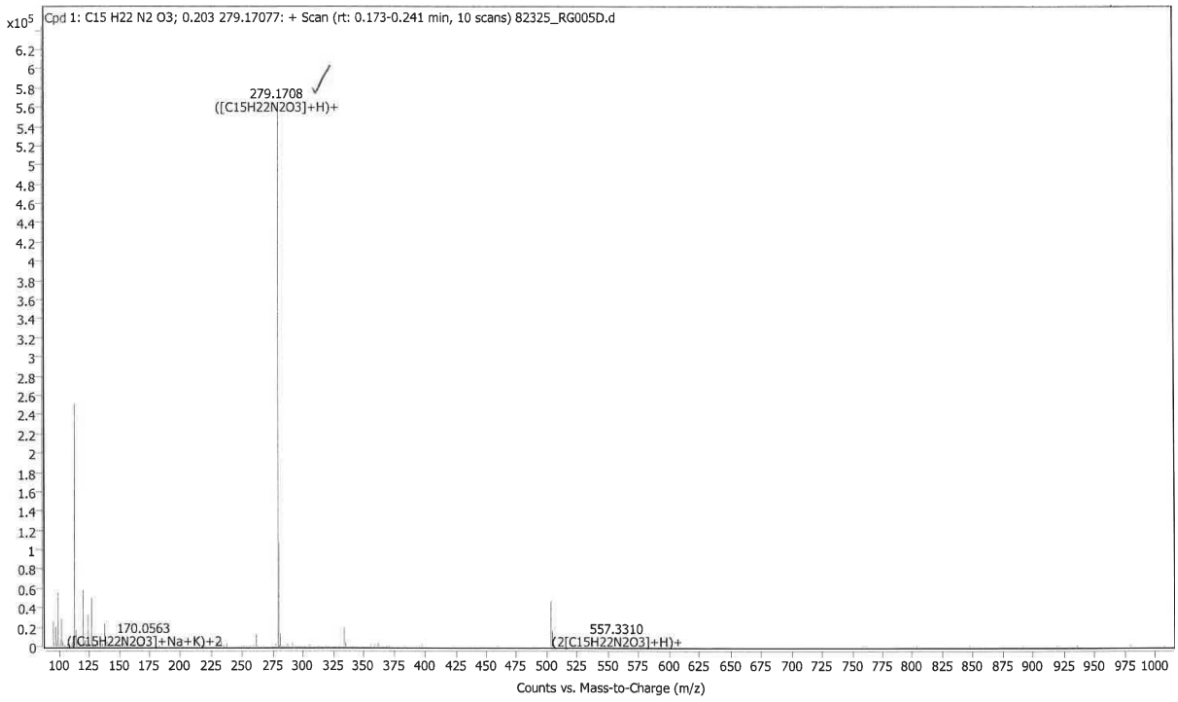

Cpd 1: C12 H16 N2 O3; 0.202 237.12403: + FBF Spectrum (rt: 0.172-0.248 min) 82272_RG-002D.d Subtract
x10^5
237.1240
([C12H16N2O3]+H)+
238.1272
([C12H16N2O3]+H)+
236.1167
[C12H16N2O3]+
239.1296
([C12H16N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

FL·TFA **D9**

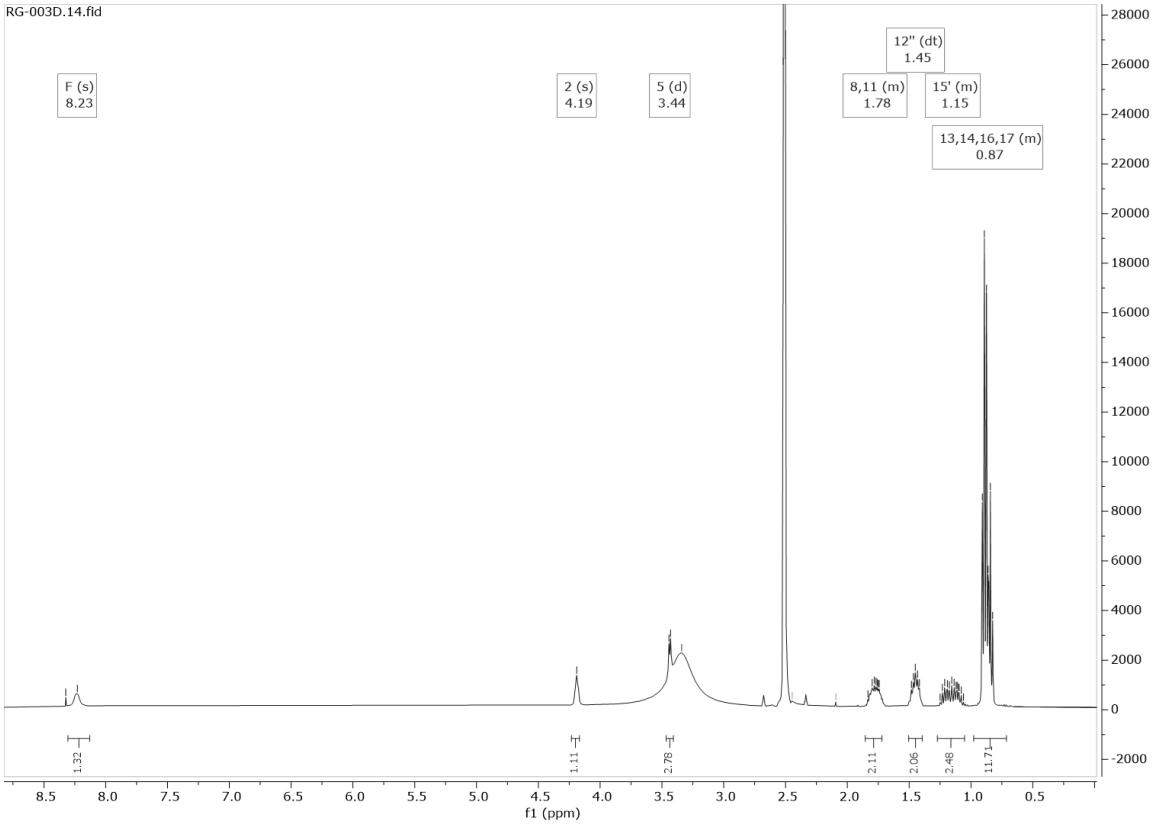

RG-005D.10.fid
1 (d)
8.75
13,14,15,16,17 (m)
7.31
2 (dd)
4.05
5 (m)
4.29
11' (dd)
2.93
11" (dd)
3.15
8" (t)
1.57
18 (m)
1.68
19,20 (dd)
0.91
f1 (ppm)

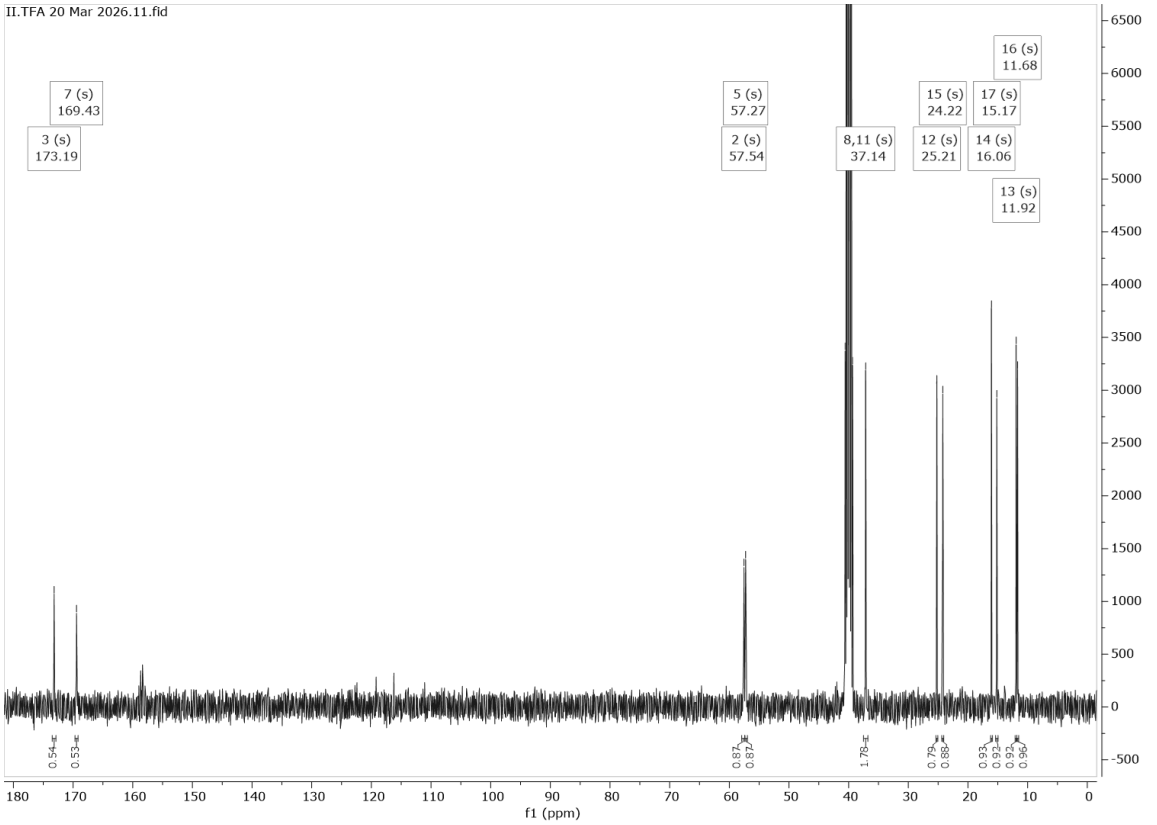

RG-005D.11.fid
7 (s)
168.11
3 (s)
173.36
15 (s)
127.17
14,16 (s)
129.56
12 (s)
134.74
13,17 (s)
128.52
5 (s)
50.47
2 (s)
53.22
8 (s)
40.02
11 (s)
36.88
19 (s)
22.81
18 (s)
24.15
20 (s)
21.30
f1 (ppm)

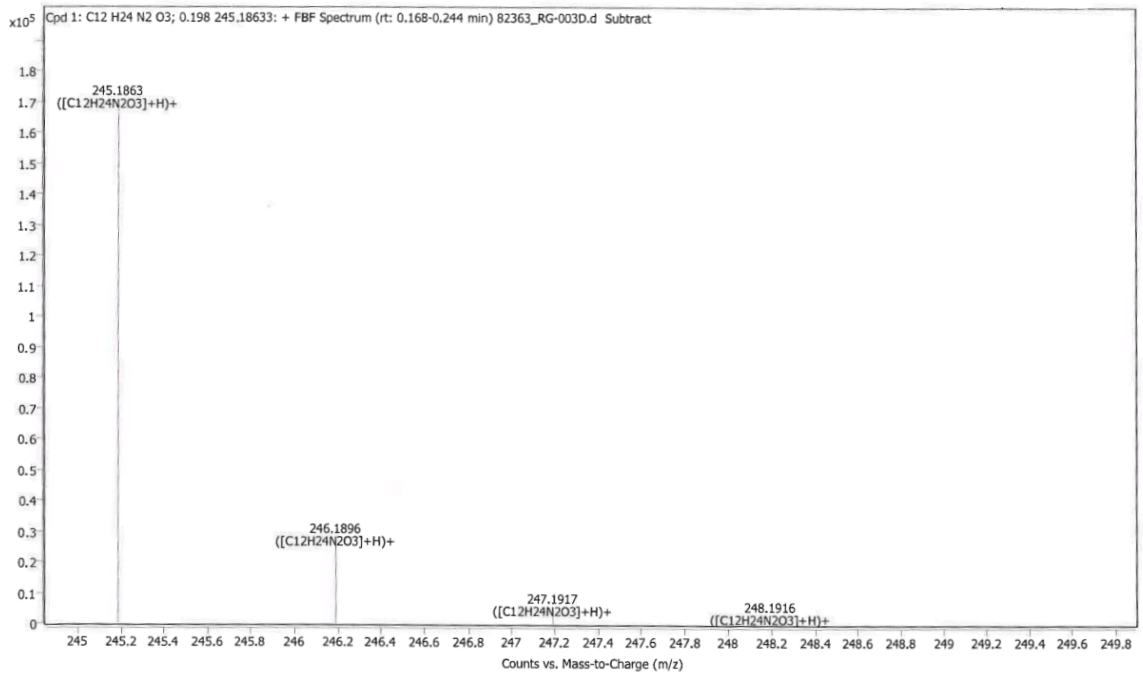

Cpd 1: C15 H22 N2 O3; 0.203 279.17077: + FBF Spectrum (rt: 0.173-0.241 min) 82325_RG005D.d Subtract
FL
279.1708
([C15H22N2O3]+H)+
280.1739
([C15H22N2O3]+H)+
281.1762
([C15H22N2O3]+H)+
282.1795
([C15H22N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

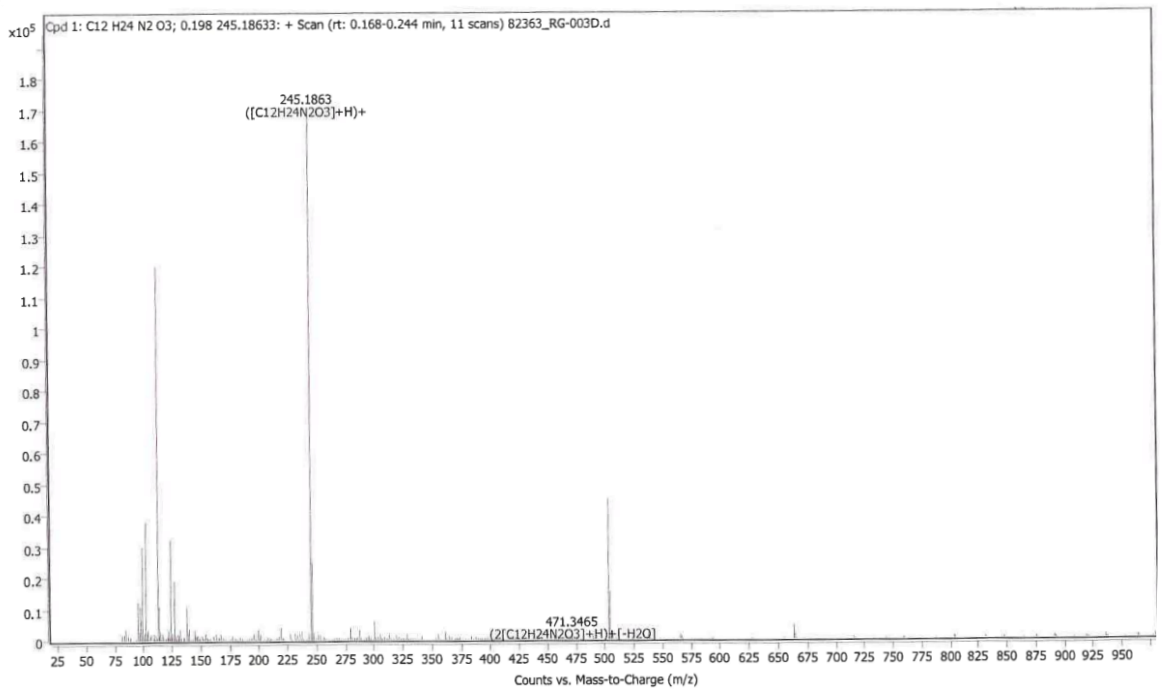

Cpd 1: C15 H22 N2 O3; 0.203 279.17077: + Scan (rt: 0.173-0.241 min, 10 scans) 82325_RG005D.d
279.1708
([C15H22N2O3]+H)+
170.0563
([C15H22N2O3]+Na+K)+2
557.3310
(2[C15H22N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

II·TFA **D10**

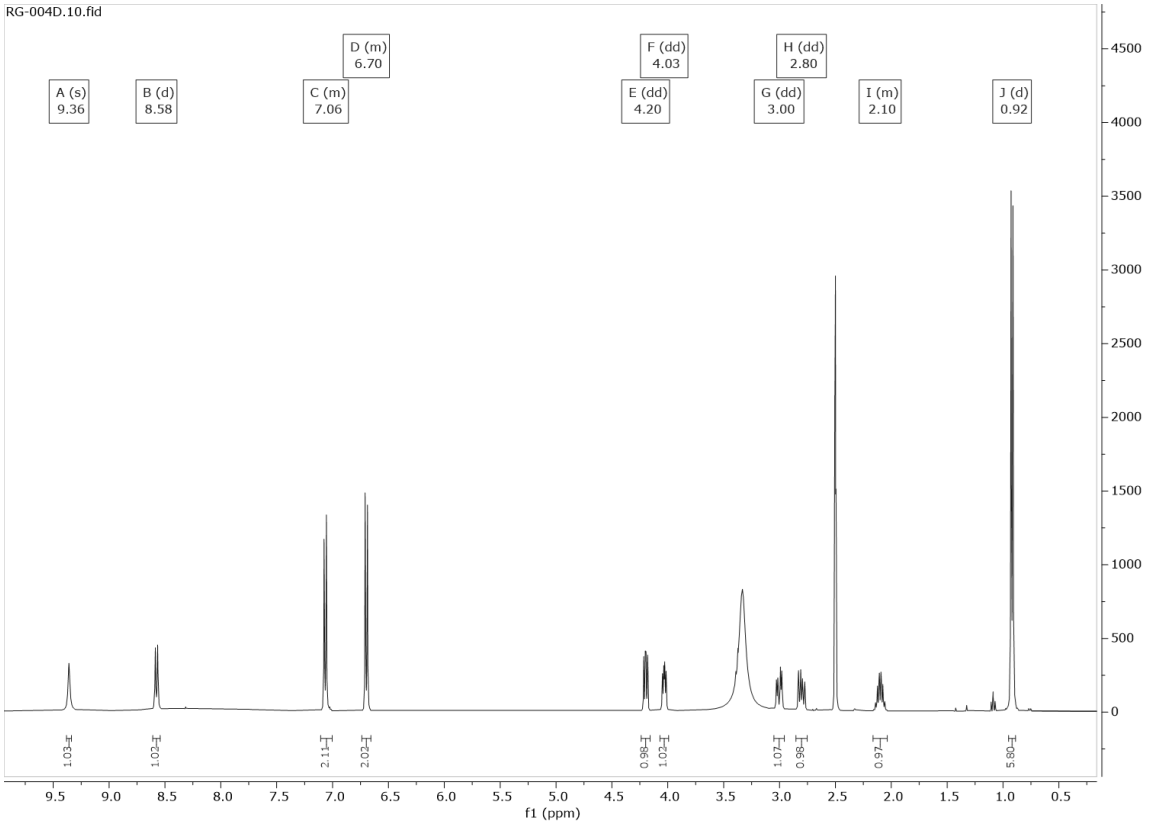

RG-003D.14.fid
F (s)
8.23
2 (s)
4.19
5 (d)
3.44
12" (dt)
1.45
8,11 (m)
1.78
15' (m)
1.15
13,14,16,17 (m)
0.87
1.32
1.11
2.78
2.11
2.06
2.48
11.71
f1 (ppm)

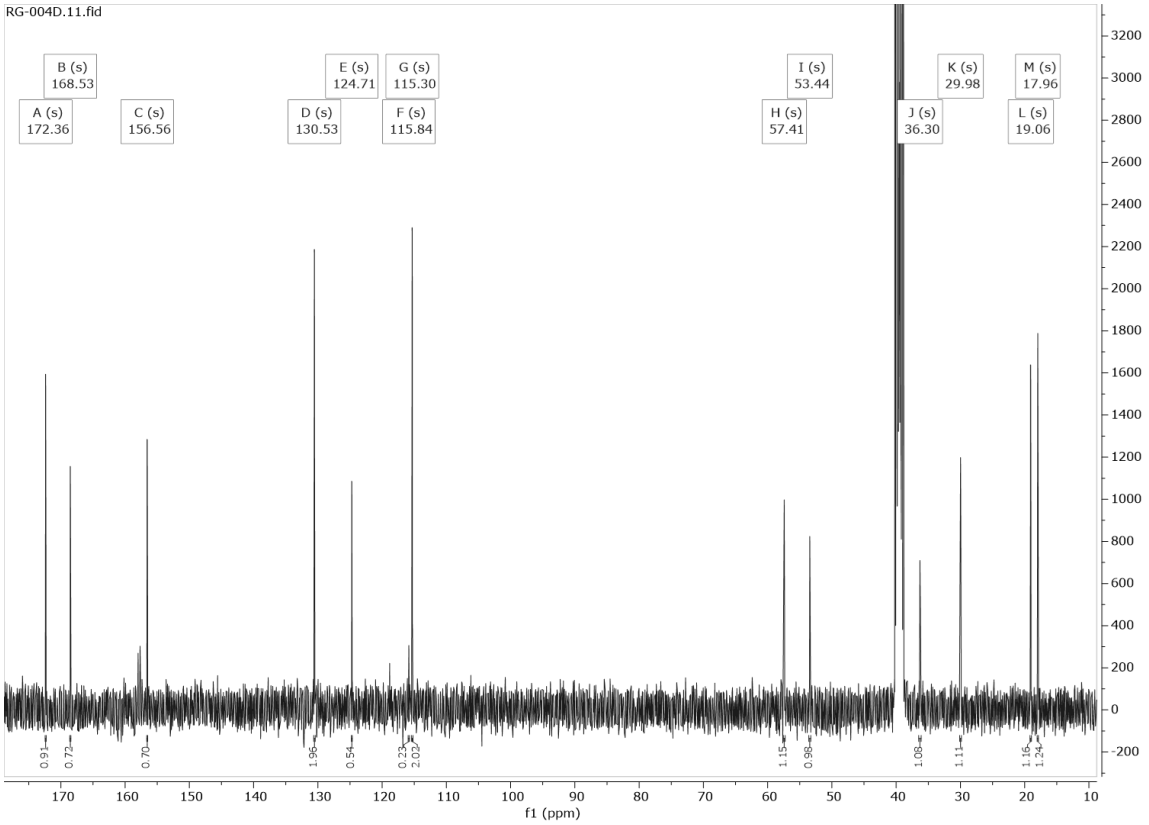

II.TFA 20 Mar 2026.11.fid
7 (s)
169.43
3 (s)
173.19
5 (s)
57.27
2 (s)
57.54
8,11 (s)
37.14
16 (s)
11.68
15 (s)
24.22
17 (s)
15.17
12 (s)
25.21
14 (s)
16.06
13 (s)
11.92
0.54
0.53
0.87
0.87
1.78
0.79
0.88
0.93
0.92
0.92
0.96
f1 (ppm)

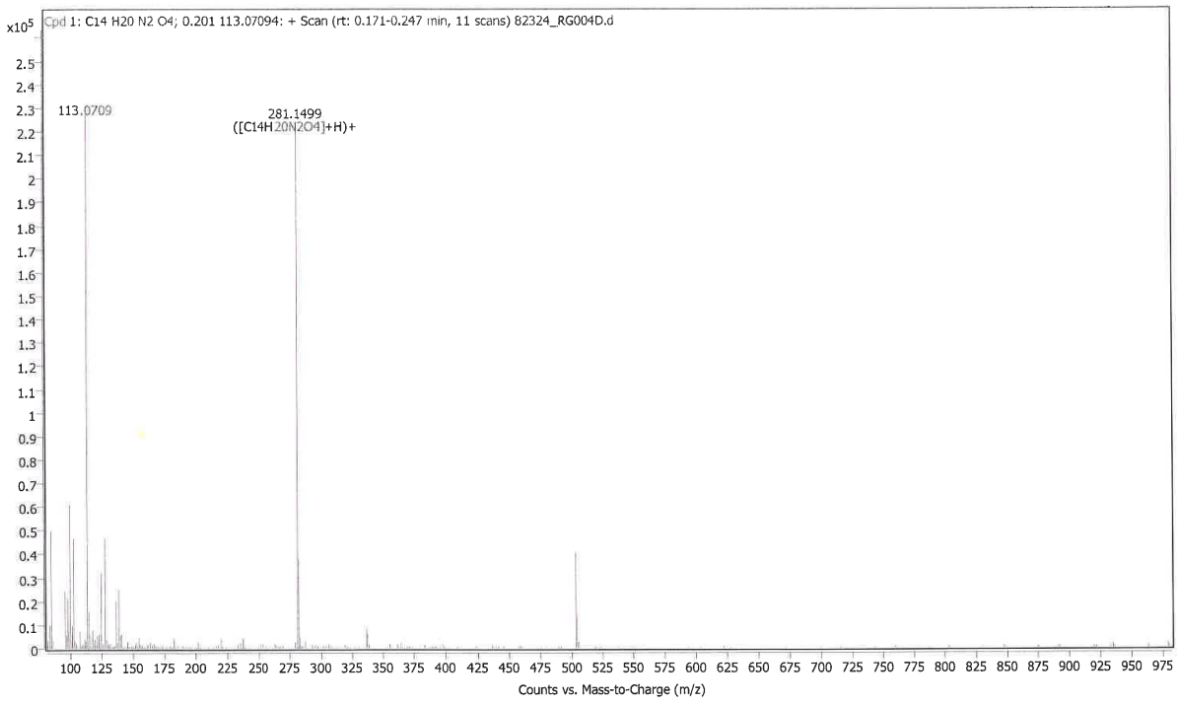

Cpd 1: C12 H24 N2 O3; 0.198 245.18633: + FBF Spectrum (rt: 0.168-0.244 min) 82363_RG-003D.d Subtract
x10^5
245.1863
([C12H24N2O3]+H)+
246.1896
([C12H24N2O3]+H)+
247.1917
([C12H24N2O3]+H)+
248.1916
([C12H24N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

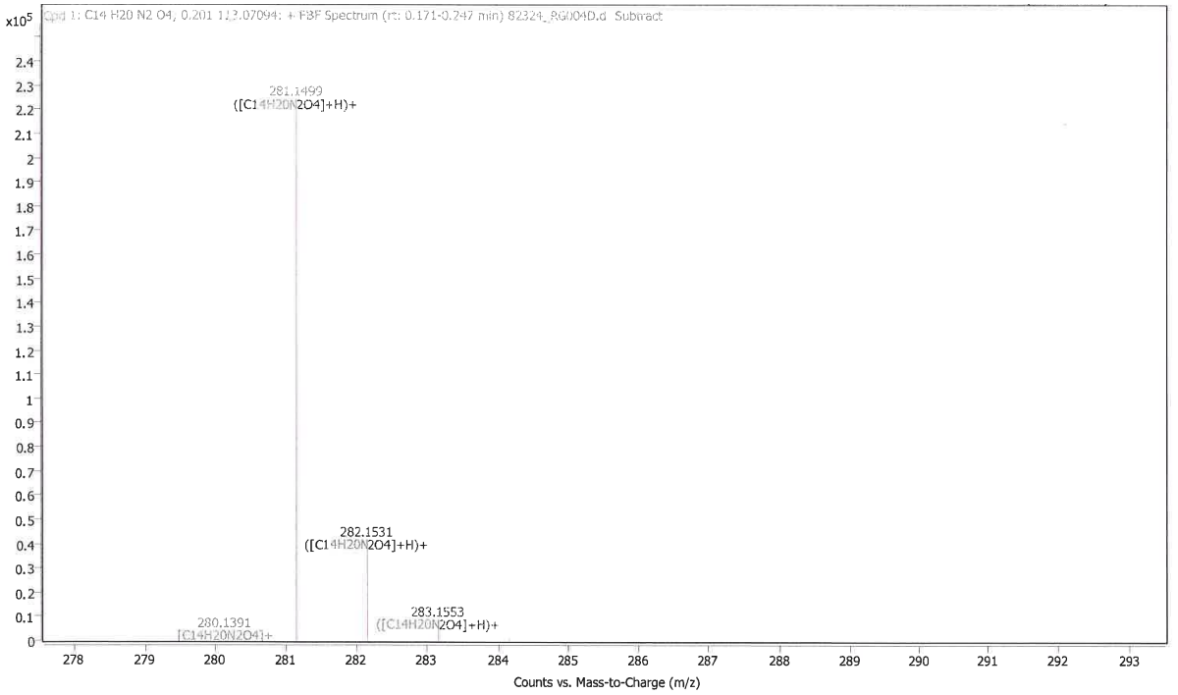

Cpd 1: C12 H24 N2 O3; 0.198 245.18633: + Scan (rt: 0.168-0.244 min, 11 scans) 82363_RG-003D.d
x10^5
245.1863
([C12H24N2O3]+H)+
471.3465
(2[C12H24N2O3]+H)+[-H2O]
Counts vs. Mass-to-Charge (m/z)

YV·TFA **D11**

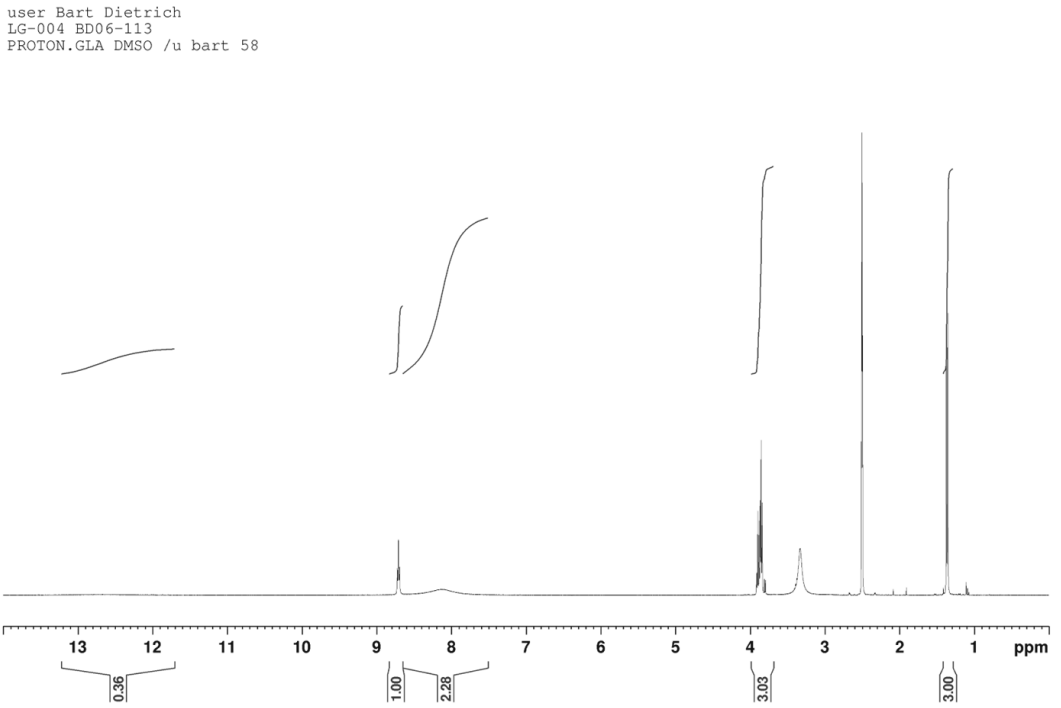

RG-004D.10.fid
A (s) 9.36
B (d) 8.58
C (m) 7.06
D (m) 6.70
E (dd) 4.20
F (dd) 4.03
G (dd) 3.00
H (dd) 2.80
I (m) 2.10
J (d) 0.92
1.03
1.03
2.11
2.03
0.98
1.02
1.07
0.98
0.97
5.80
f1 (ppm)

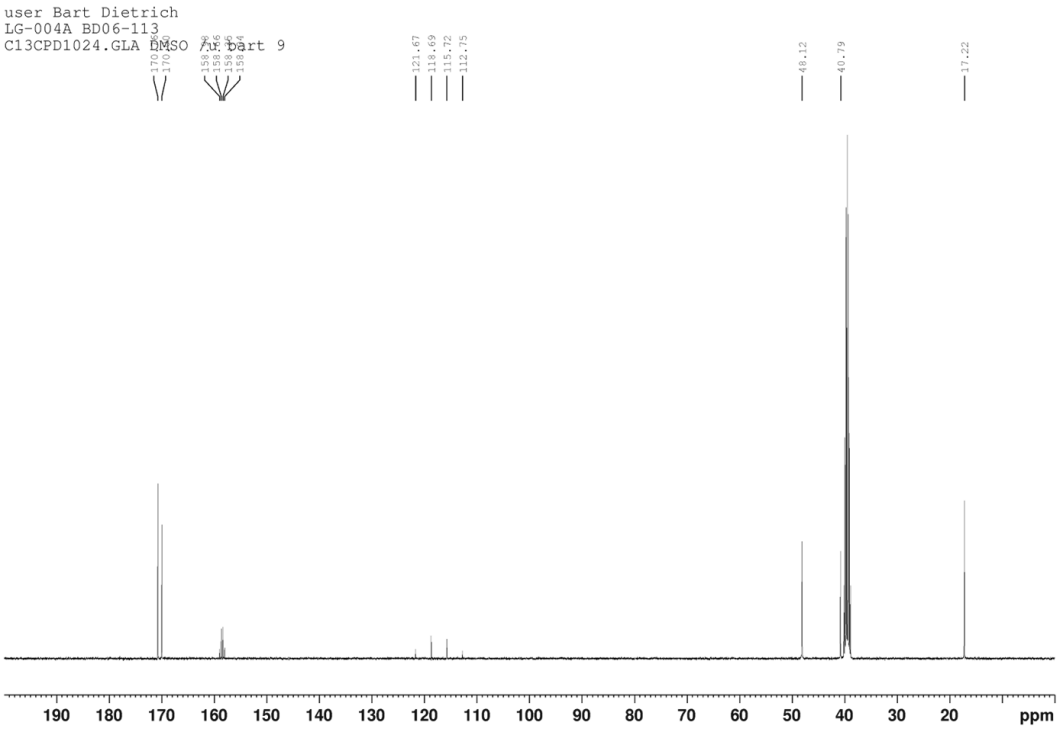

RG-004D.11.fid
A (s) 172.36
B (s) 168.53
C (s) 156.56
D (s) 130.53
E (s) 124.71
F (s) 115.84
G (s) 115.30
H (s) 57.41
I (s) 53.44
J (s) 36.30
K (s) 29.98
L (s) 19.06
M (s) 17.96
0.91
0.72
0.70
1.96
0.54
0.23
2.03
1.15
0.98
1.08
1.11
1.16
1.24
f1 (ppm)

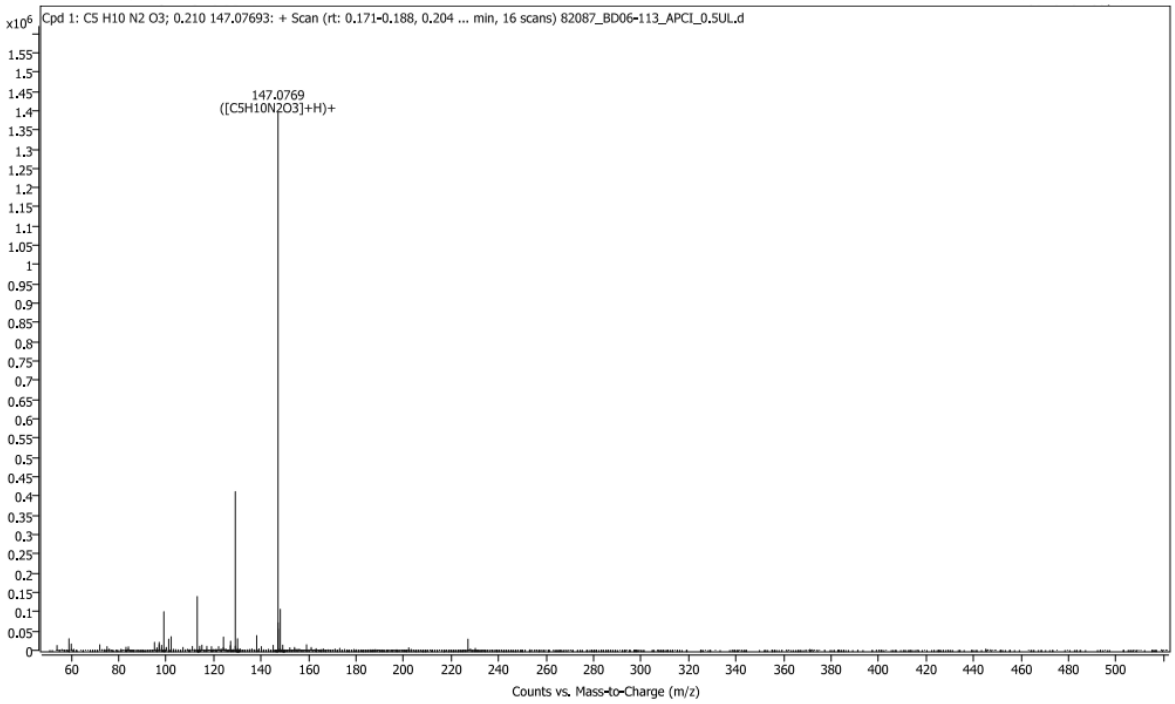

Cpd 1: C14 H20 N2 O4; 0.201 113.07094: + Scan (rt: 0.171-0.247 min, 11 scans) 82324_RG004D.d
x10^5
113.0709
281.1499
([C14H20N2O4]+H)+
Counts vs. Mass-to-Charge (m/z)

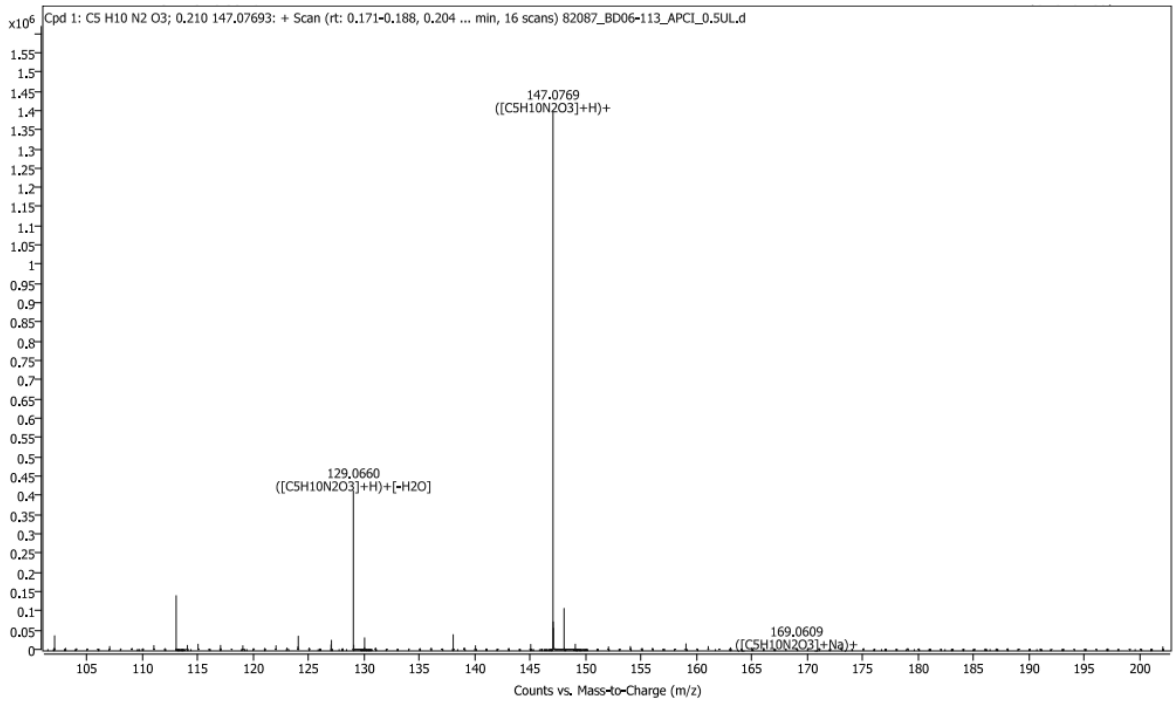

Cpd 1: C14 H20 N2 O4; 0.201 113.07094: + F3F Spectrum (rt: 0.171-0.247 min) 82324_RG004D.d Subtract
x10^5
281.1499
([C14H20N2O4]+H)+
282.1531
([C14H20N2O4]+H)+
280.1391
[C14H20N2O4]+
283.1553
([C14H20N2O4]+H)+
Counts vs. Mass-to-Charge (m/z)

AG·TFA **D12**

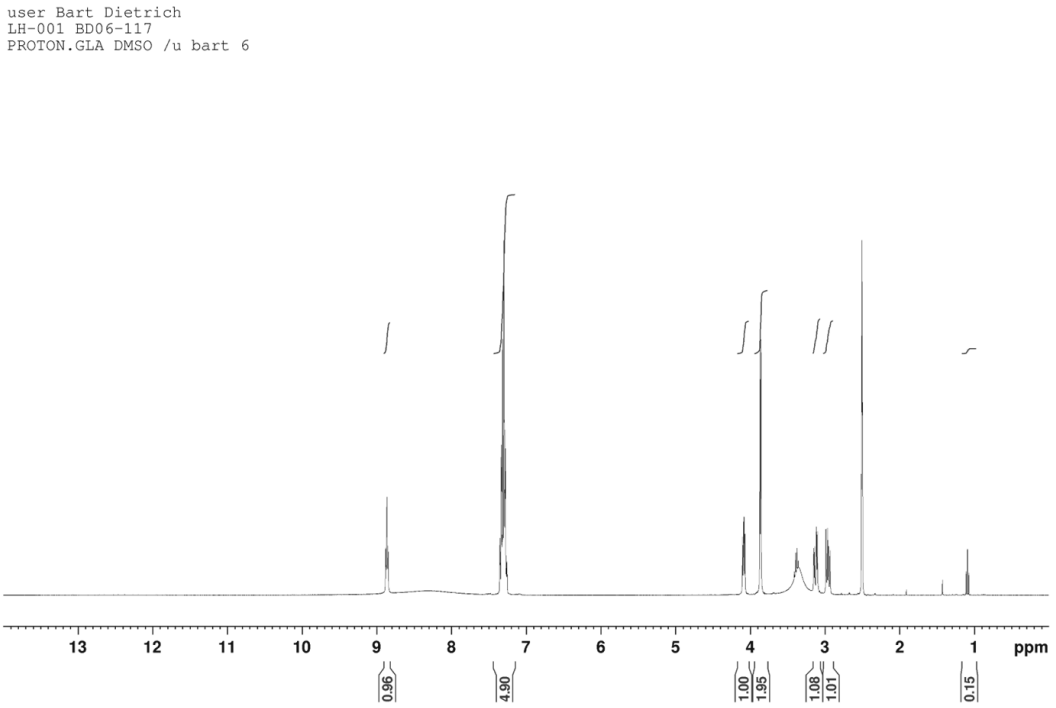

user Bart Dietrich
LG-004 BD06-113
PROTON.GLA DMSO /u bart 58
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
0.36
1.00
2.28
3.03
3.00

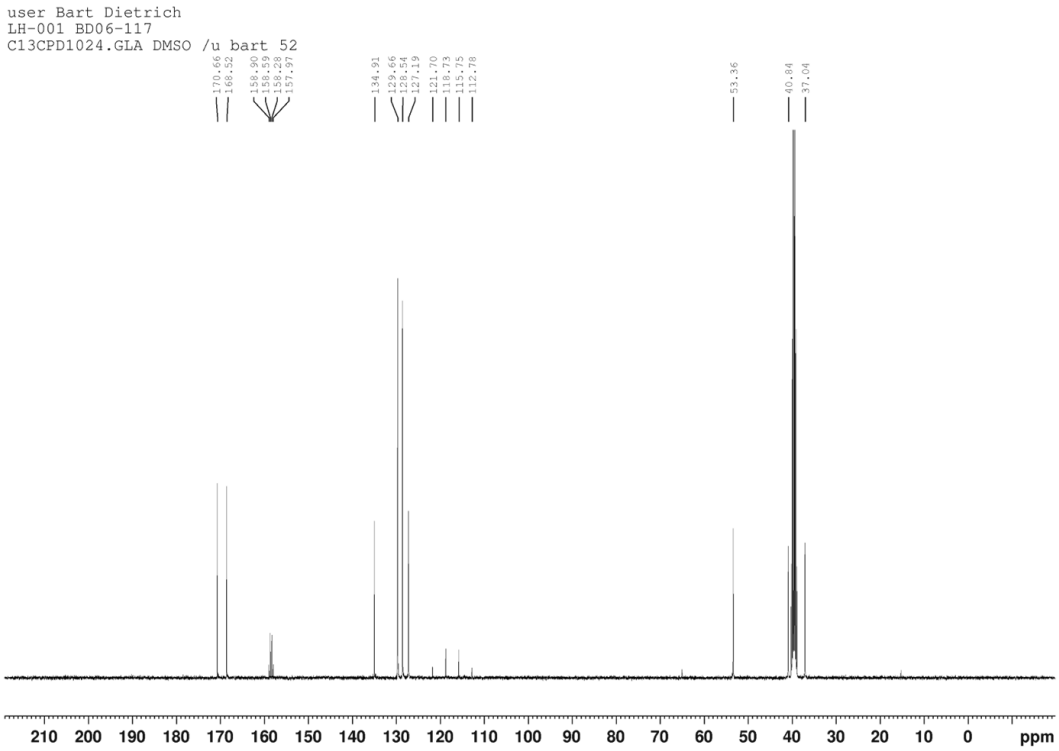

user Bart Dietrich
LG-004A BD06-113
C13CPD1024.GLA DMSO /u bart 9
170.03
169.78
158.78
158.56
158.34
158.04
121.67
118.69
115.72
112.75
48.12
40.79
17.22
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

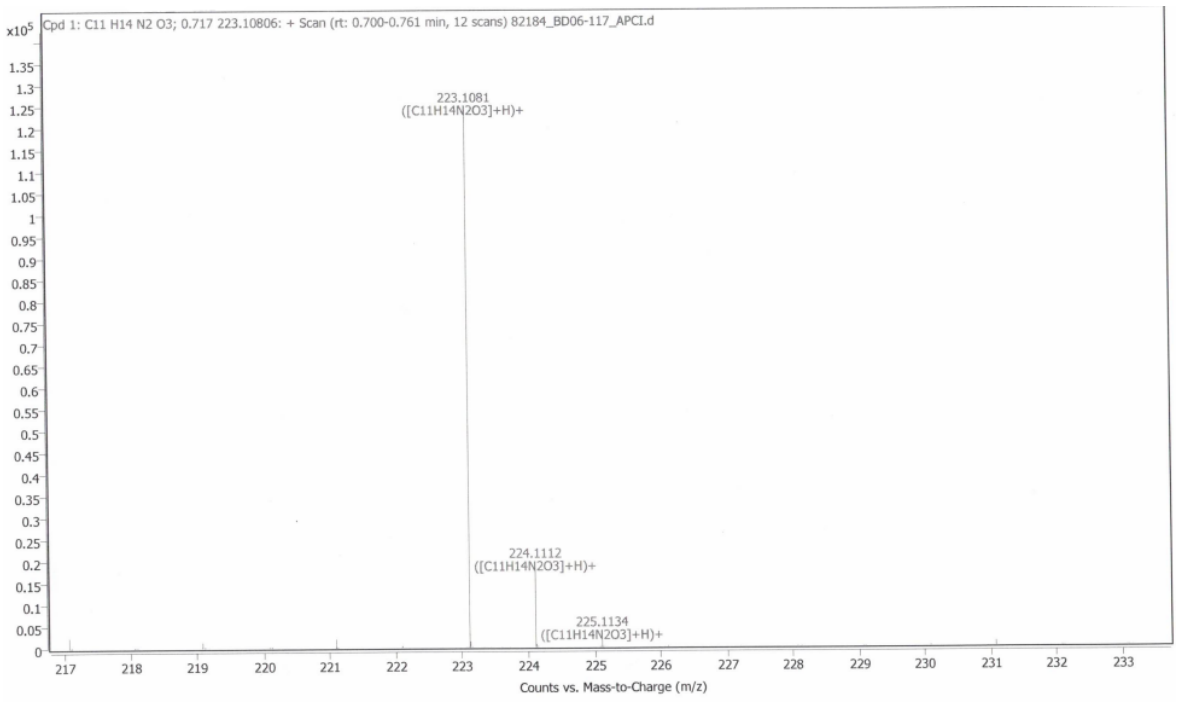

Cpd 1: C5 H10 N2 O3; 0.210 147.07693: + Scan (rt: 0.171-0.188, 0.204 ... min, 16 scans) 82087_BD06-113_APCI_0.5UL.d
x10^6
147.0769
([C5H10N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

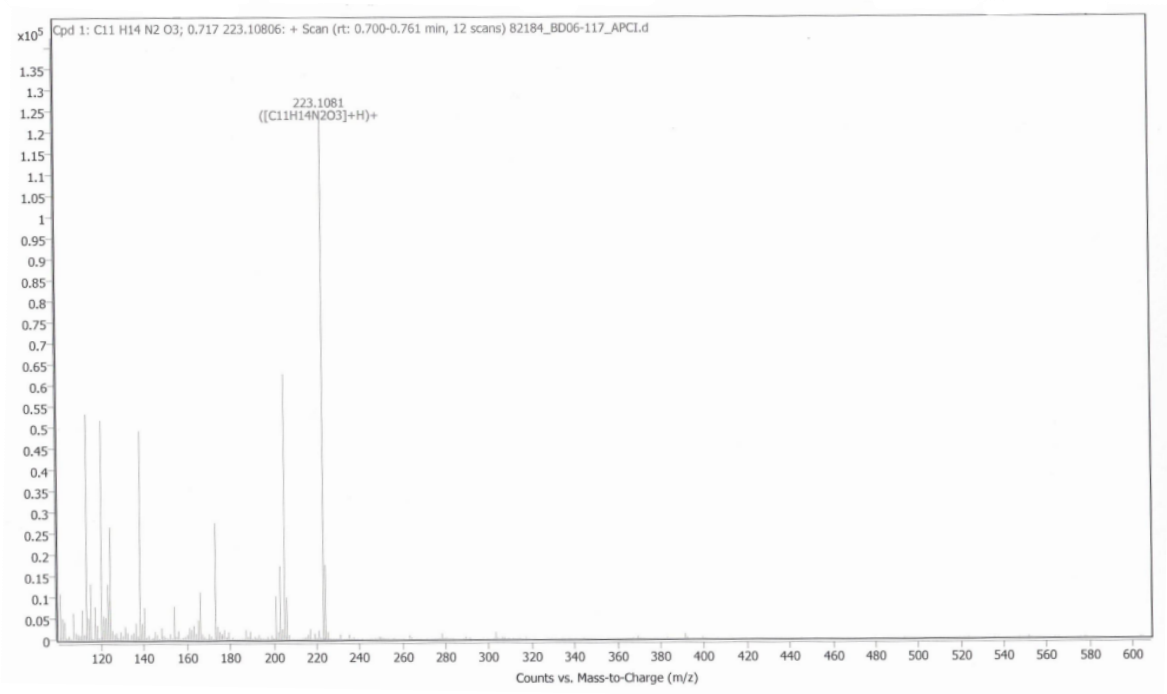

Cpd 1: C5 H10 N2 O3; 0.210 147.07693: + Scan (rt: 0.171-0.188, 0.204 ... min, 16 scans) 82087_BD06-113_APCI_0.5UL.d
x10^6
147.0769
([C5H10N2O3]+H)+
129.0660
([C5H10N2O3]+H)+[-H2O]
169.0609
([C5H10N2O3]+Na)+
Counts vs. Mass-to-Charge (m/z)

FG·TFA **D13**

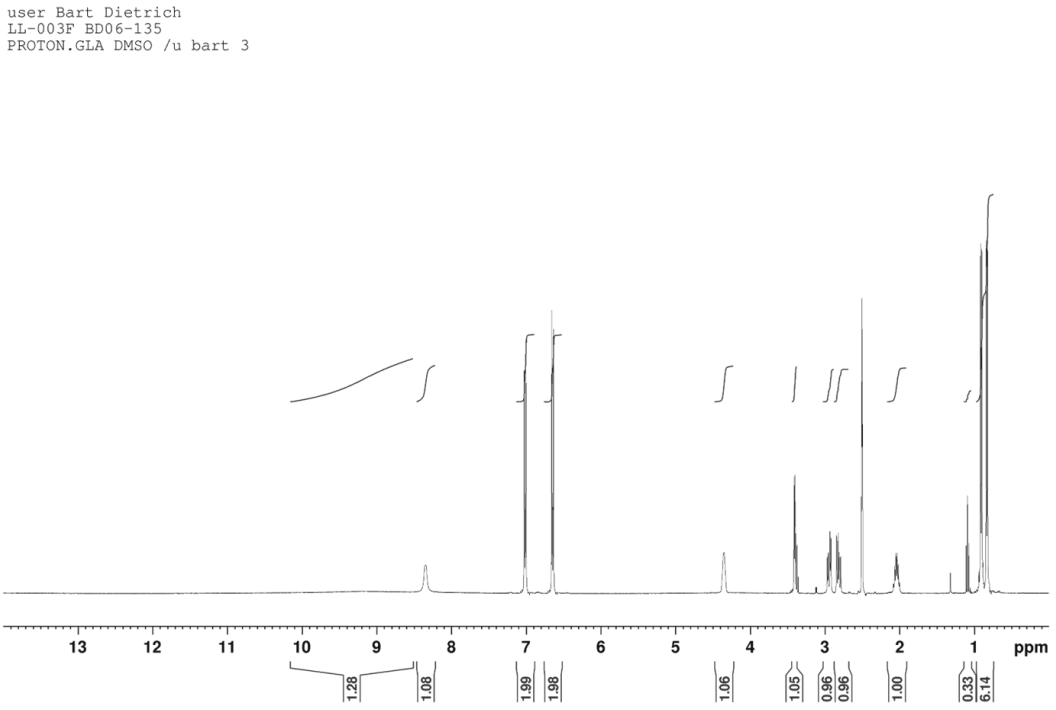

user Bart Dietrich
LH-001 BD06-117
PROTON.GLA DMSO /u bart 6
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
0.96
4.90
1.00
1.95
1.08
1.01
0.15

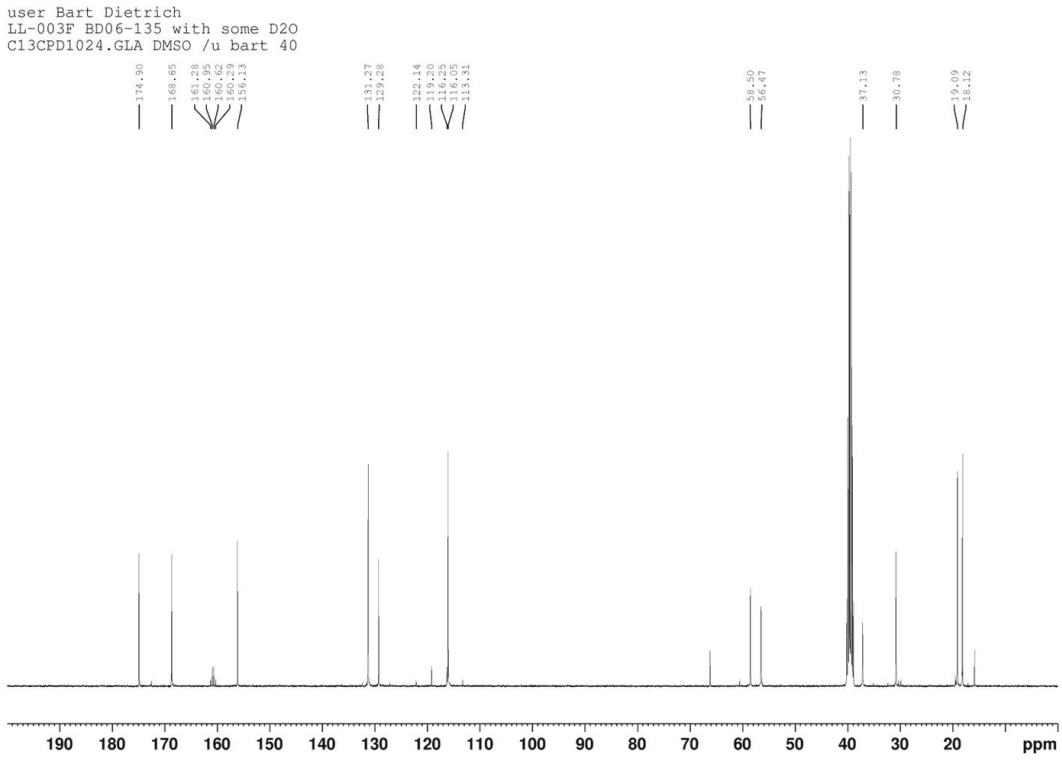

user Bart Dietrich
LH-001 BD06-117
C13CPD1024.GLA DMSO /u bart 52
170.66
168.52
158.90
158.59
158.28
157.97
134.91
129.66
128.54
127.19
121.70
116.73
115.75
112.78
53.36
40.84
37.04
210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 ppm

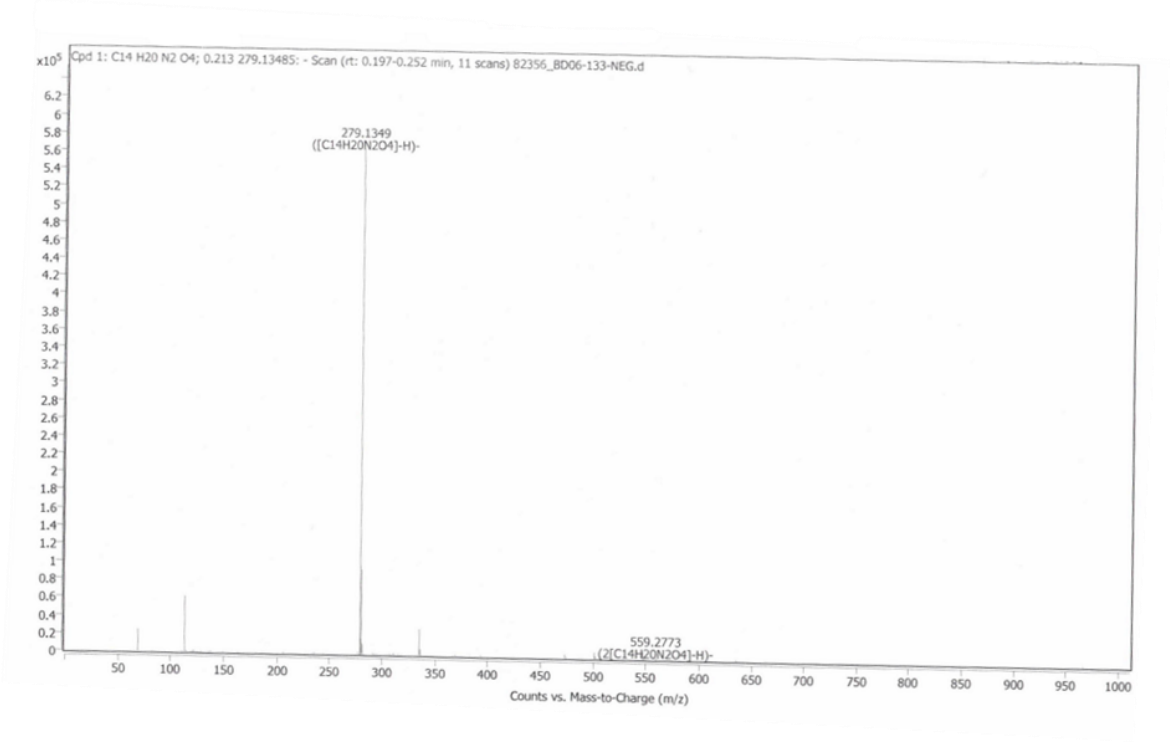

Cpd 1: C11 H14 N2 O3; 0.717 223.10806: + Scan (rt: 0.700-0.761 min, 12 scans) 82184_BD06-117_APCI.d
x10^5
223.1081
([C11H14N2O3]+H)+
224.1112
([C11H14N2O3]+H)+
225.1134
([C11H14N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

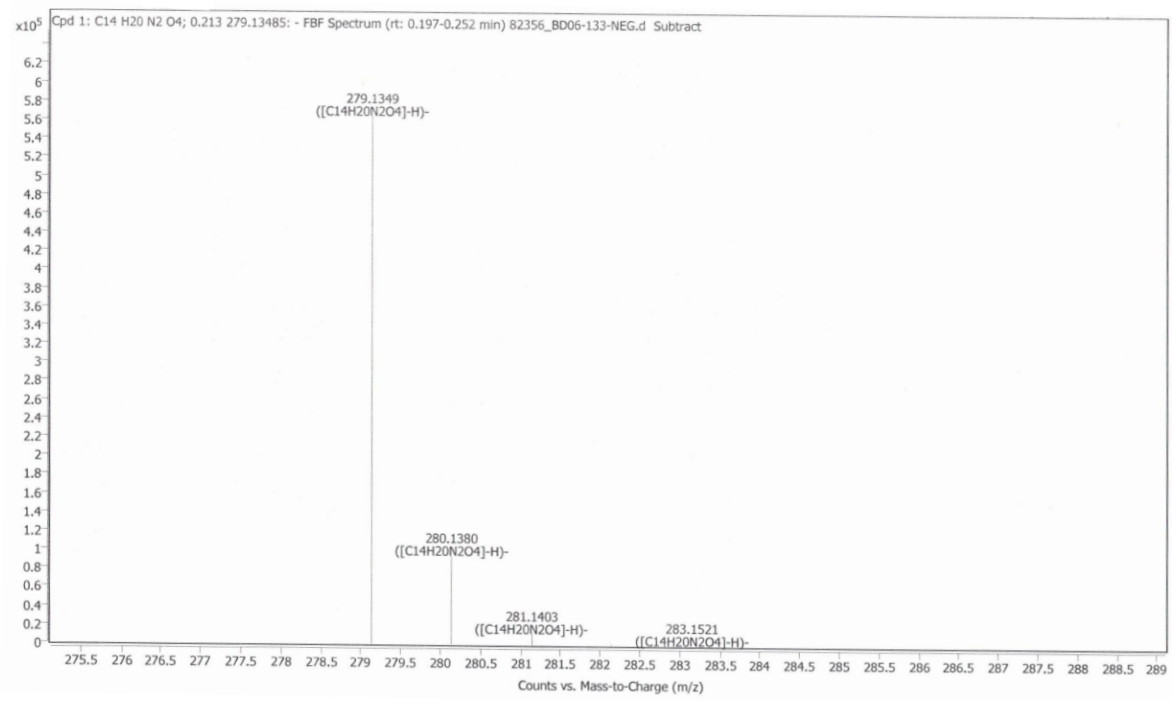

Cpd 1: C11 H14 N2 O3; 0.717 223.10806: + Scan (rt: 0.700-0.761 min, 12 scans) 82184_BD06-117_APCI.d
x10^5
223.1081
([C11H14N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

VY·TFA **D14**

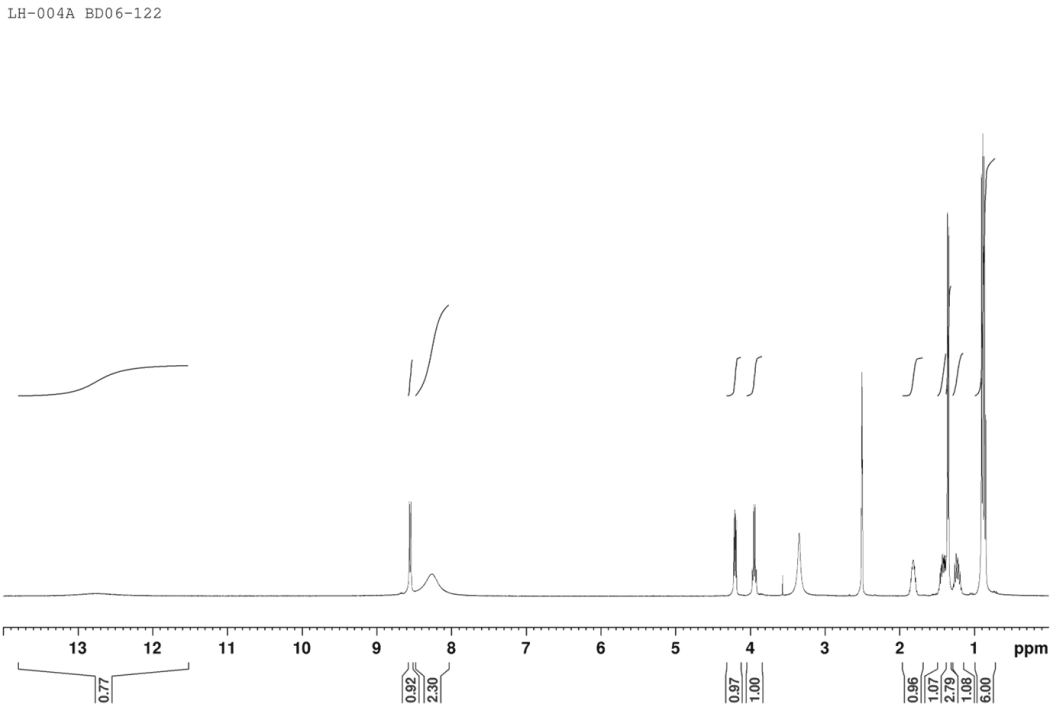

user Bart Dietrich
LL-003F BD06-135
PROTON.GLA DMSO /u bart 3
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
1.28 1.08 1.99 1.98 1.06 1.05 0.96 0.96 1.00 0.33 6.14

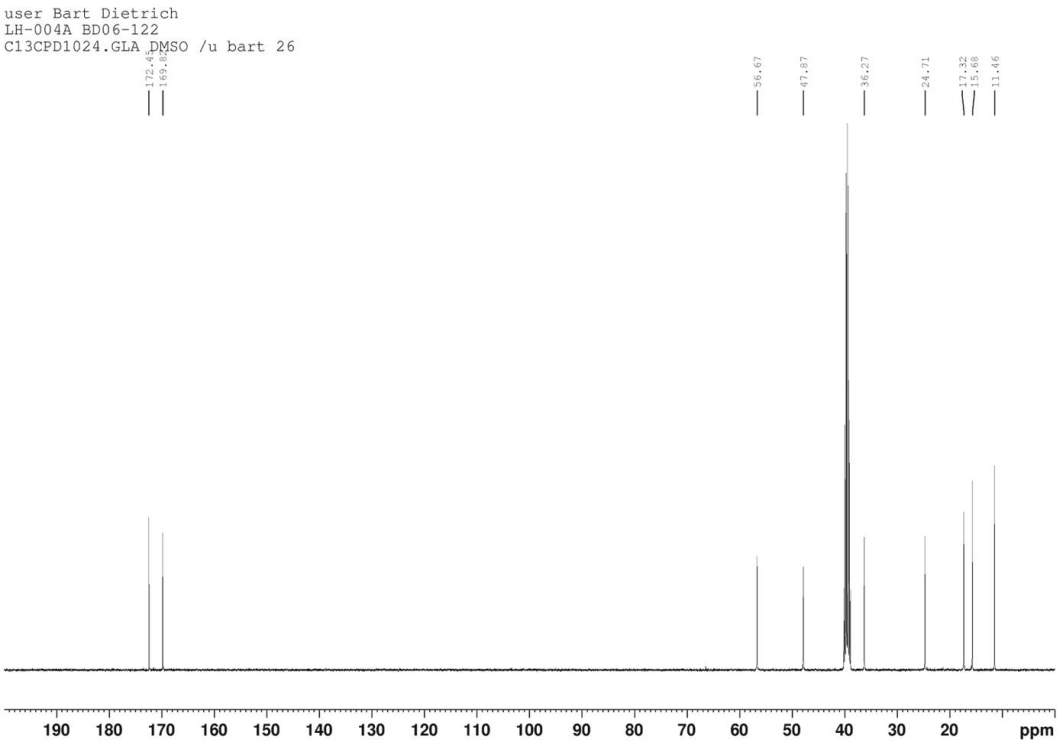

user Bart Dietrich
LL-003F BD06-135 with some D2O
C13CPD1024.GLA DMSO /u bart 40
174.90
168.65
161.28
160.95
160.62
160.29
156.13
131.27
129.28
122.14
119.20
116.35
116.05
113.31
58.50
56.47
37.13
30.78
19.09
18.12
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

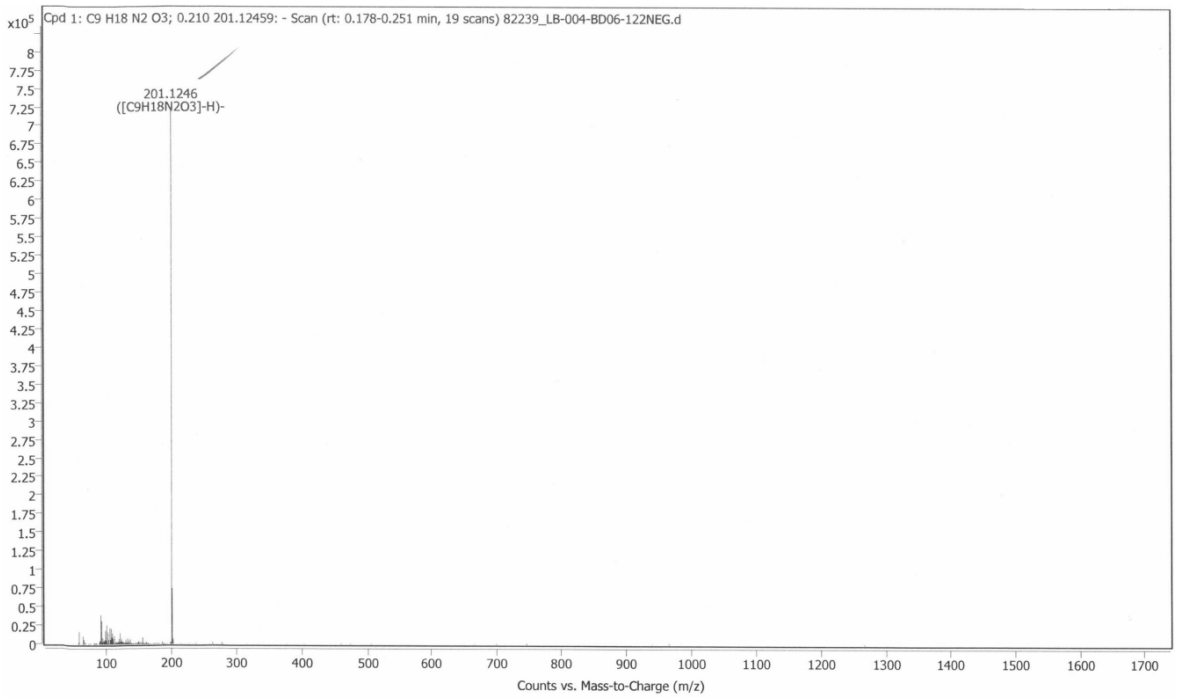

x10^5
Cpd 1: C14 H20 N2 O4; 0.213 279.13485: - Scan (rt: 0.197-0.252 min, 11 scans) 82356_BD06-133-NEG.d
279.1349
([C14H20N2O4]-H)-
559.2773
(2[C14H20N2O4]-H)-
Counts vs. Mass-to-Charge (m/z)

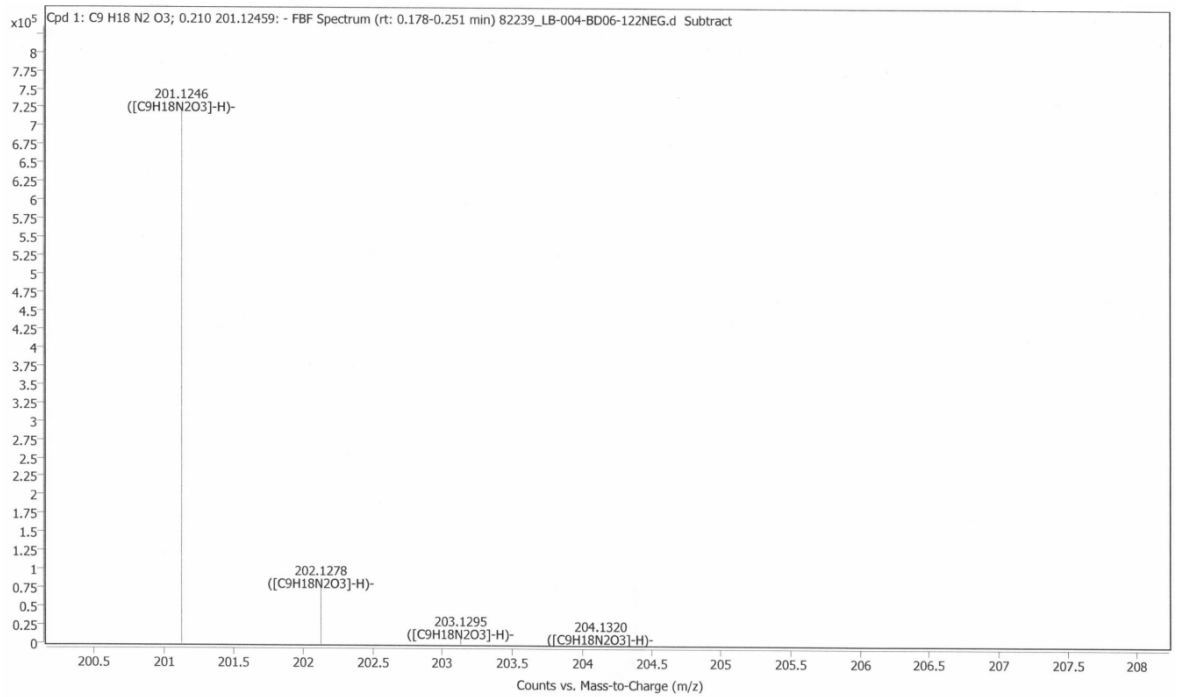

x10^5
Cpd 1: C14 H20 N2 O4; 0.213 279.13485: - FBF Spectrum (rt: 0.197-0.252 min) 82356_BD06-133-NEG.d Subtract
279.1349
([C14H20N2O4]-H)-
280.1380
([C14H20N2O4]-H)-
281.1403
([C14H20N2O4]-H)-
283.1521
([C14H20N2O4]-H)-
Counts vs. Mass-to-Charge (m/z)

AI·HCl **D15**

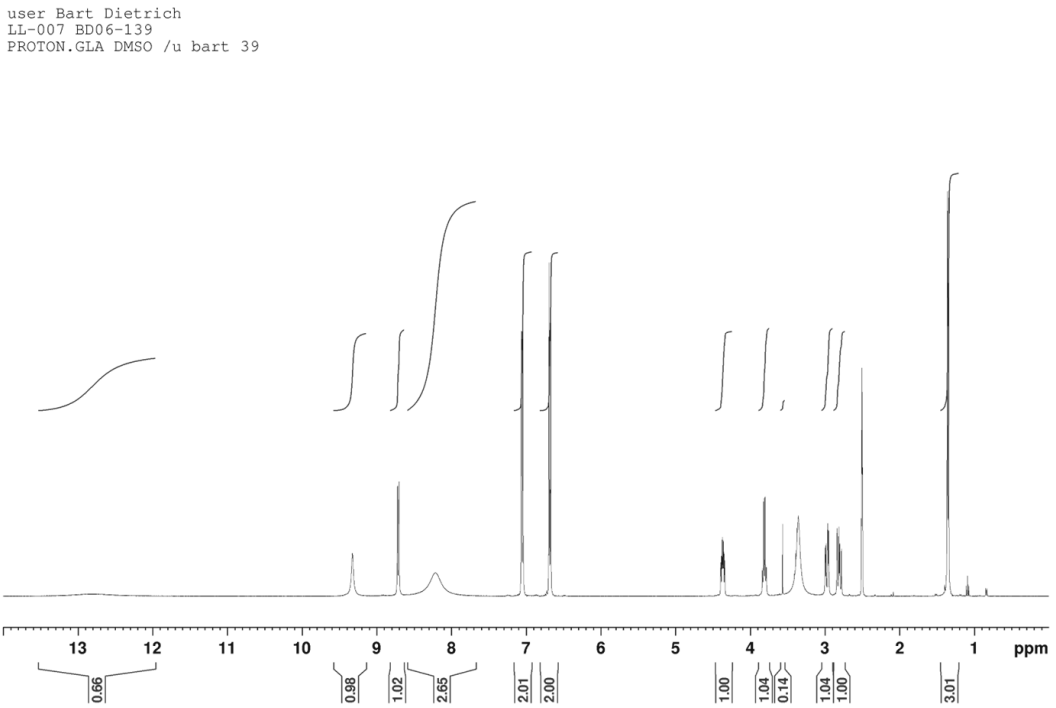

LH-004A BD06-122
13
12
11
10
9
8
7
6
5
4
3
2
1
ppm
0.77
0.92
2.30
0.97
1.00
0.96
1.07
2.79
1.08
6.00

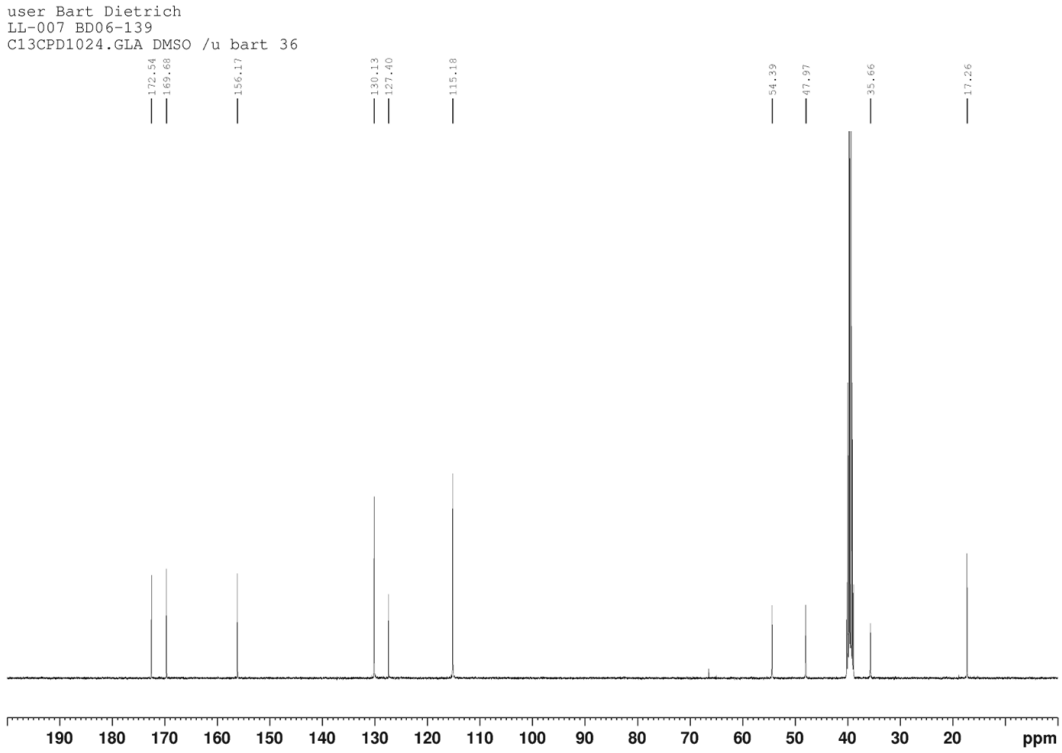

user Bart Dietrich
LH-004A BD06-122
C13CPD1024.GLA DMSO /u bart 26
172.41
169.82
56.67
47.87
36.27
24.71
17.32
15.68
11.46
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
ppm

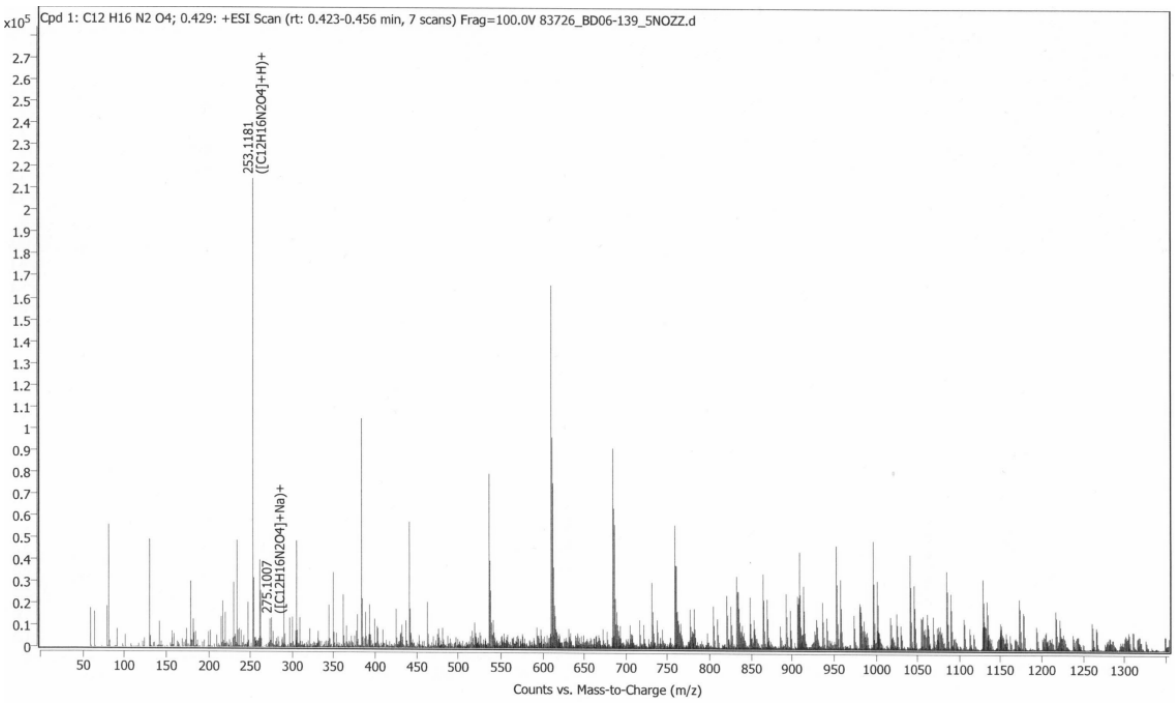

Cpd 1: C9 H18 N2 O3; 0.210 201.12459: - Scan (rt: 0.178-0.251 min, 19 scans) 82239_LB-004-BD06-122NEG.d
x10^5
201.1246
([C9H18N2O3]-H)-
Counts vs. Mass-to-Charge (m/z)

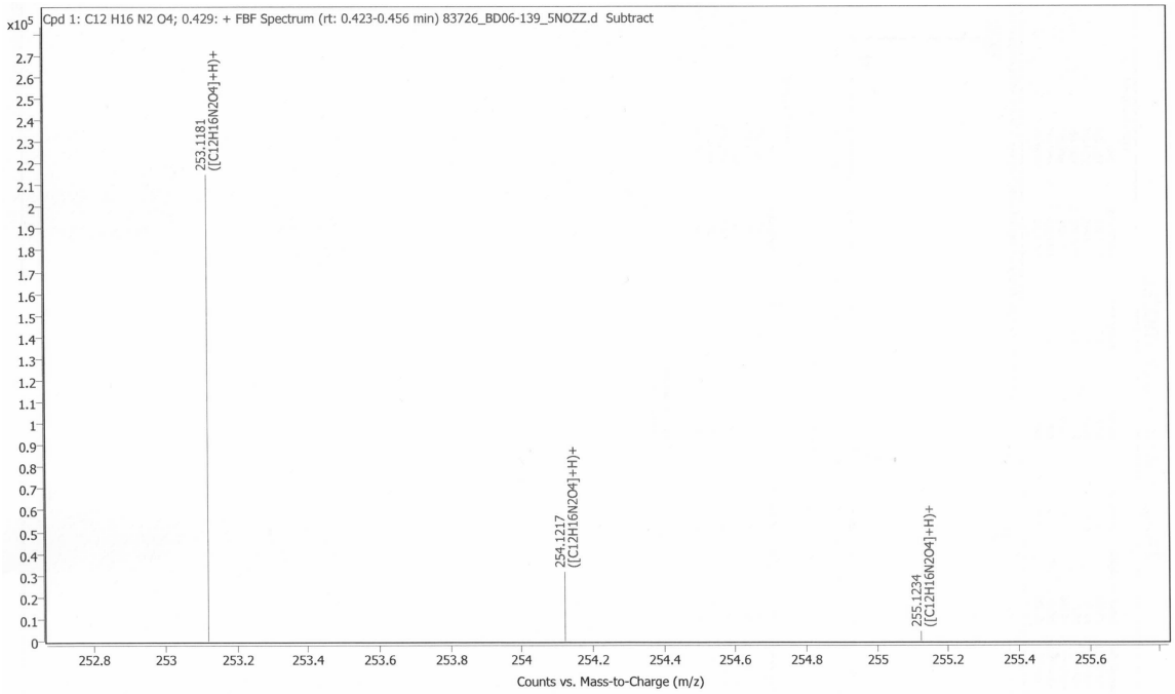

Cpd 1: C9 H18 N2 O3; 0.210 201.12459: - FBF Spectrum (rt: 0.178-0.251 min) 82239_LB-004-BD06-122NEG.d Subtract
x10^5
201.1246
([C9H18N2O3]-H)-
202.1278
([C9H18N2O3]-H)-
203.1295
([C9H18N2O3]-H)-
204.1320
([C9H18N2O3]-H)-
Counts vs. Mass-to-Charge (m/z)

AY·HCl **D16**

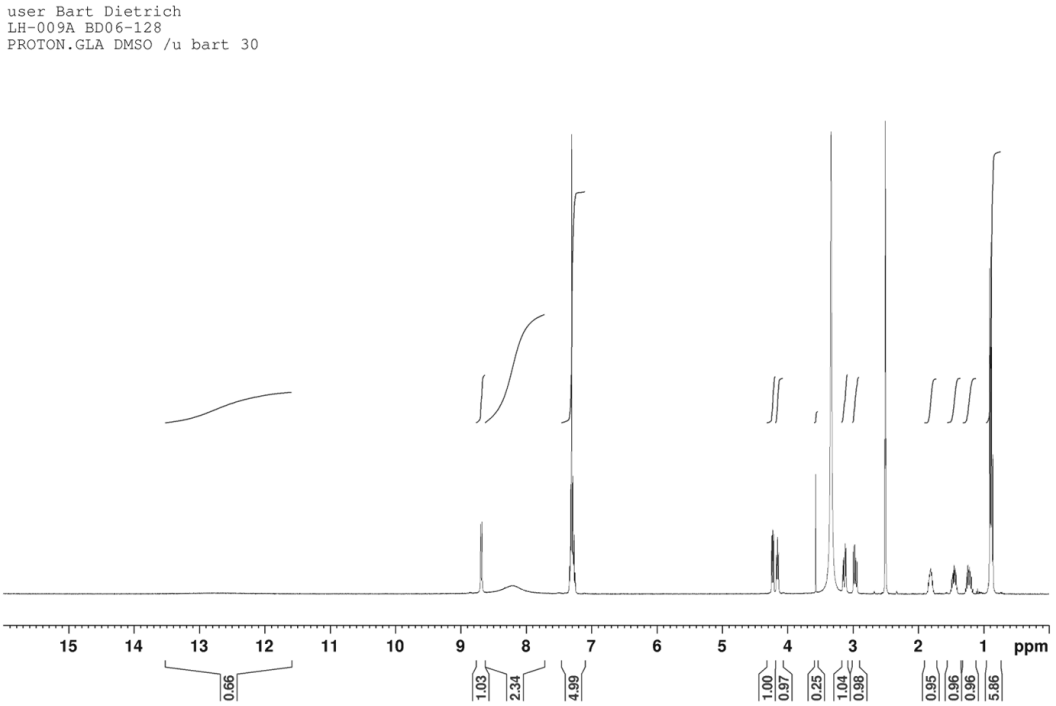

user Bart Dietrich
LL-007 BD06-139
PROTON.GLA DMSO /u bart 39
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
0.66
0.98
1.02
2.65
2.01
2.00
1.00
1.04
0.14
1.04
1.00
3.01

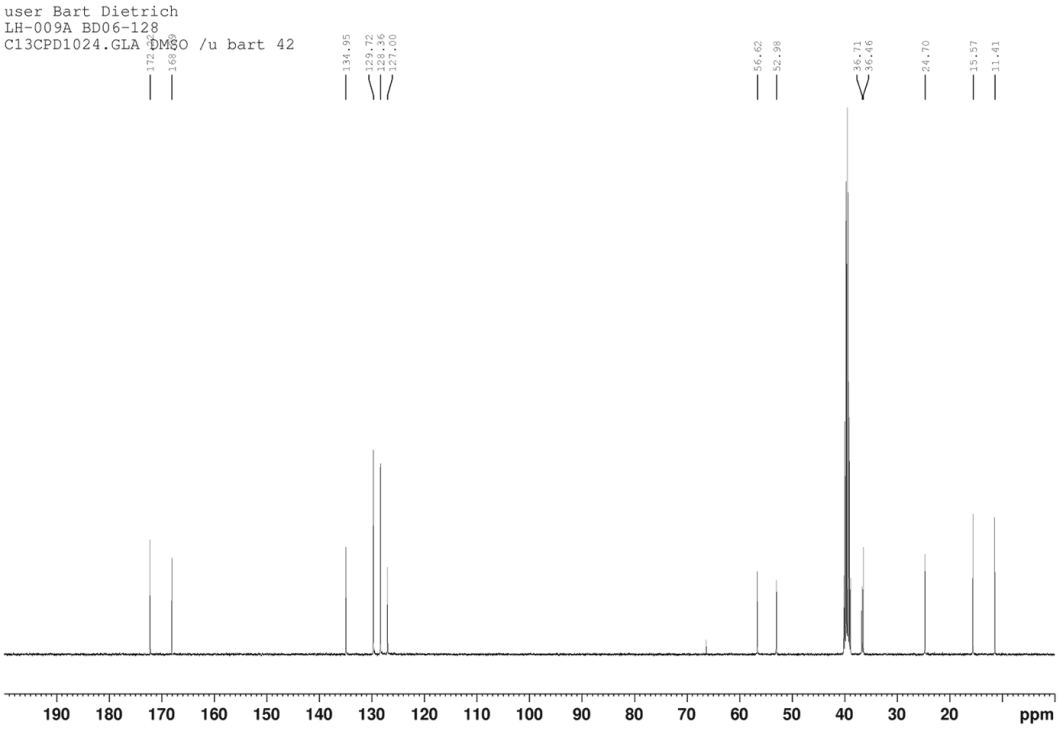

user Bart Dietrich
LL-007 BD06-139
C13CPD1024.GLA DMSO /u bart 36
172.54
169.68
156.17
130.13
127.40
115.18
54.39
47.97
35.66
17.26
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

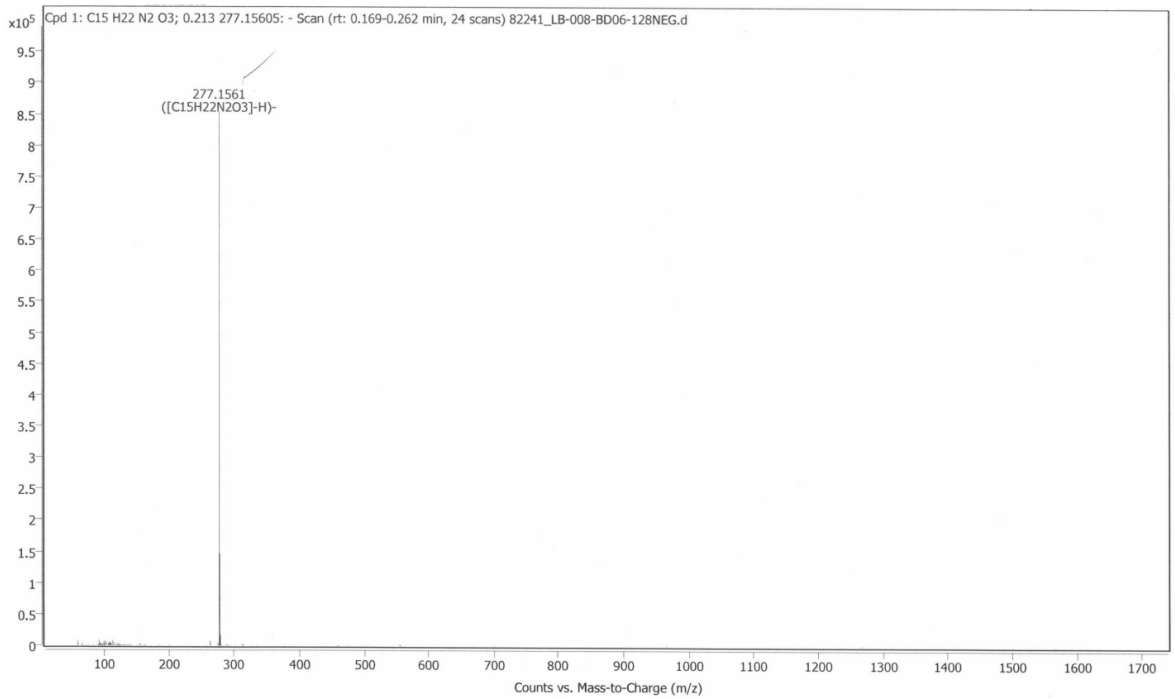

Cpd 1: C12 H16 N2 O4; 0.429: +ESI Scan (rt: 0.423-0.456 min, 7 scans) Frag=100.0V 83726_BD06-139_5NOZZ.d
x10^5
253.1181
([C12H16N2O4]+H)+
275.1007
([C12H16N2O4]+Na)+
Counts vs. Mass-to-Charge (m/z)

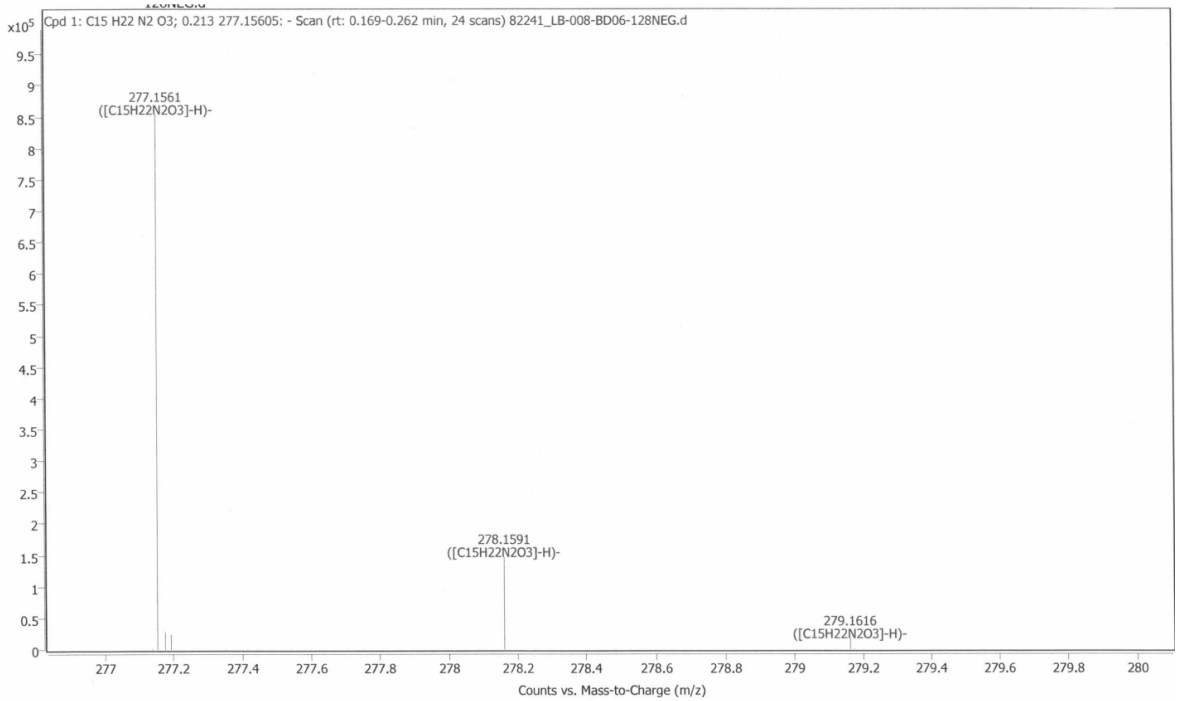

Cpd 1: C12 H16 N2 O4; 0.429: + FBF Spectrum (rt: 0.423-0.456 min) 83726_BD06-139_5NOZZ.d Subtract
x10^5
253.1181
([C12H16N2O4]+H)+
254.1217
([C12H16N2O4]+H)+
255.1234
([C12H16N2O4]+H)+
Counts vs. Mass-to-Charge (m/z)

FI·HCl **D17**

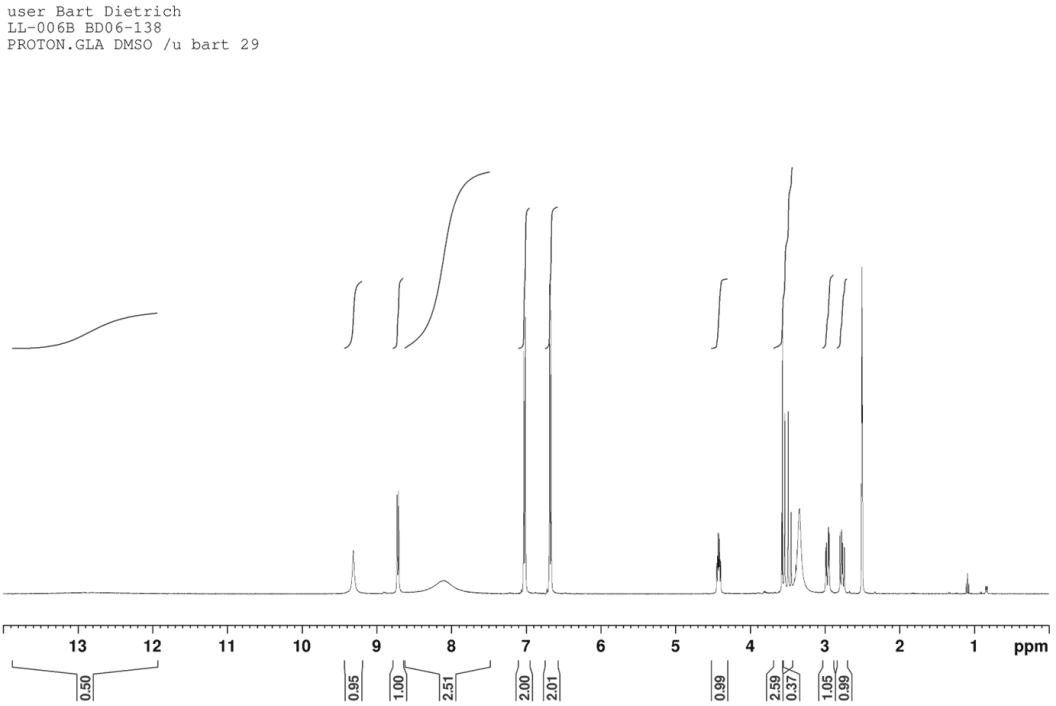

user Bart Dietrich
LH-009A BD06-128
PROTON.GLA DMSO /u bart 30
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
0.66
1.03
2.34
4.99
1.00
0.97
0.25
1.04
0.98
0.95
0.96
0.96
5.86

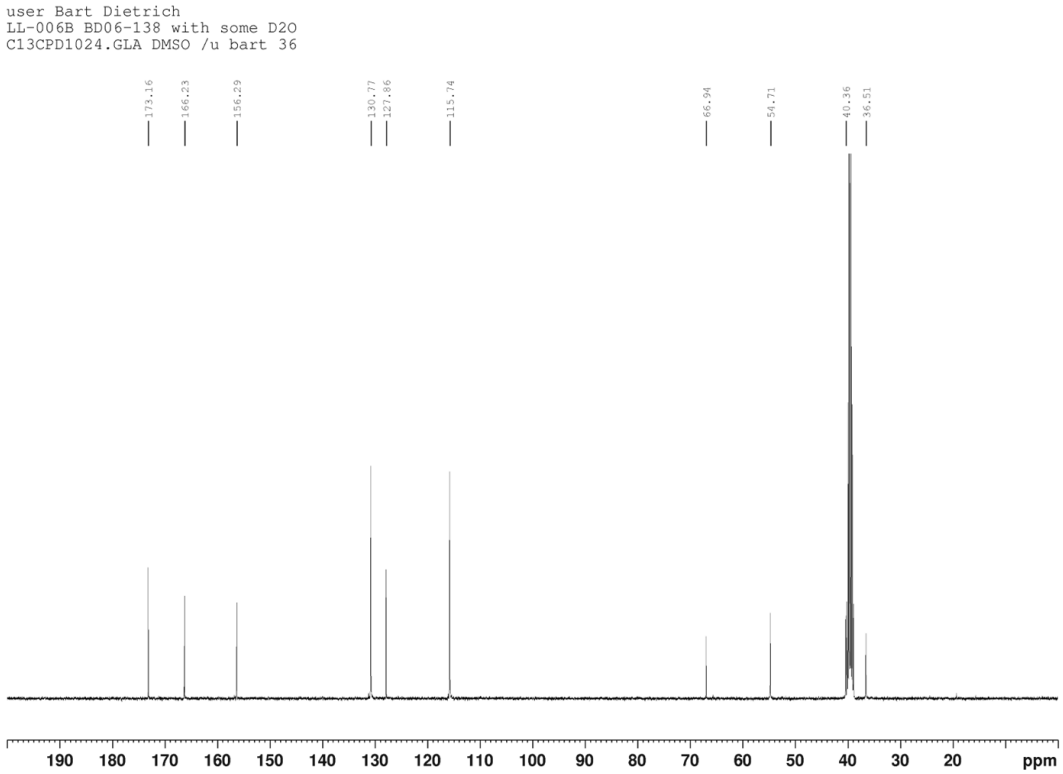

user Bart Dietrich
LH-009A BD06-128
C13CPD1024.GLA DMSO /u bart 42
172.03
168.83
134.95
129.72
128.36
127.00
56.62
52.98
36.71
36.46
24.70
15.57
11.41
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

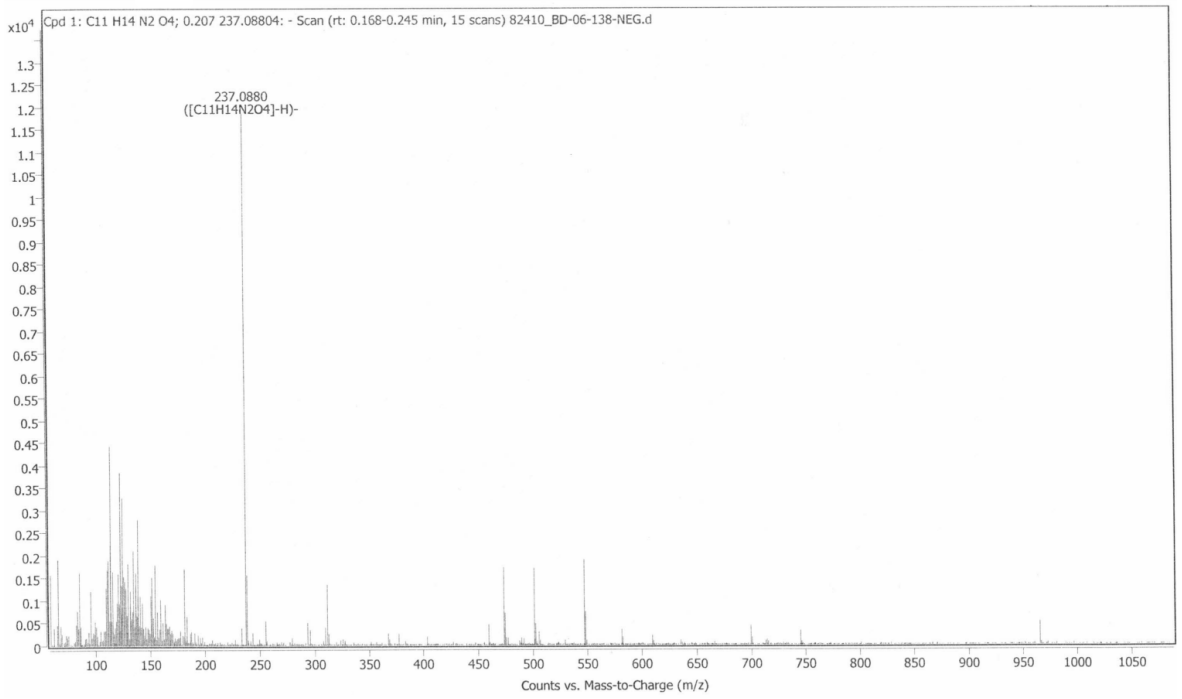

Cpd 1: C15 H22 N2 O3; 0.213 277.15605: - Scan (rt: 0.169-0.262 min, 24 scans) 82241_LB-008-BD06-128NEG.d
x10^5
277.1561
([C15H22N2O3]-H)-
Counts vs. Mass-to-Charge (m/z)

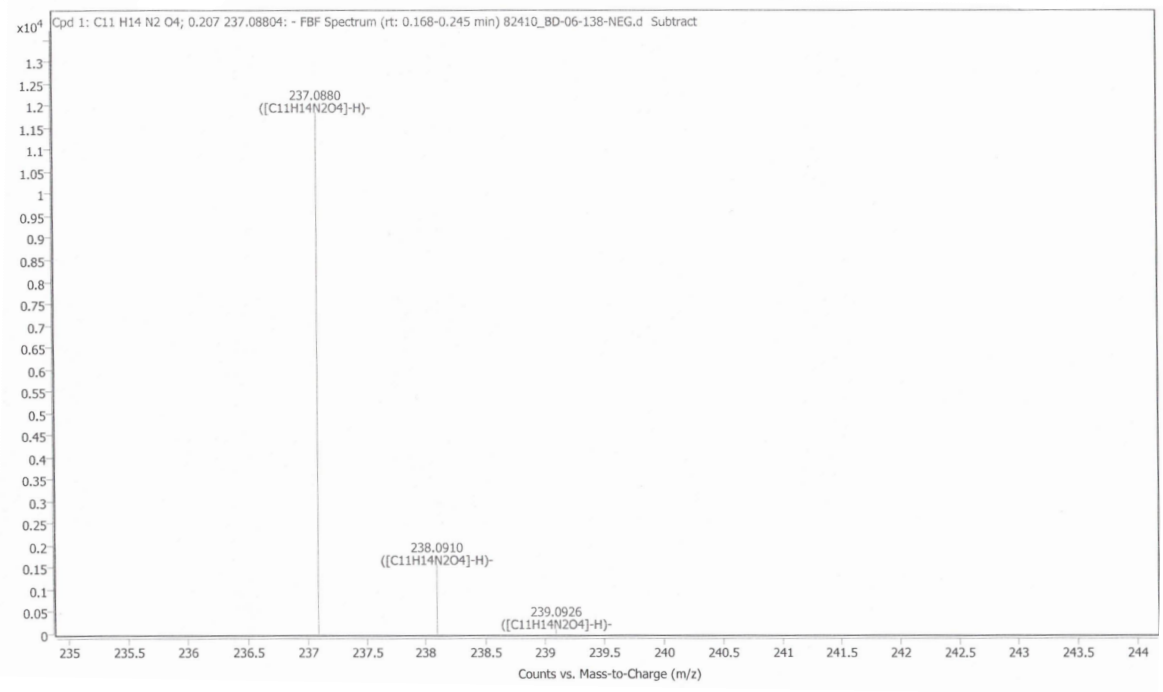

Cpd 1: C15 H22 N2 O3; 0.213 277.15605: - Scan (rt: 0.169-0.262 min, 24 scans) 82241_LB-008-BD06-128NEG.d
x10^5
277.1561
([C15H22N2O3]-H)-
278.1591
([C15H22N2O3]-H)-
279.1616
([C15H22N2O3]-H)-
Counts vs. Mass-to-Charge (m/z)

GY·HCl **D18**

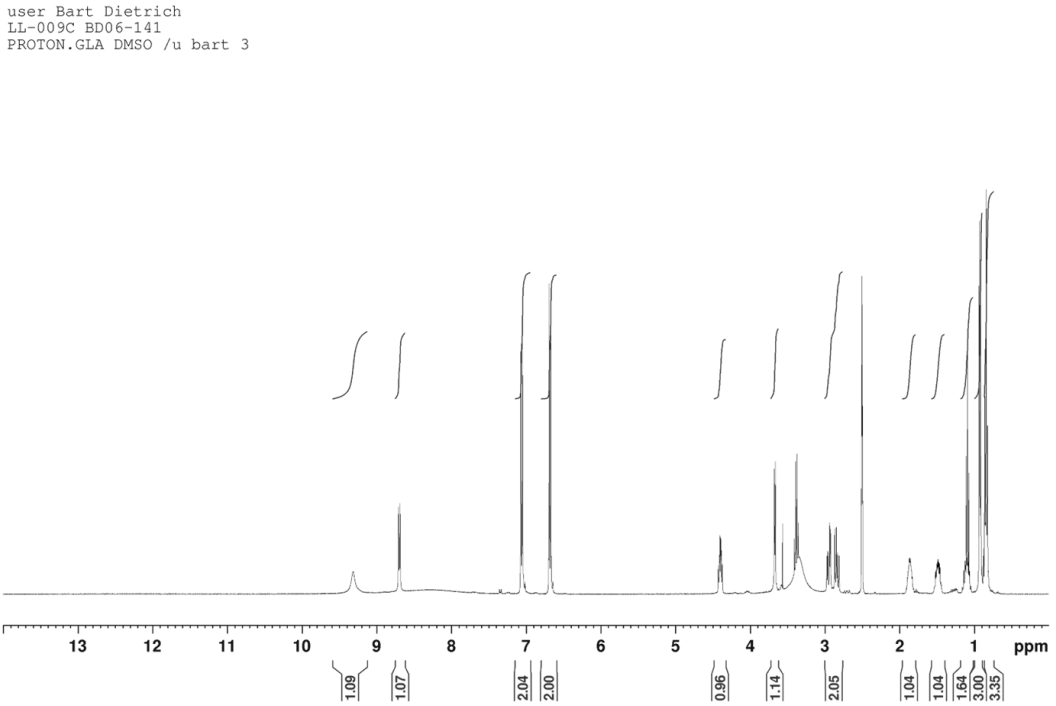

user Bart Dietrich
LL-006B BD06-138
PROTON.GLA DMSO /u bart 29
13
12
11
10
9
8
7
6
5
4
3
2
1
ppm
0.50
0.95
1.00
2.51
2.00
2.01
0.99
2.59
0.37
1.05
0.99

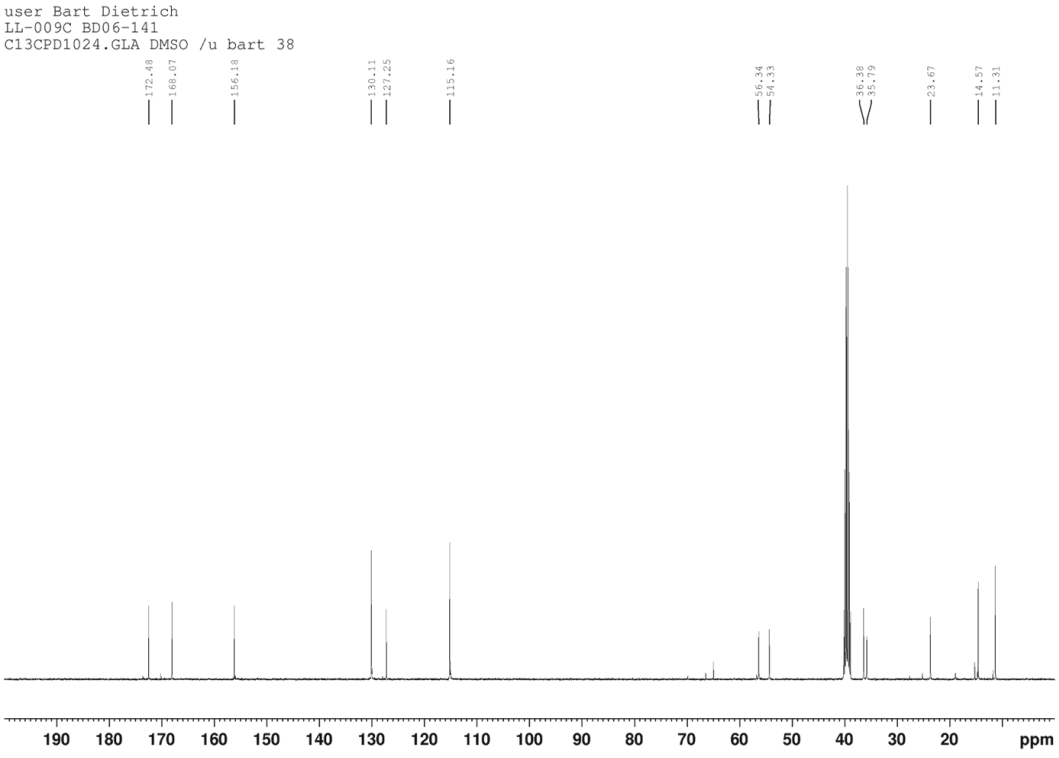

user Bart Dietrich
LL-006B BD06-138 with some D2O
C13CPD1024.GLA DMSO /u bart 36
173.16
166.23
156.29
130.77
127.86
115.74
66.94
54.71
40.36
36.51
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
ppm

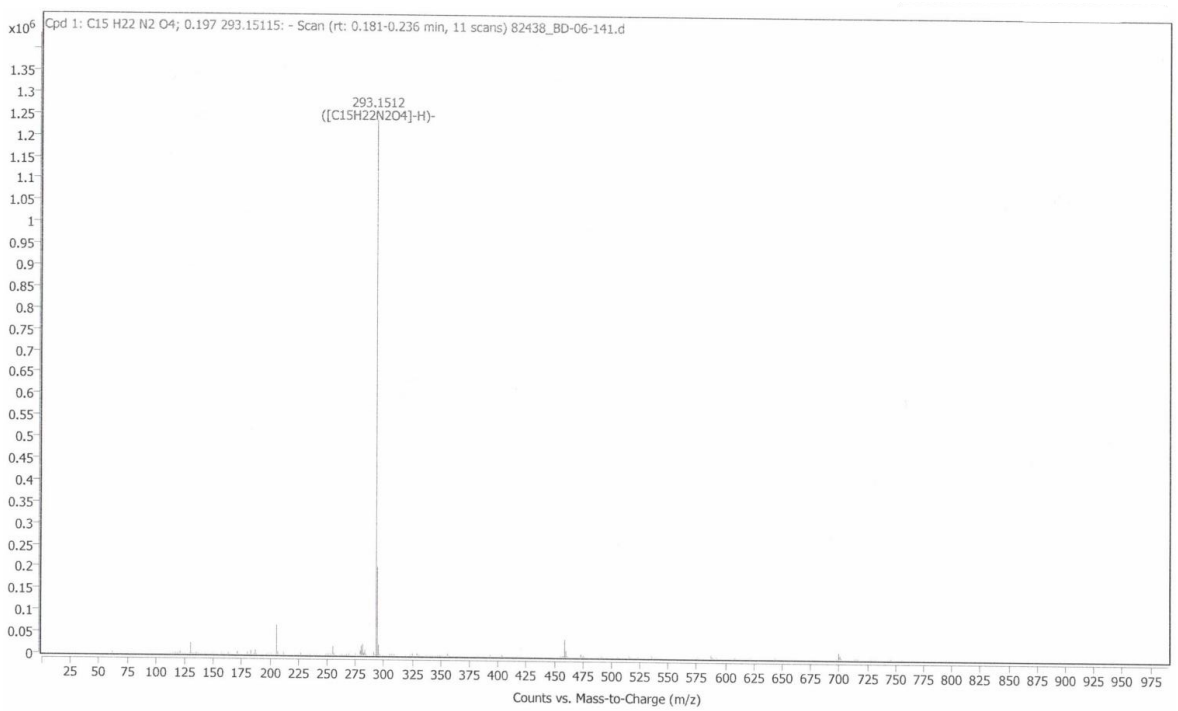

Cpd 1: C11 H14 N2 O4; 0.207 237.08804: - Scan (rt: 0.168-0.245 min, 15 scans) 82410_BD-06-138-NEG.d
x10^4
237.0880
([C11H14N2O4]-H)-
Counts vs. Mass-to-Charge (m/z)

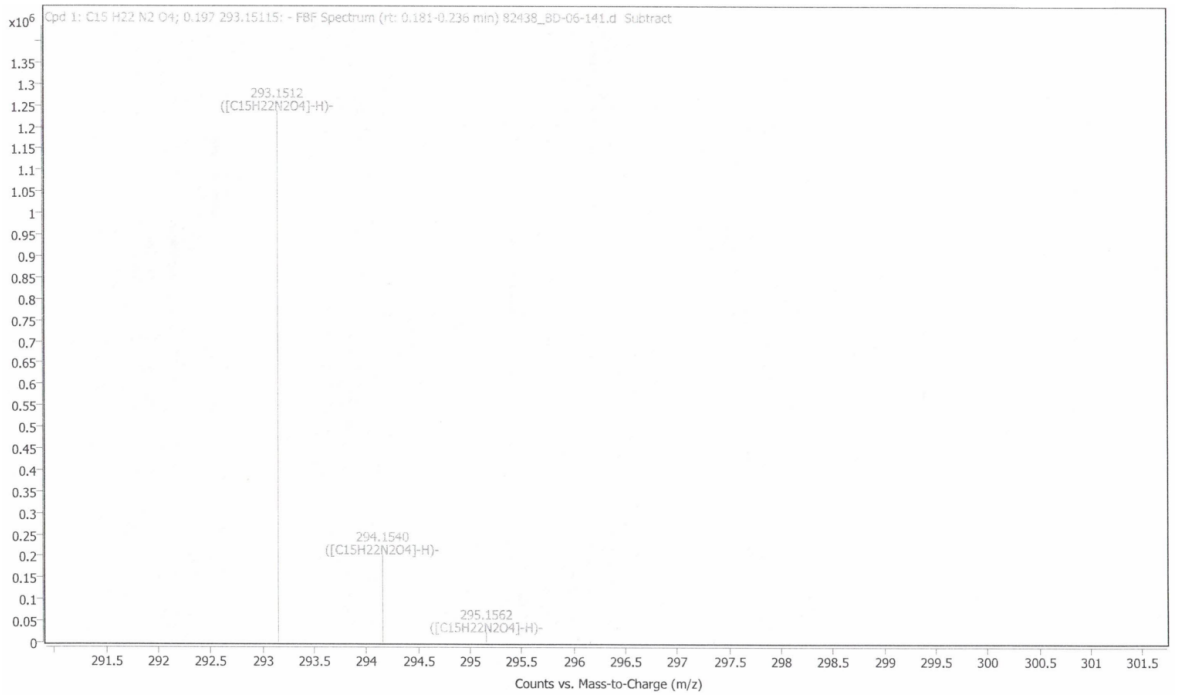

Cpd 1: C11 H14 N2 O4; 0.207 237.08804: - FBF Spectrum (rt: 0.168-0.245 min) 82410_BD-06-138-NEG.d Subtract
x10^4
237.0880
([C11H14N2O4]-H)-
238.0910
([C11H14N2O4]-H)-
239.0926
([C11H14N2O4]-H)-
Counts vs. Mass-to-Charge (m/z)

IY·HCl **D19**

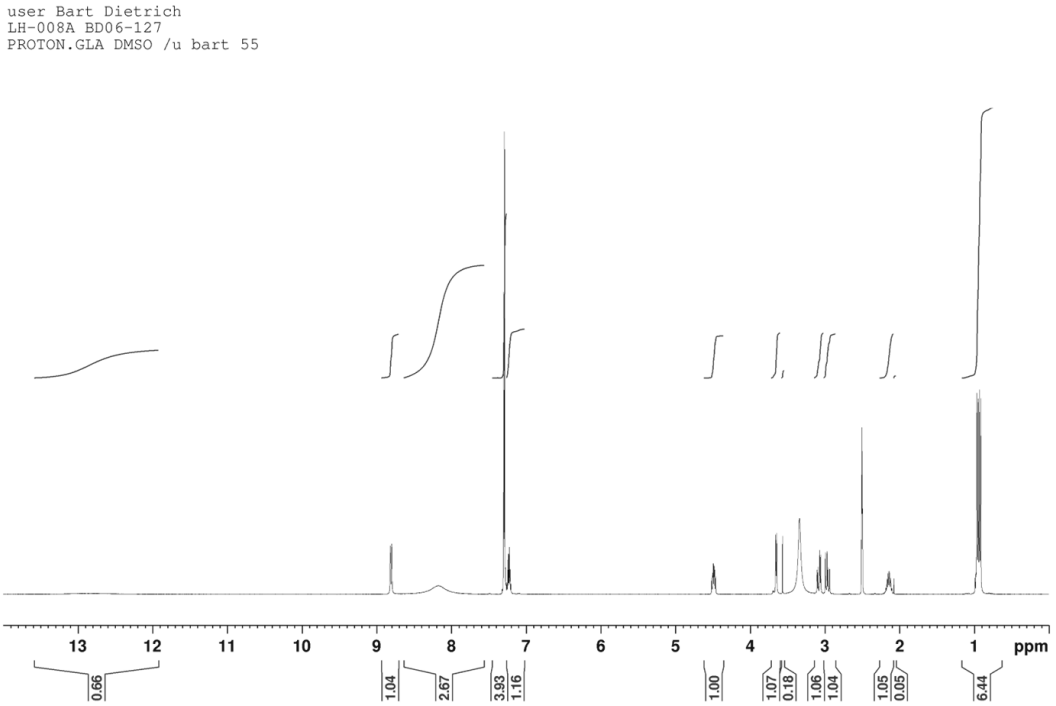

user Bart Dietrich
LL-009C BD06-141
PROTON.GLA DMSO /u bart 3
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
1.09 1.07 2.04 2.00 0.96 1.14 2.05 1.04 1.04 1.64 3.00 3.35

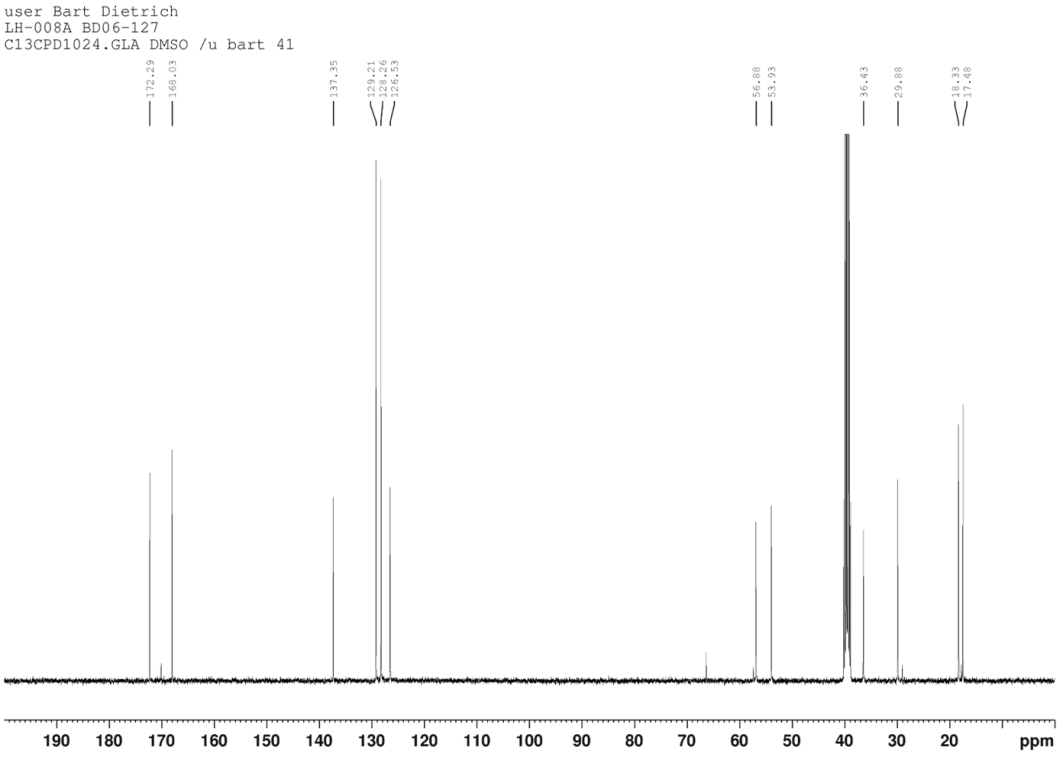

user Bart Dietrich
LL-009C BD06-141
C13CPD1024.GLA DMSO /u bart 38
172.48 168.07 156.18 130.11 127.25 115.16 56.34 54.33 36.38 35.79 23.67 14.57 11.31
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

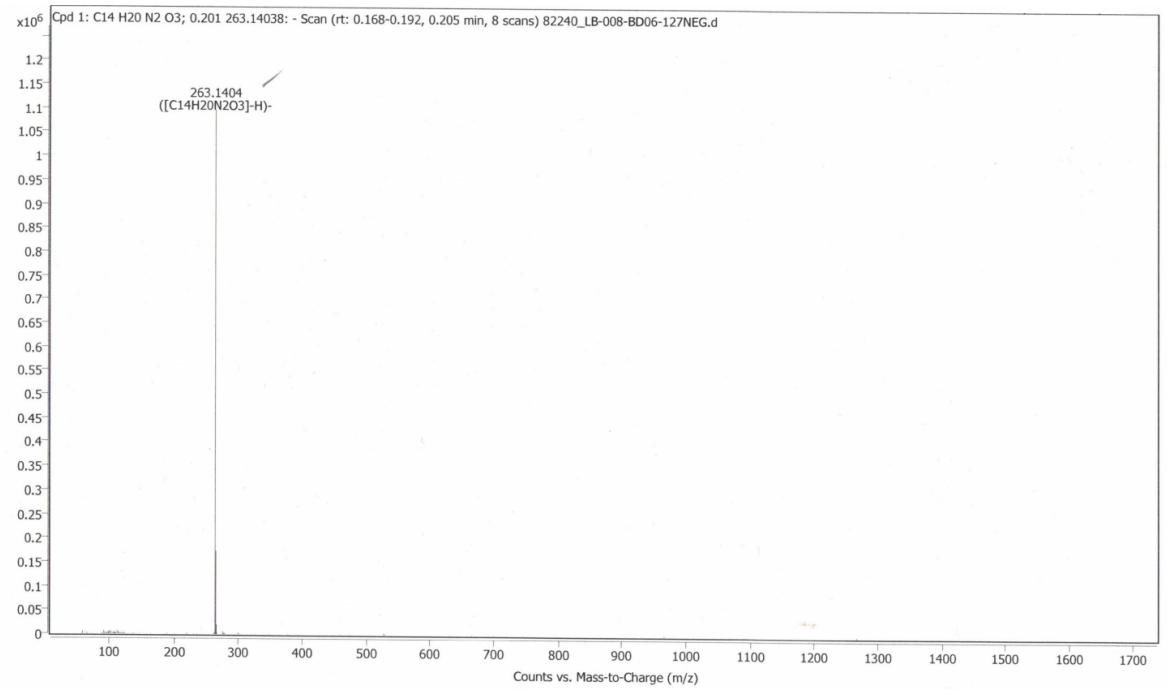

Cpd 1: C15 H22 N2 O4; 0.197 293.15115: - Scan (rt: 0.181-0.236 min, 11 scans) 82438_BD-06-141.d
x10^6
293.1512
([C15H22N2O4]-H)-
Counts vs. Mass-to-Charge (m/z)

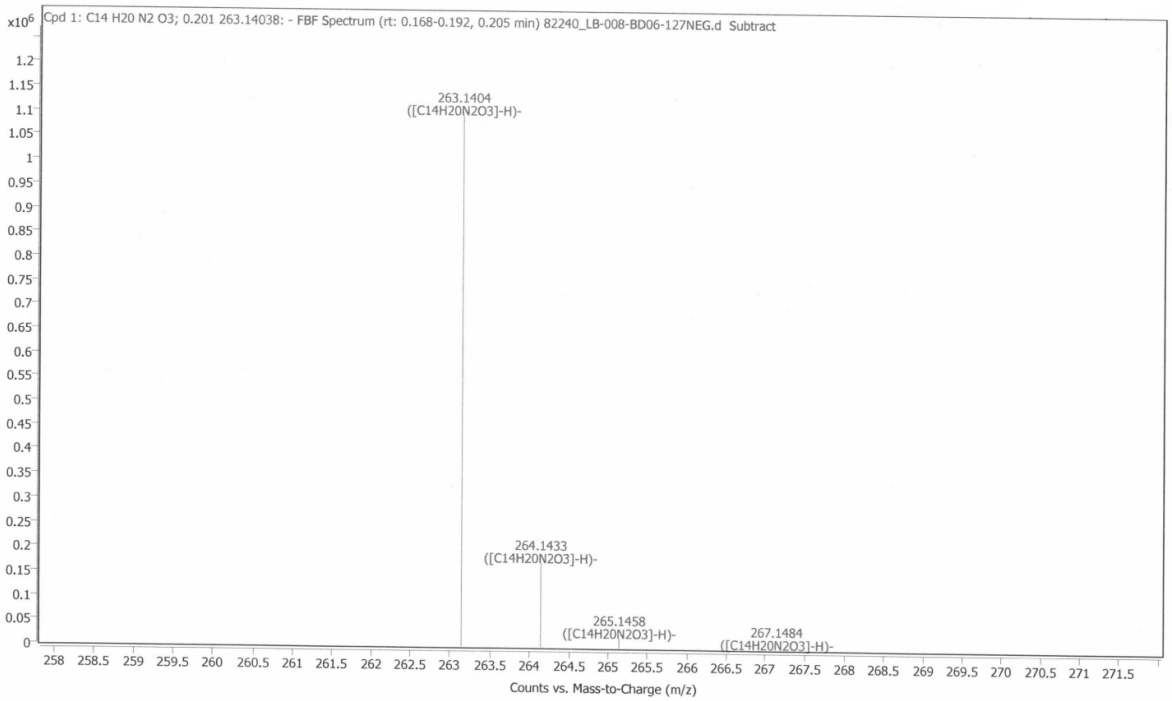

Cpd 1: C15 H22 N2 O4; 0.197 293.15115: - FBF Spectrum (rt: 0.181-0.236 min) 82438_BD-06-141.d Subtract
x10^6
293.1512
([C15H22N2O4]-H)-
294.1540
([C15H22N2O4]-H)-
295.1562
([C15H22N2O4]-H)-
Counts vs. Mass-to-Charge (m/z)

VF·HCl **D20**

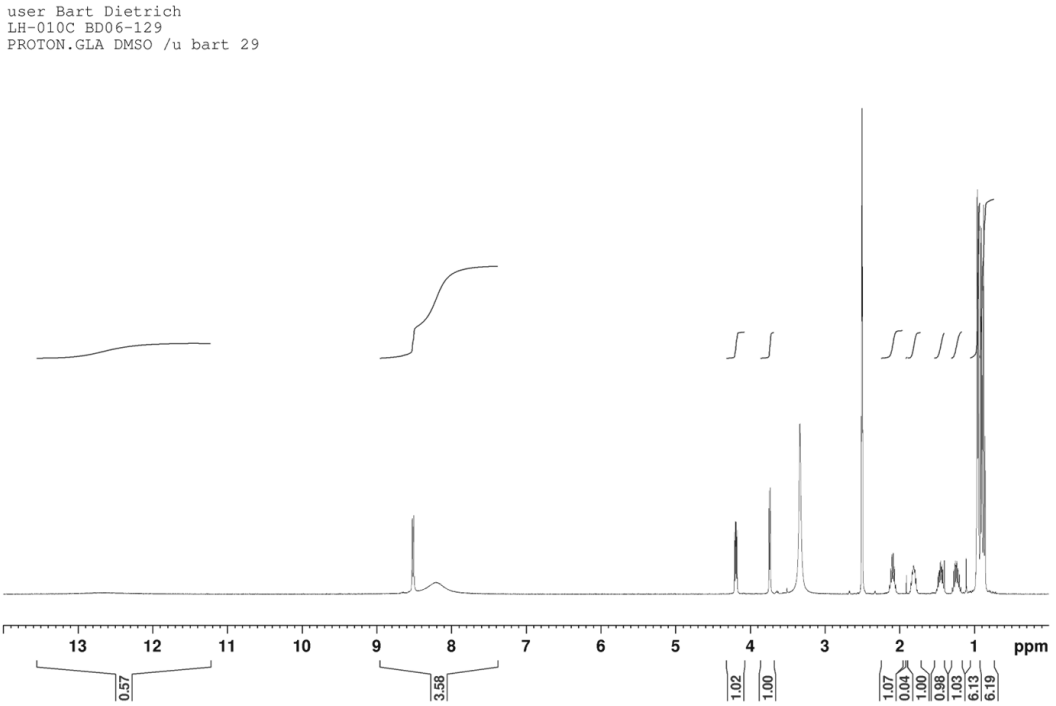

user Bart Dietrich
LH-008A BD06-127
PROTON.GLA DMSO /u bart 55
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
0.66
1.04
2.67
3.93
1.16
1.00
1.07
0.18
1.06
1.04
1.05
0.05
6.44

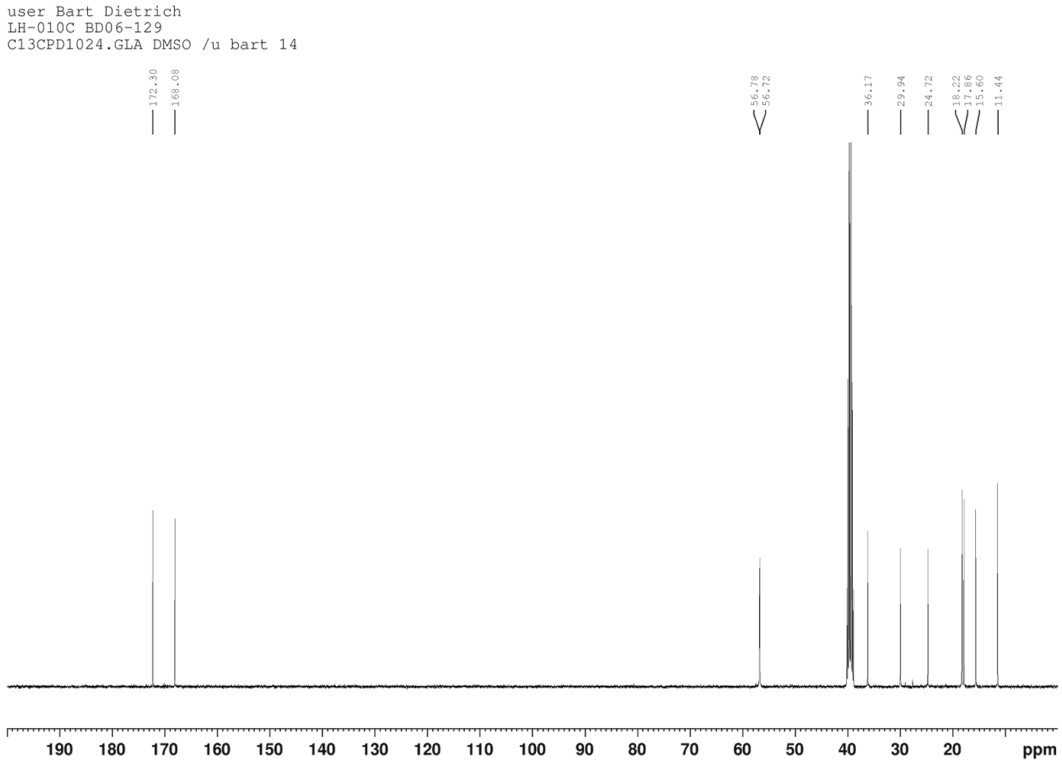

user Bart Dietrich
LH-008A BD06-127
C13CPD1024.GLA DMSO /u bart 41
172.29
168.03
137.35
129.21
128.26
126.53
56.88
53.93
36.43
29.88
18.33
17.48
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

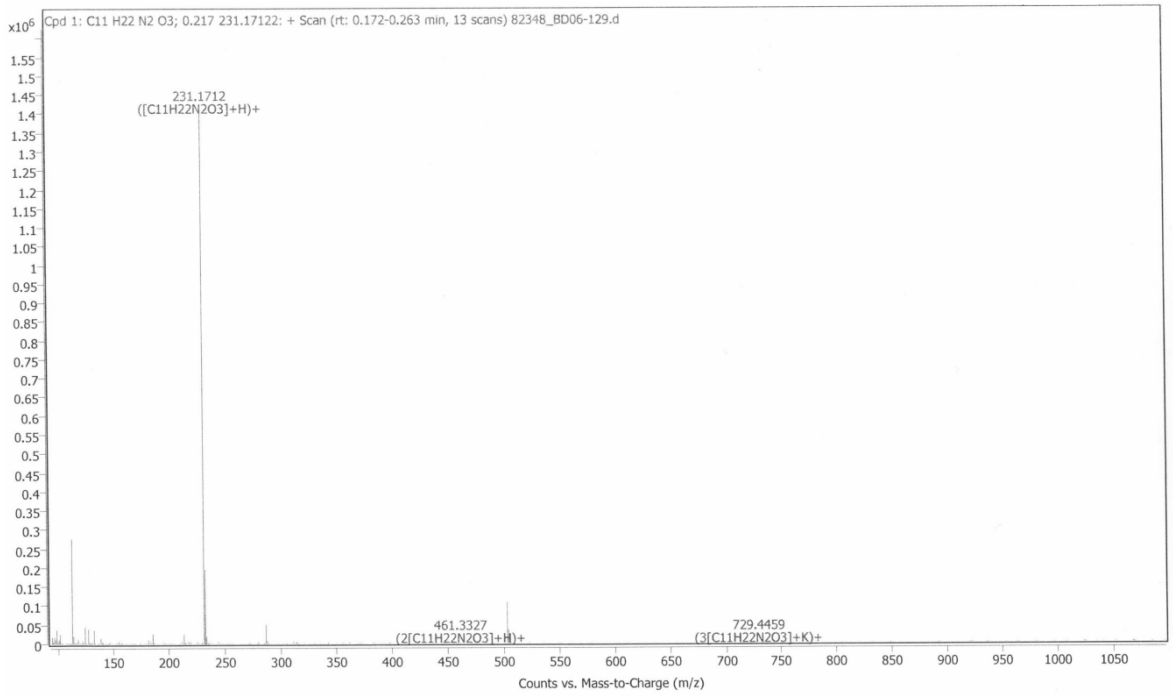

x10^6
Cpd 1: C14 H20 N2 O3; 0.201 263.14038: - Scan (rt: 0.168-0.192, 0.205 min, 8 scans) 82240_LB-008-BD06-127NEG.d
263.1404
([C14H20N2O3]-H)-
Counts vs. Mass-to-Charge (m/z)

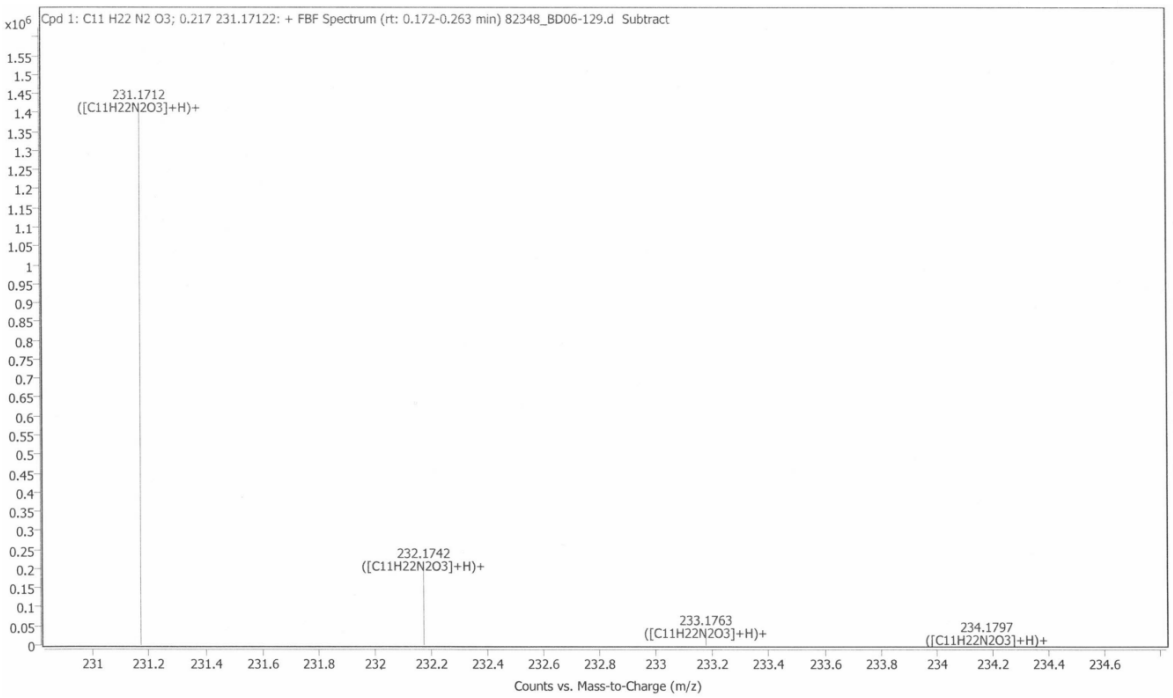

x10^6
Cpd 1: C14 H20 N2 O3; 0.201 263.14038: - FBF Spectrum (rt: 0.168-0.192, 0.205 min) 82240_LB-008-BD06-127NEG.d Subtract
263.1404
([C14H20N2O3]-H)-
264.1433
([C14H20N2O3]-H)-
265.1458
([C14H20N2O3]-H)-
267.1484
([C14H20N2O3]-H)-
Counts vs. Mass-to-Charge (m/z)

VI·HCl **D21**

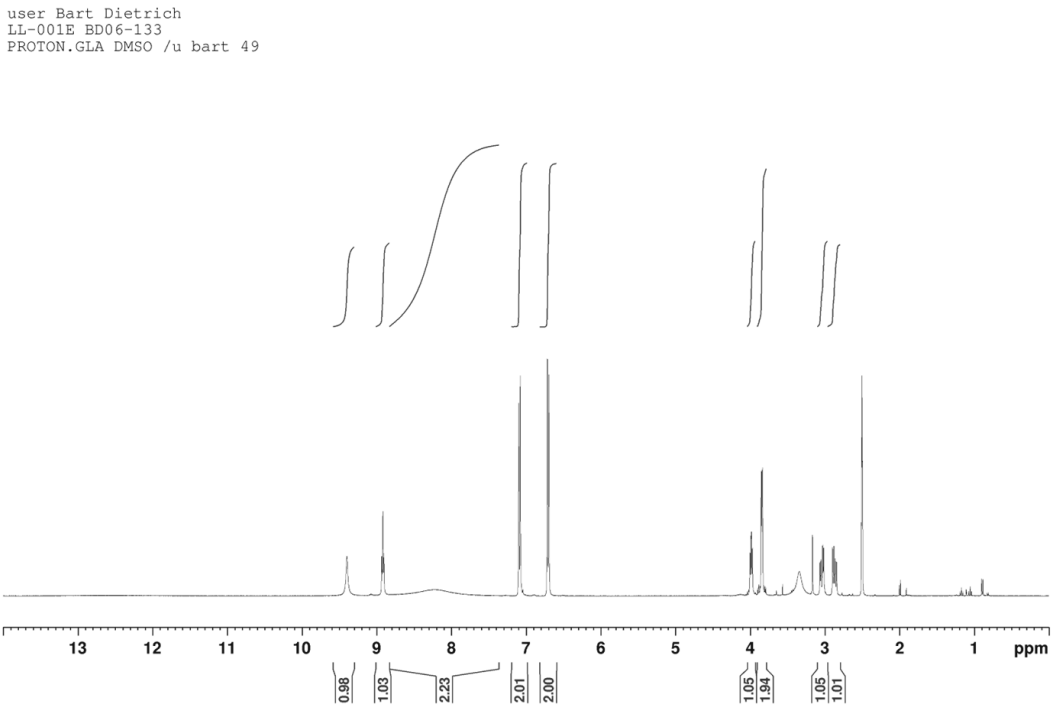

user Bart Dietrich
LH-010C BD06-129
PROTON.GLA DMSO /u bart 29
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
0.57
3.58
1.02
1.00
1.07
0.04
1.00
0.98
1.03
6.13
6.19

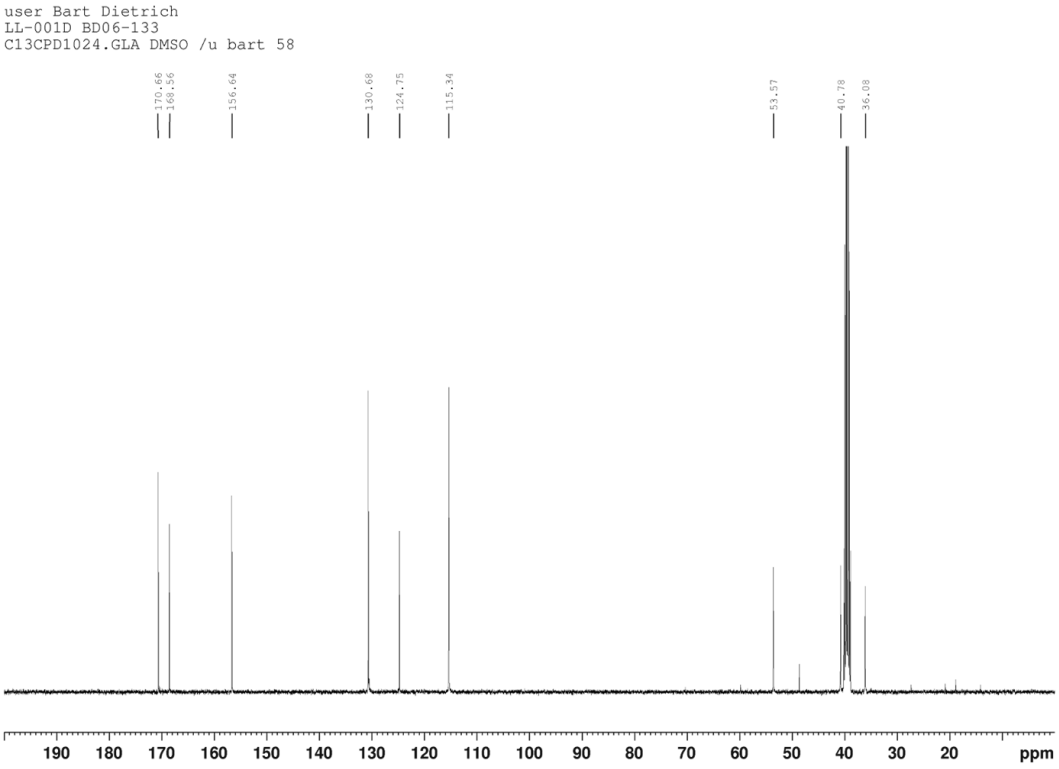

user Bart Dietrich
LH-010C BD06-129
C13CPD1024.GLA DMSO /u bart 14
172.30
168.08
56.78
56.72
36.17
29.94
24.72
18.22
17.86
15.60
11.44
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

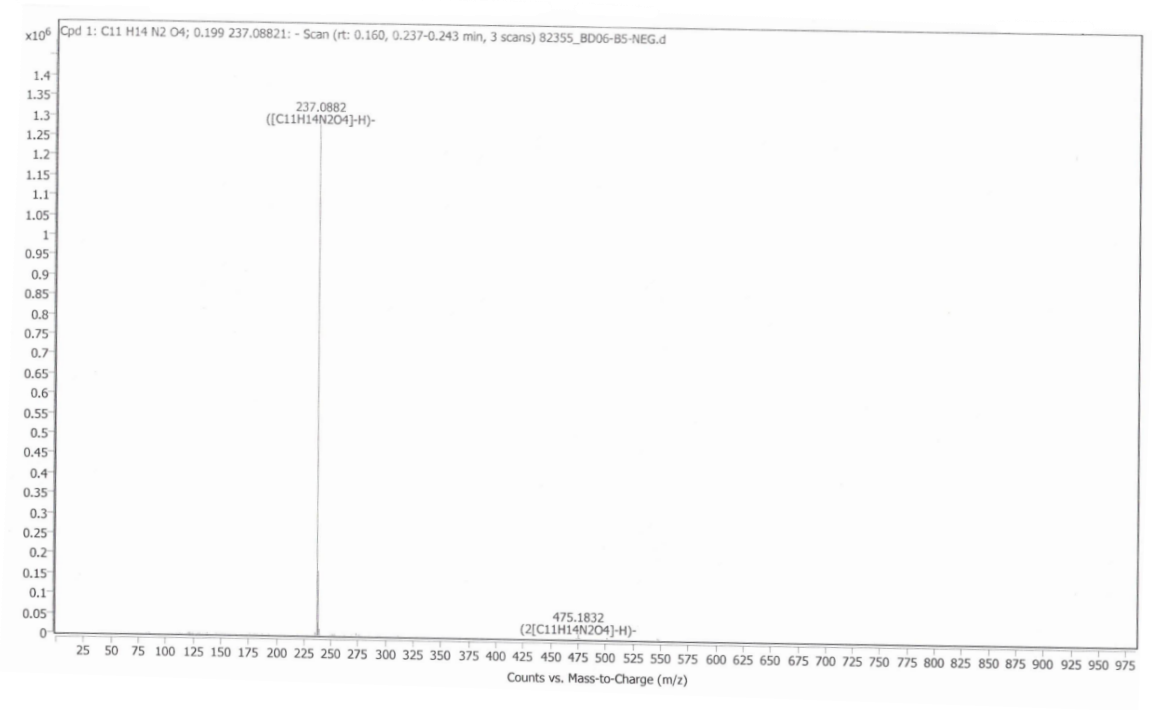

Cpd 1: C11 H22 N2 O3; 0.217 231.17122: + Scan (rt: 0.172-0.263 min, 13 scans) 82348_BD06-129.d
x10^6
231.1712
([C11H22N2O3]+H)+
461.3327
(2[C11H22N2O3]+H)+
729.4459
(3[C11H22N2O3]+K)+
Counts vs. Mass-to-Charge (m/z)

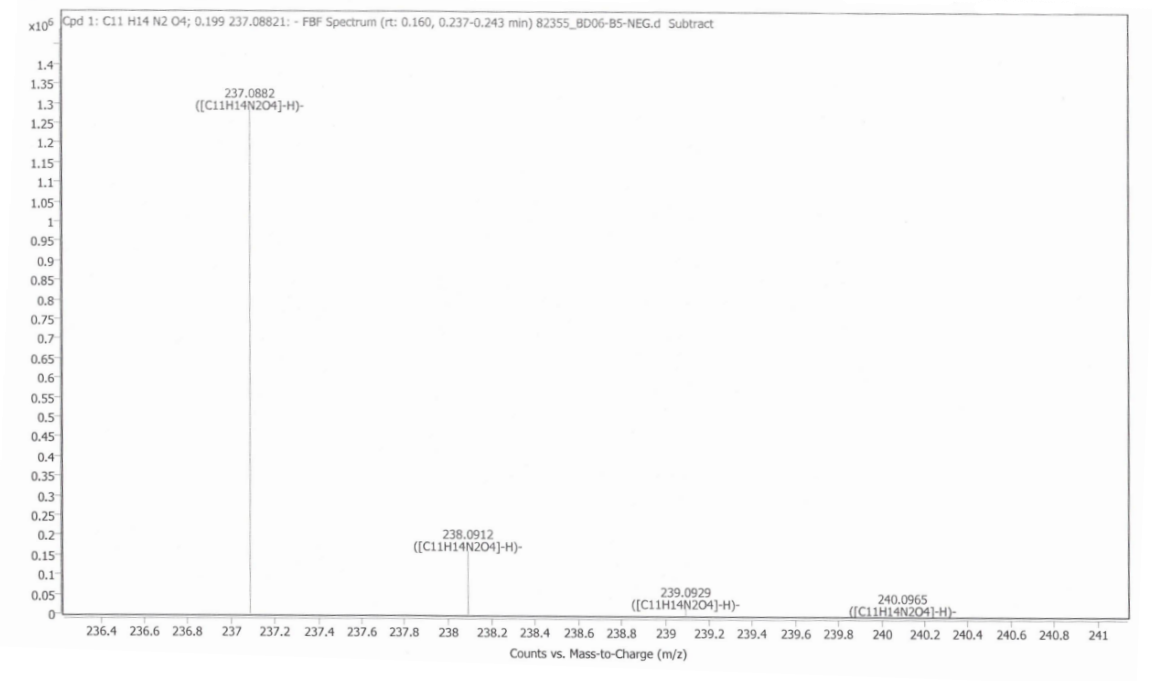

Cpd 1: C11 H22 N2 O3; 0.217 231.17122: + FBF Spectrum (rt: 0.172-0.263 min) 82348_BD06-129.d Subtract
x10^6
231.1712
([C11H22N2O3]+H)+
232.1742
([C11H22N2O3]+H)+
233.1763
([C11H22N2O3]+H)+
234.1797
([C11H22N2O3]+H)+
Counts vs. Mass-to-Charge (m/z)

YG·HCl **D22**

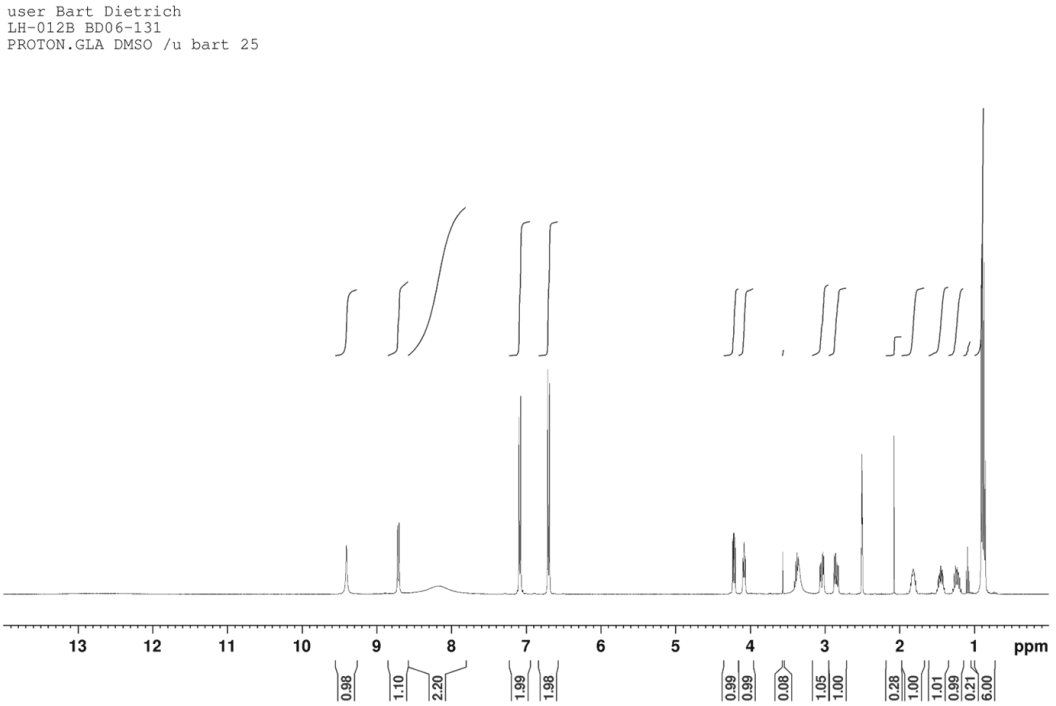

user Bart Dietrich
LL-001E BD06-133
PROTON.GLA DMSO /u bart 49
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
0.98 1.03 2.23 2.01 2.00 1.05 1.94 1.05 1.01

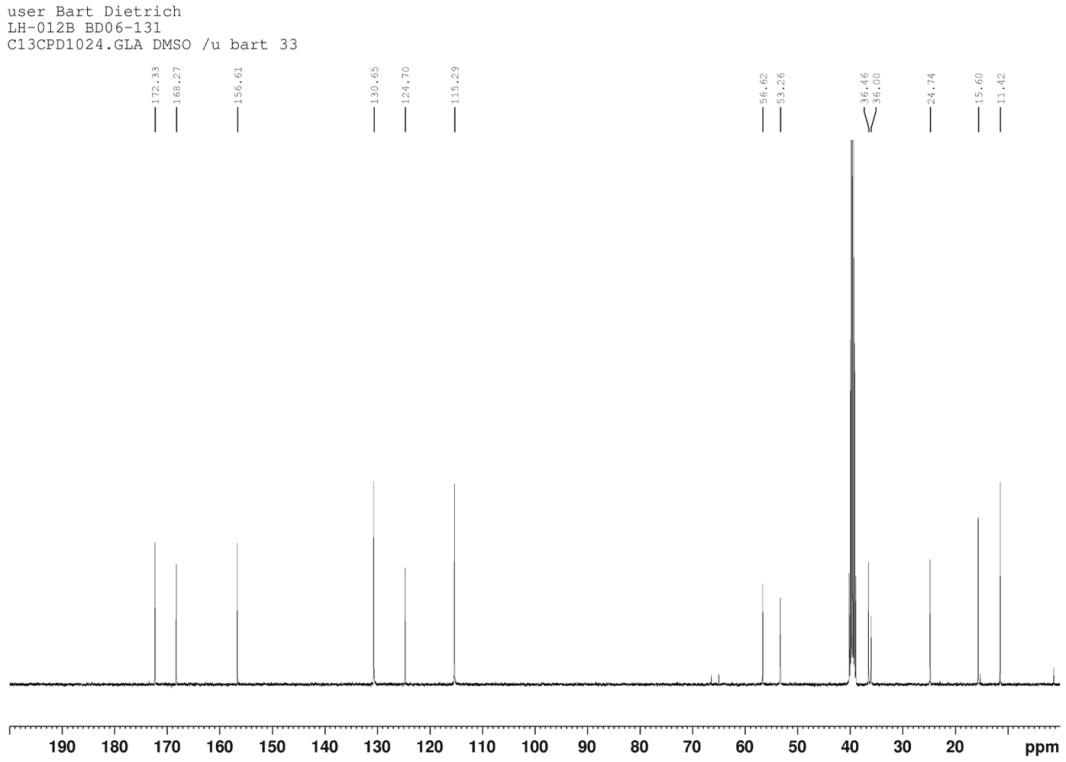

user Bart Dietrich
LL-001D BD06-133
C13CPD1024.GLA DMSO /u bart 58
170.66
168.56
156.64
130.68
124.75
115.34
53.57
40.78
36.08
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

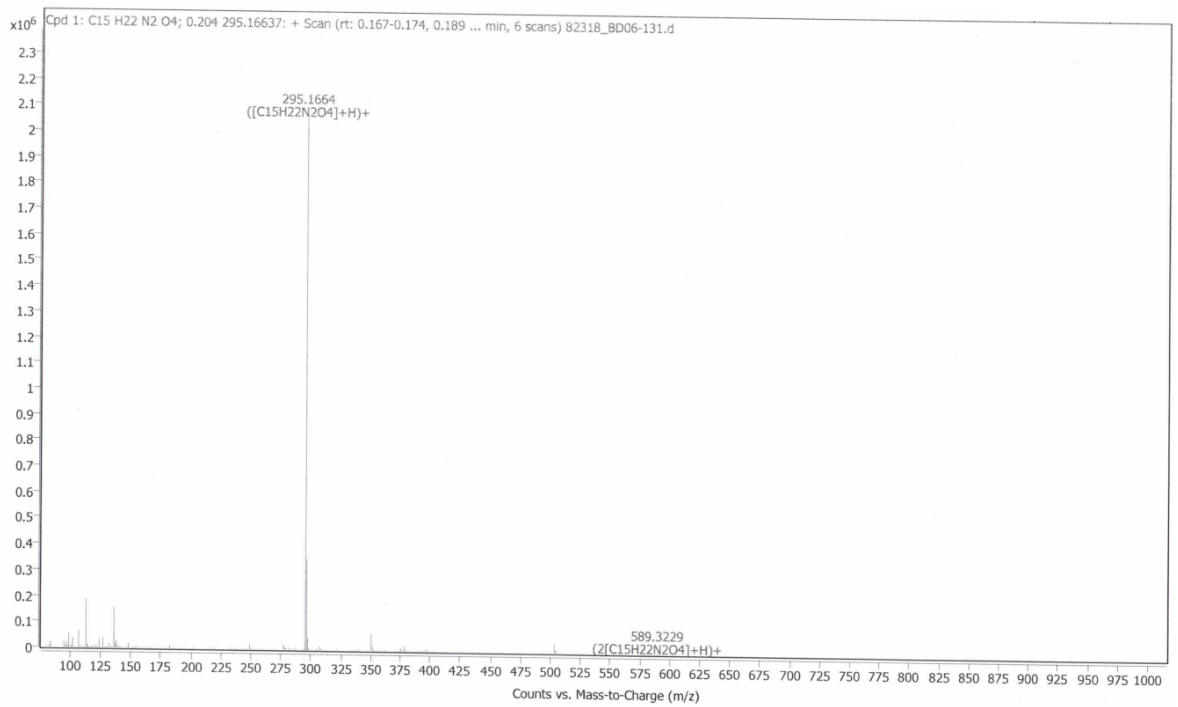

Cpd 1: C11 H14 N2 O4; 0.199 237.08821: - Scan (rt: 0.160, 0.237-0.243 min, 3 scans) 82355_BD06-B5-NEG.d
x10^6
237.0882
([C11H14N2O4]-H)-
475.1832
(2[C11H14N2O4]-H)-
Counts vs. Mass-to-Charge (m/z)

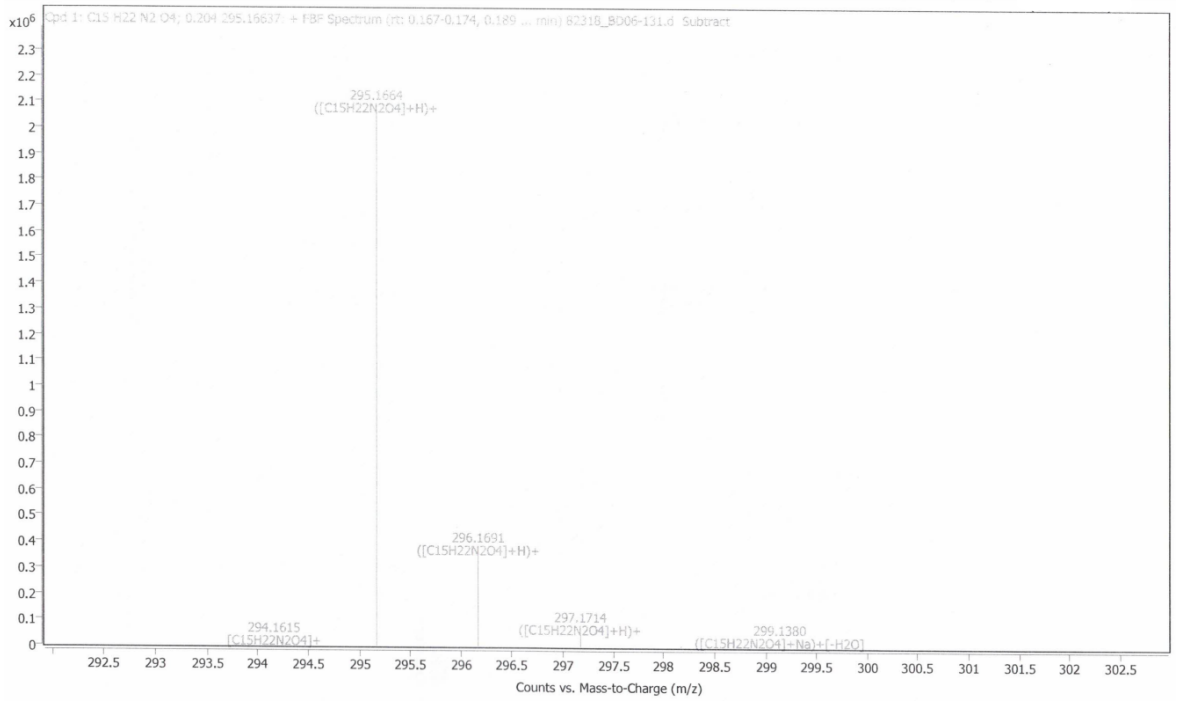

Cpd 1: C11 H14 N2 O4; 0.199 237.08821: - FBF Spectrum (rt: 0.160, 0.237-0.243 min) 82355_BD06-B5-NEG.d Subtract
x10^6
237.0882
([C11H14N2O4]-H)-
238.0912
([C11H14N2O4]-H)-
239.0929
([C11H14N2O4]-H)-
240.0965
([C11H14N2O4]-H)-
Counts vs. Mass-to-Charge (m/z)

YI·HCl **D23**


user Bart Dietrich
LH-012B BD06-131
PROTON.GLA DMSO /u bart 25
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
0.98 1.10 2.20 1.99 1.98 0.99 0.99 0.08 1.05 1.00 0.28 1.00 1.01 0.99 0.21 6.00

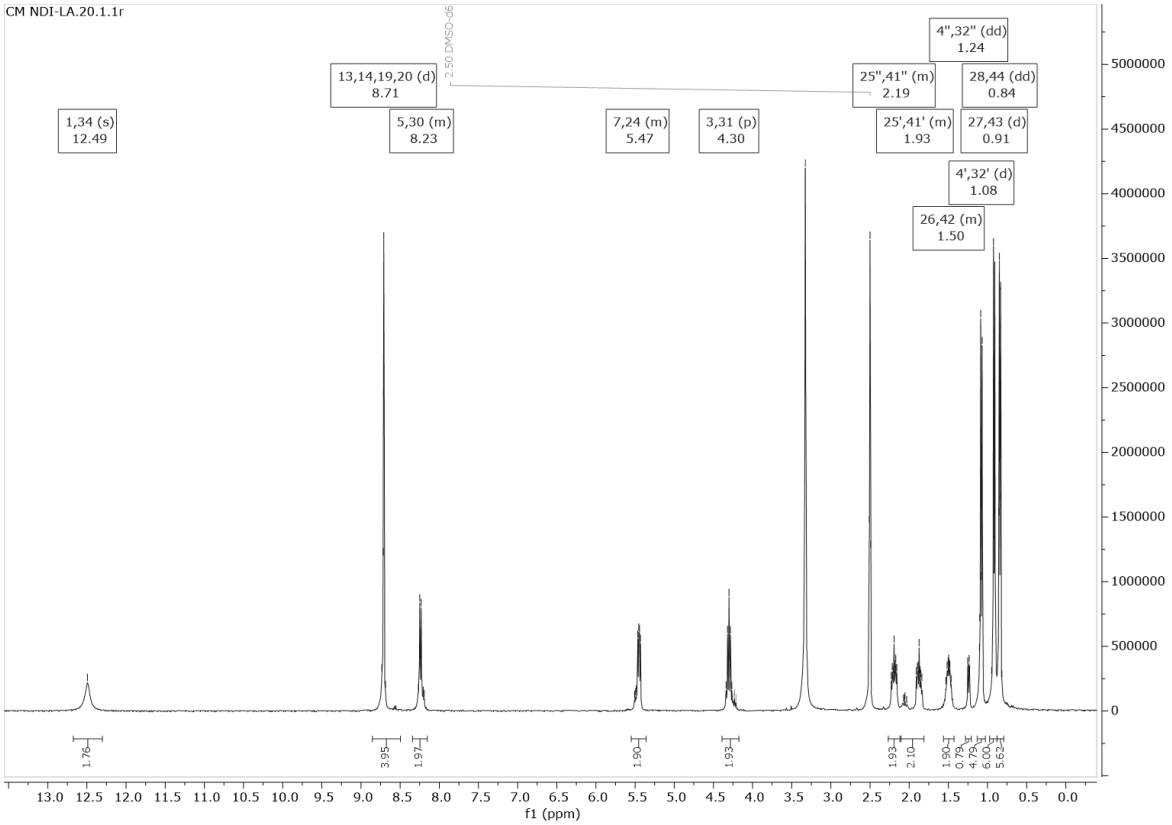

user Bart Dietrich
LH-012B BD06-131
C13CPD1024.GLA DMSO /u bart 33
172.33
168.27
156.61
130.65
124.70
115.29
56.62
53.26
36.46
36.00
24.74
15.60
11.42
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

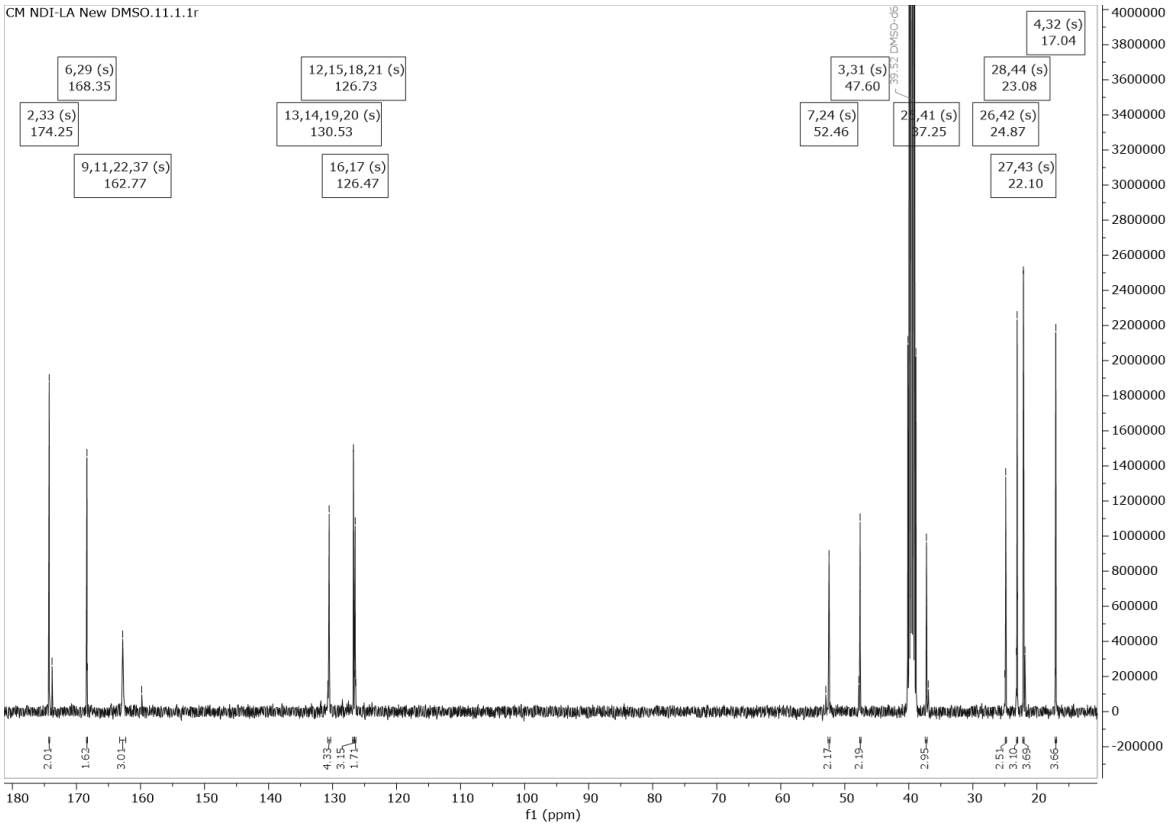

Cpd 1: C15 H22 N2 O4; 0.204 295.16637: + Scan (rt: 0.167-0.174, 0.189 ... min, 6 scans) 82318_BD06-131.d
x10^6
295.1664
([C15H22N2O4]+H)+
589.3229
(2[C15H22N2O4]+H)+
Counts vs. Mass-to-Charge (m/z)

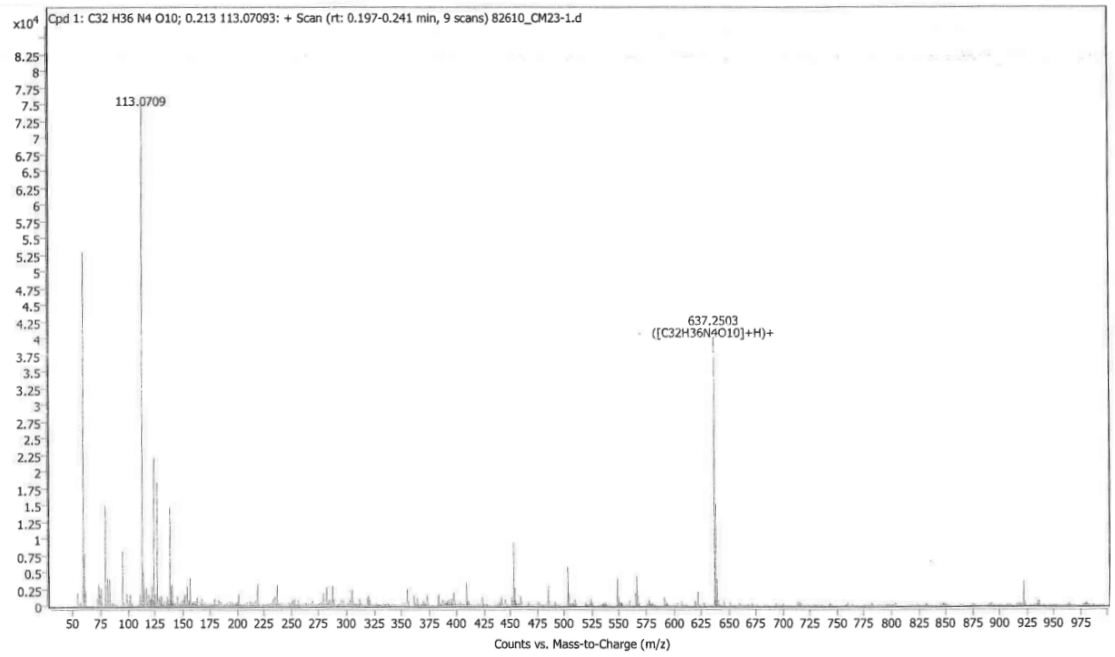

Cpd 1: C15 H22 N2 O4; 0.204 295.16637: + FBF Spectrum (rt: 0.167-0.174, 0.189 ... min) 82318_BD06-131.d Subtract
x10^6
295.1664
([C15H22N2O4]+H)+
296.1691
([C15H22N2O4]+H)+
294.1615
[C15H22N2O4]+
297.1714
([C15H22N2O4]+H)+
299.1380
([C15H22N2O4]+Na)+[-H2O]
Counts vs. Mass-to-Charge (m/z)

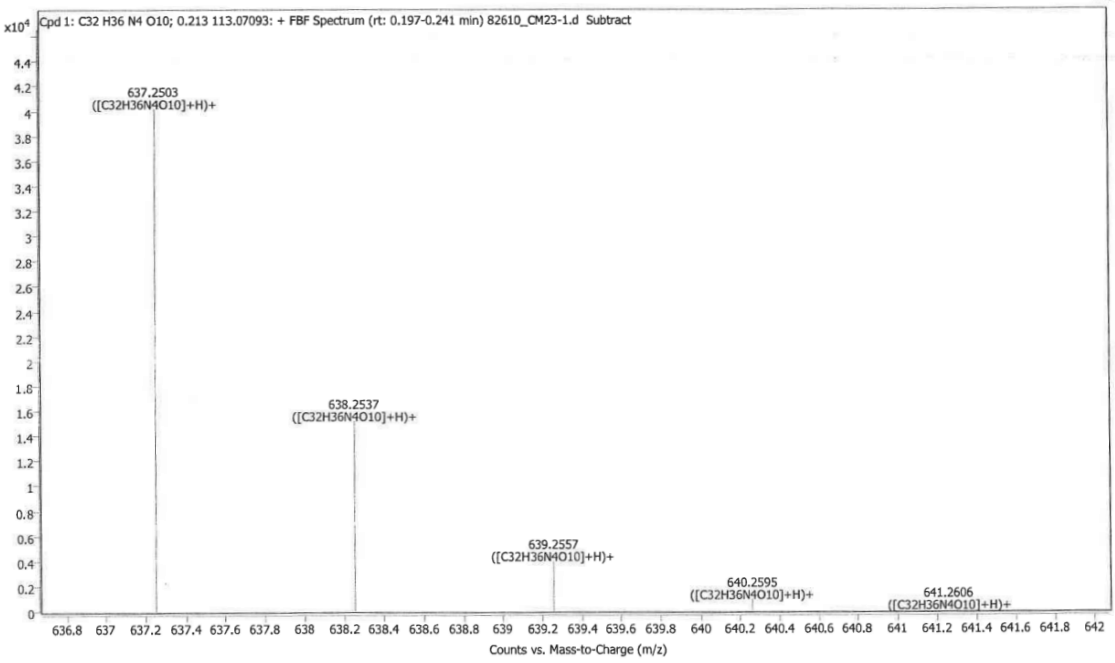
A 3.2.3 Rylene Products

NDI-LA **48**

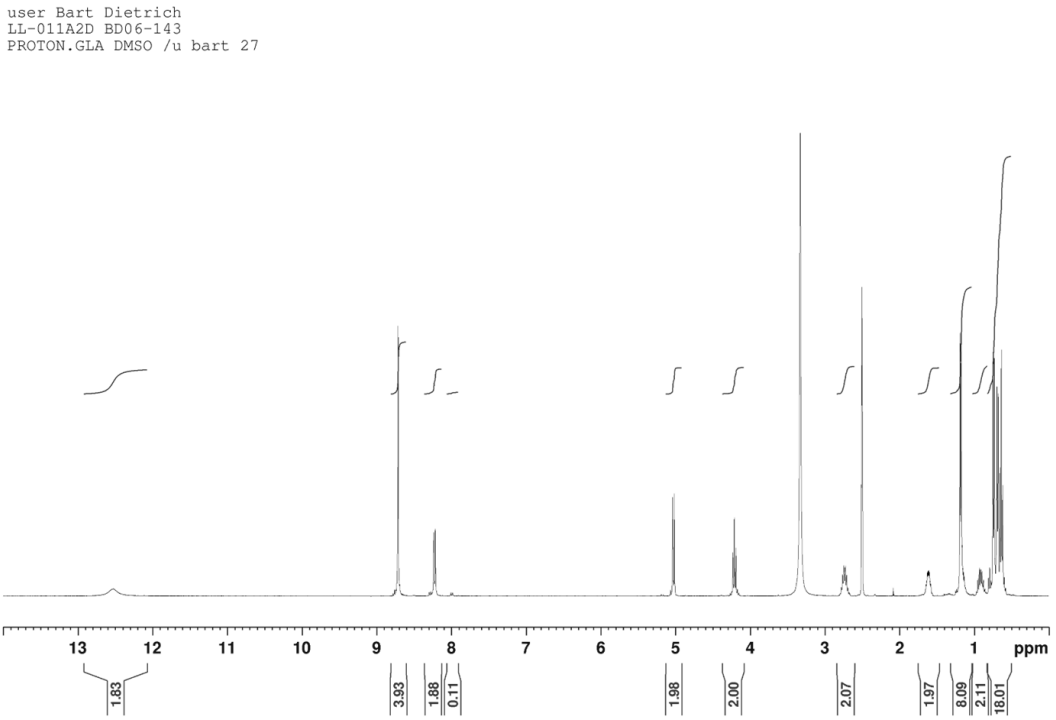


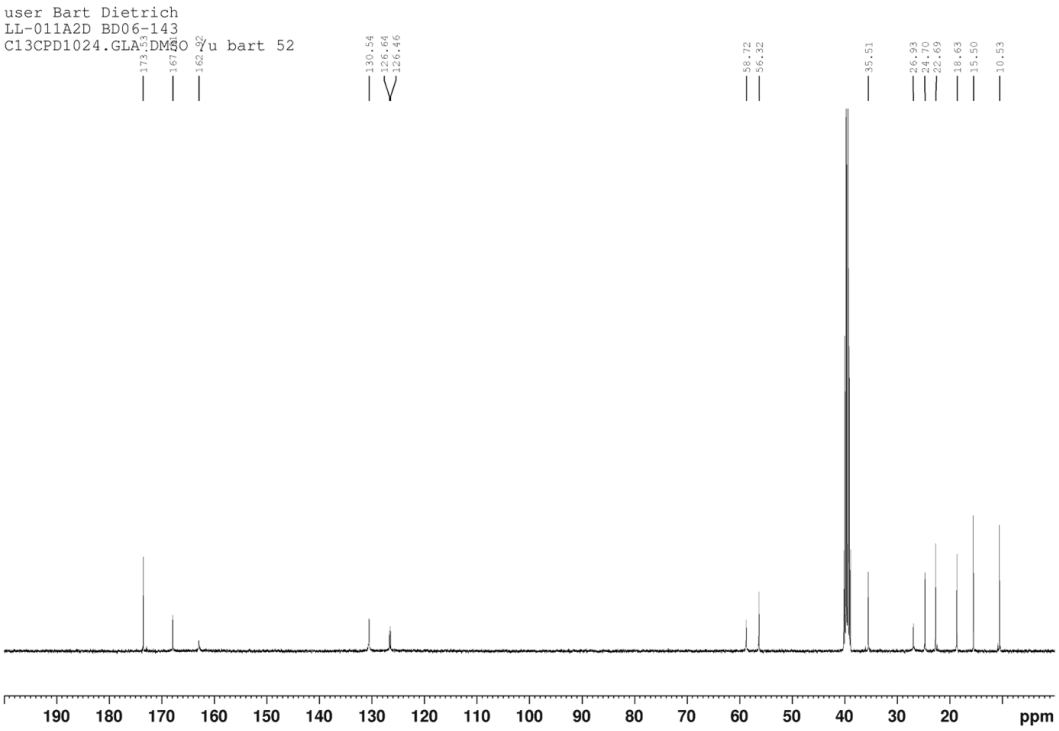


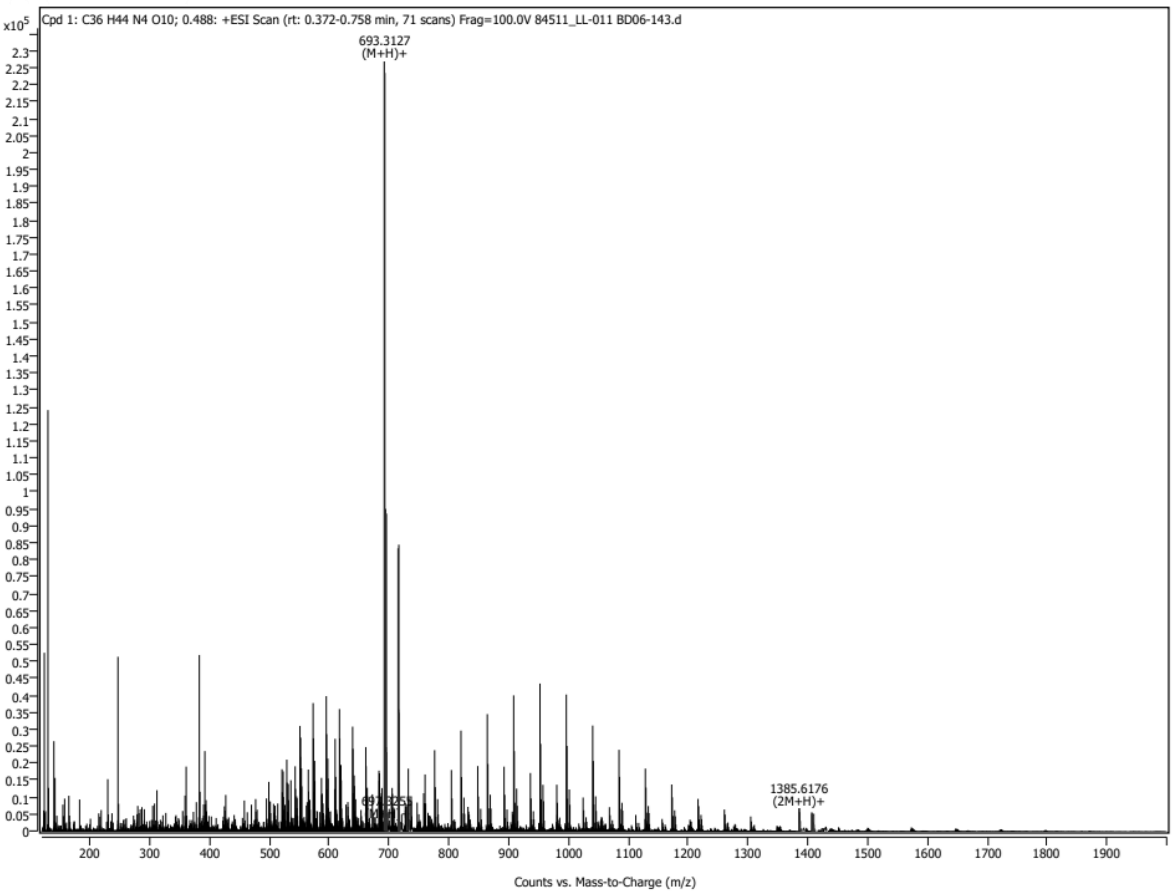

Cpd 1: C32 H36 N4 O10; 0.213 113.07093: + Scan (rt: 0.197-0.241 min, 9 scans) 82610_CM23-1.d
x10^4
113.0709
637.2503
([C32H36N4O10]+H)+
Counts vs. Mass-to-Charge (m/z)

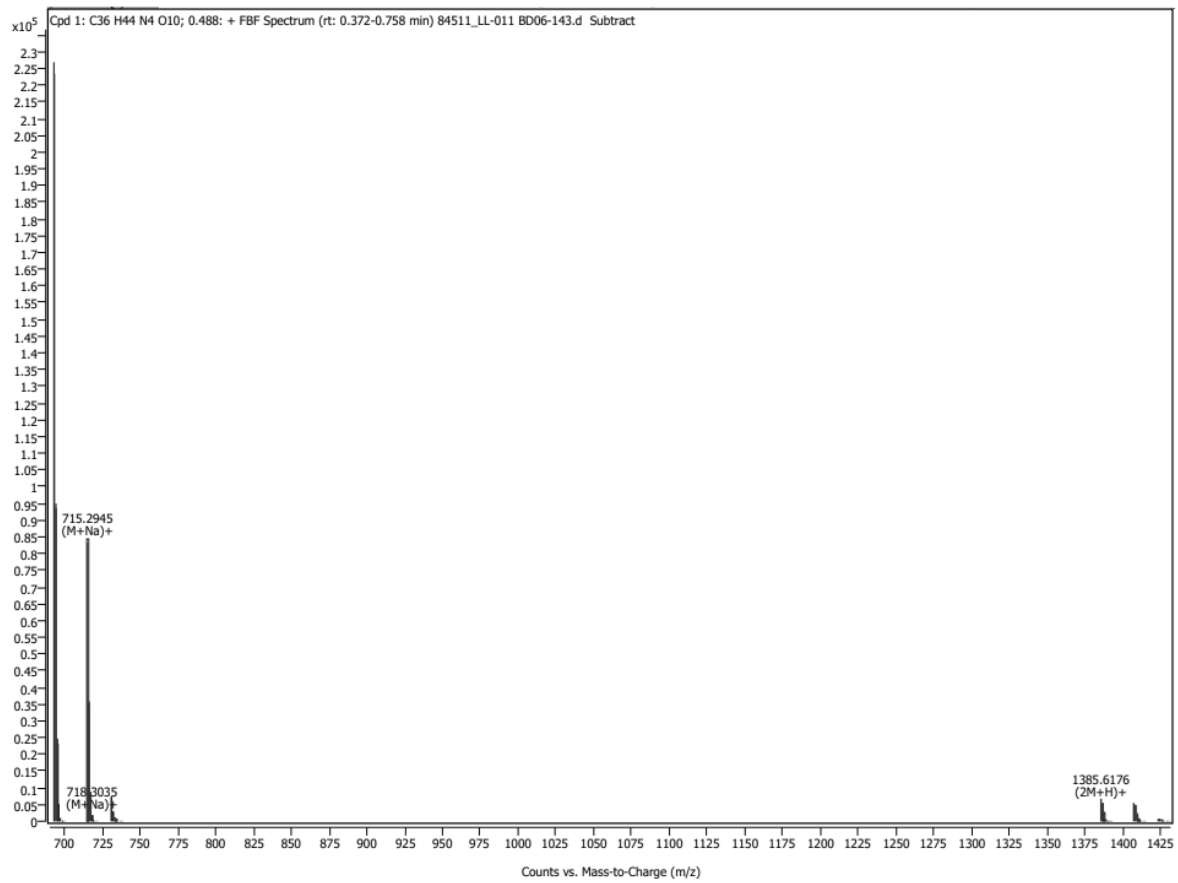

Cpd 1: C32 H36 N4 O10; 0.213 113.07093: + FBF Spectrum (rt: 0.197-0.241 min) 82610_CM23-1.d Subtract
x10^4
637.2503
([C32H36N4O10]+H)+
638.2537
([C32H36N4O10]+H)+
639.2557
([C32H36N4O10]+H)+
640.2595
([C32H36N4O10]+H)+
641.2606
([C32H36N4O10]+H)+
Counts vs. Mass-to-Charge (m/z)

NDI-VI **56**

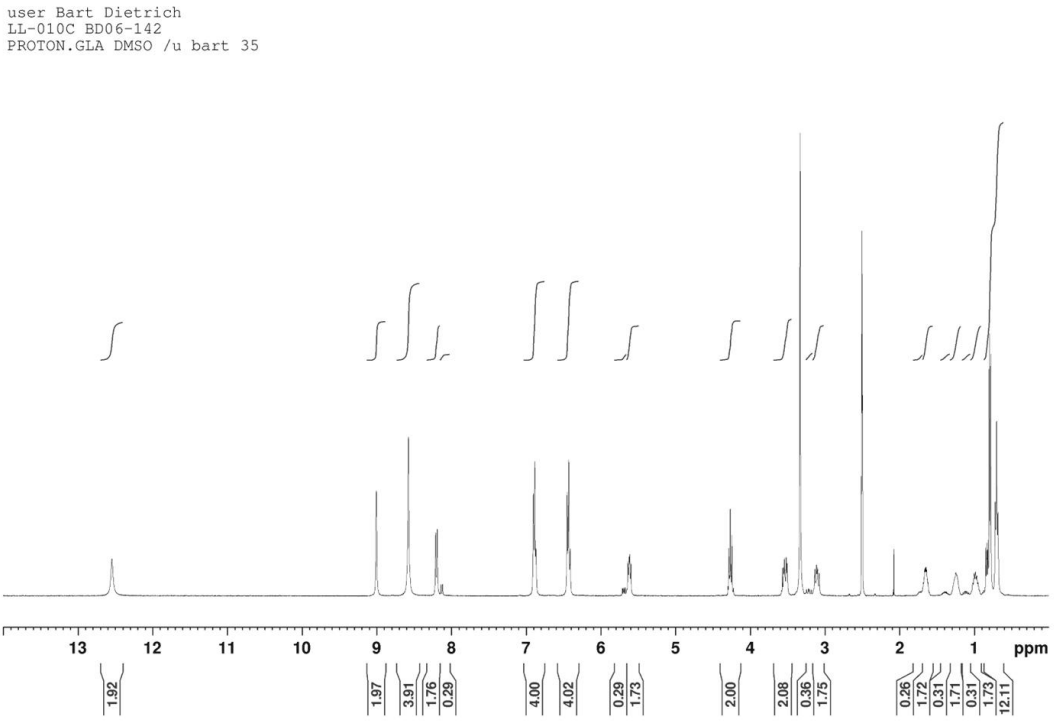

user Bart Dietrich
LL-011A2D BD06-143
PROTON.GLA DMSO /u bart 27
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
1.83
3.93
1.88
0.11
1.98
2.00
2.07
1.97
8.09
2.11
18.01

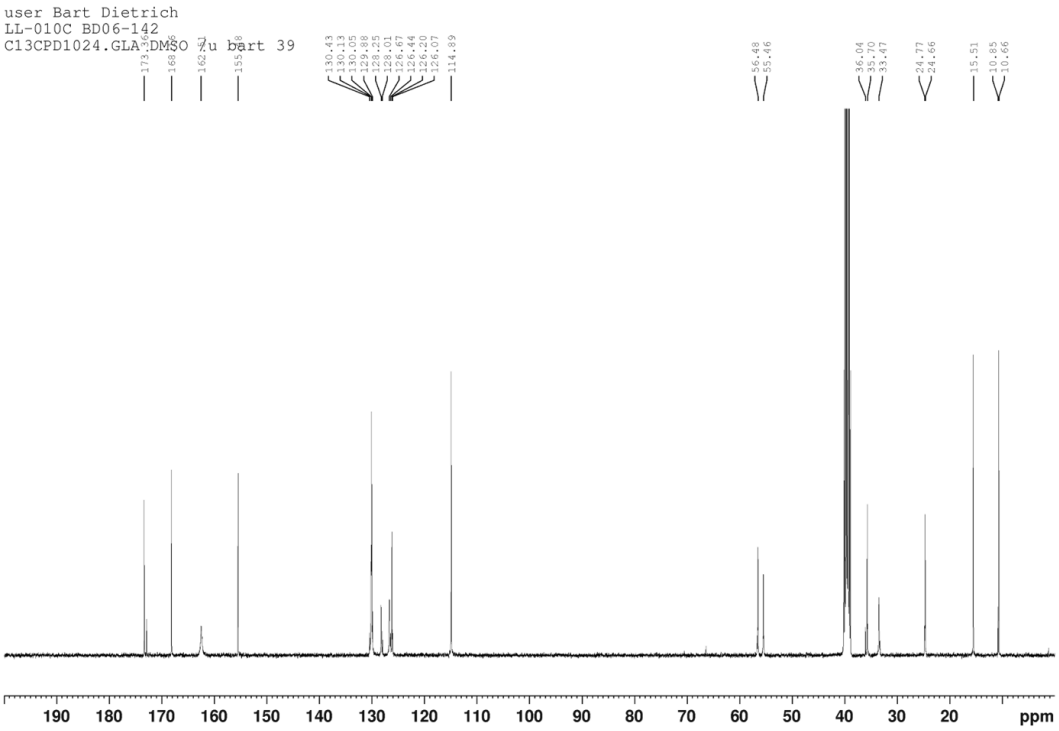

user Bart Dietrich
LL-011A2D BD06-143
C13CPD1024.GLA DMSO /u bart 52
173.35
167.80
162.32
130.54
126.64
126.46
58.72
56.32
35.51
26.93
24.70
22.69
18.63
15.50
10.53
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

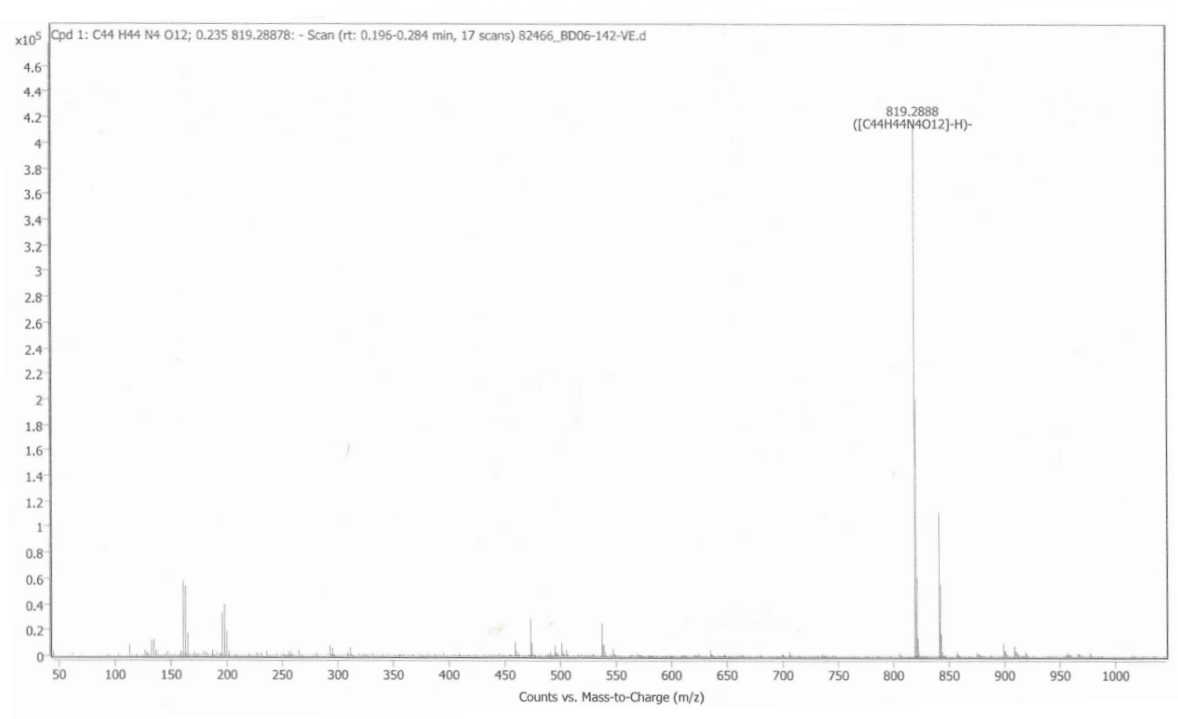

Cpd 1: C36 H44 N4 O10; 0.488: +ESI Scan (rt: 0.372-0.758 min, 71 scans) Frag=100.0V 84511_LL-011 BD06-143.d
693.3127
(M+H)+
1385.6176
(2M+H)+
Counts vs. Mass-to-Charge (m/z)

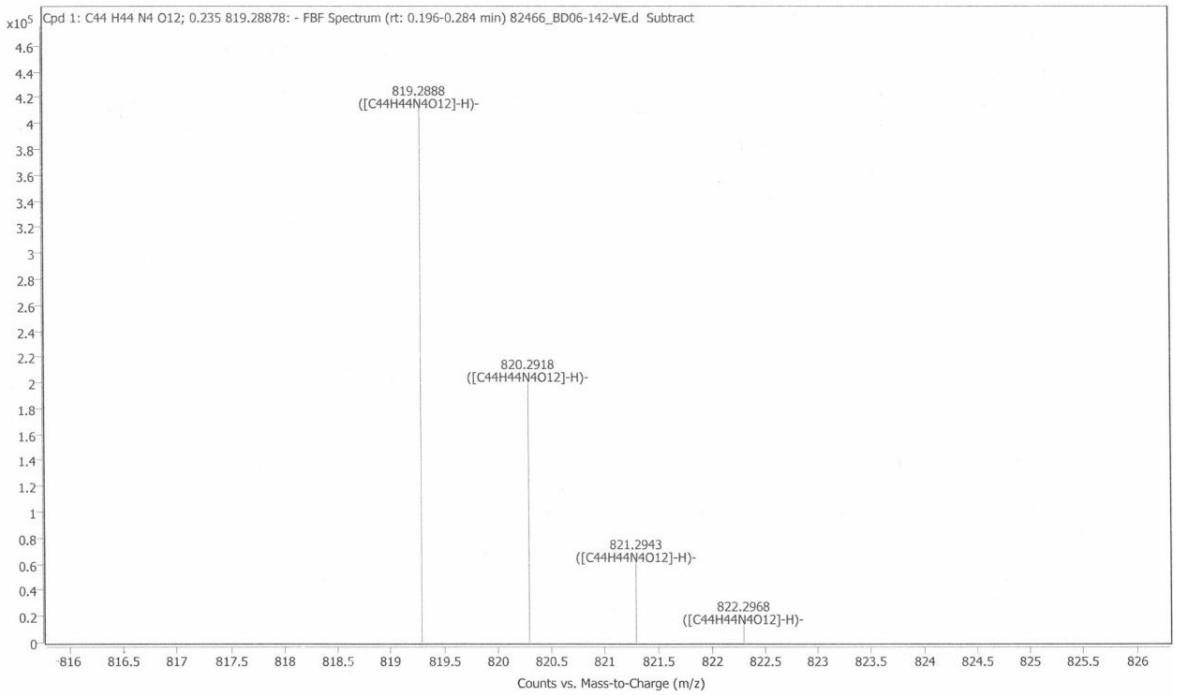

Cpd 1: C36 H44 N4 O10; 0.488: + FBF Spectrum (rt: 0.372-0.758 min) 84511_LL-011 BD06-143.d Subtract
715.2945
(M+Na)+
1385.6176
(2M+H)+
Counts vs. Mass-to-Charge (m/z)

NDI-YI **49**


user Bart Dietrich
LL-010C BD06-142
PROTON.GLA DMSO /u bart 35
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
1.92 1.97 3.91 1.76 0.29 4.00 4.02 0.29 1.73 2.00 2.08 0.36 1.75 0.26 1.72 0.31 1.71 0.31 1.73 12.11

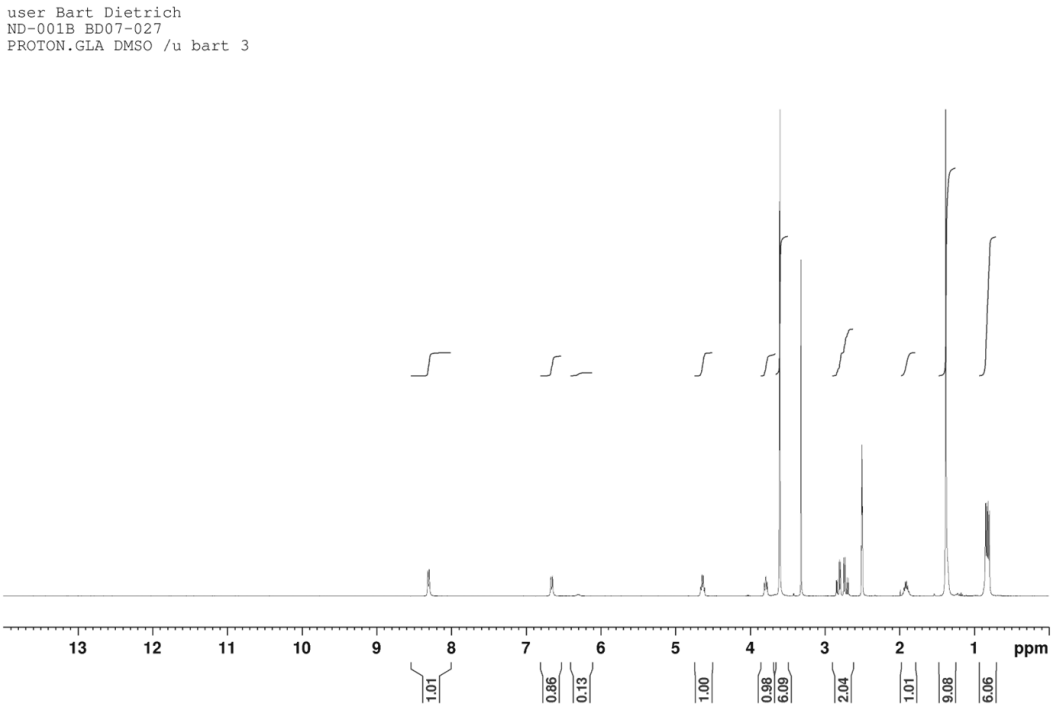

user Bart Dietrich
LL-010C BD06-142
C13CPD1024.GLA DMSO /u bart 39
130.43 130.13 130.05 129.88 128.25 128.01 126.47 126.46 126.20 126.07 114.89
56.48 55.46
36.04 35.70 33.47
24.77 24.66
15.51 10.85 10.66
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

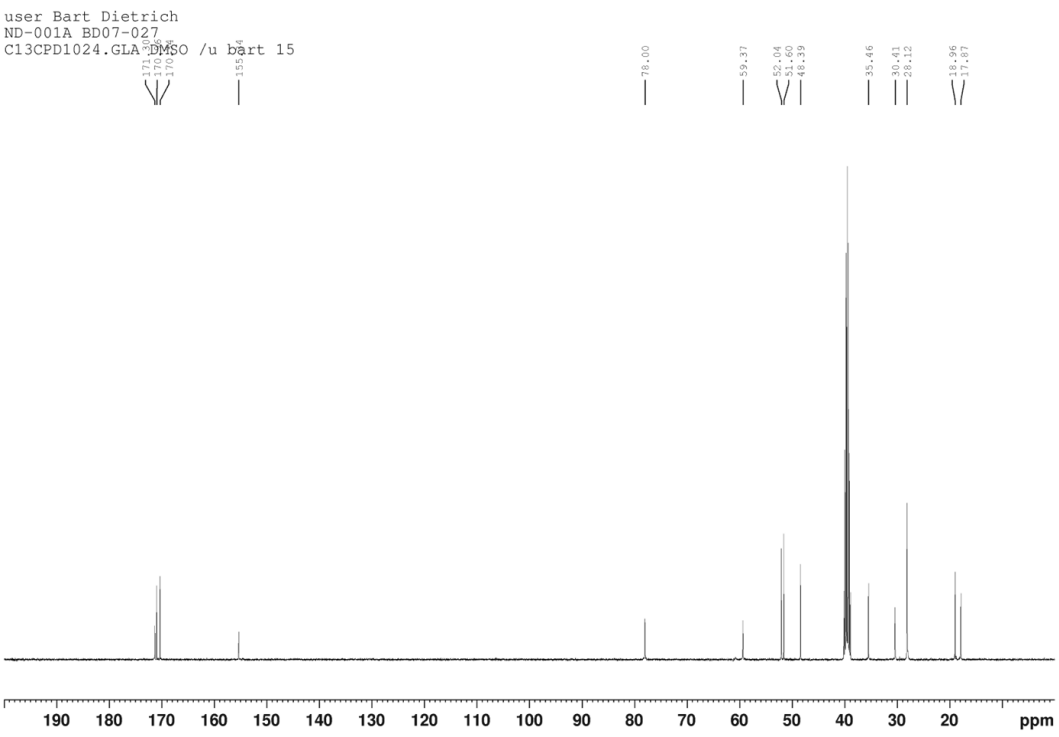

Cpd 1: C44 H44 N4 O12; 0.235 819.28878: - Scan (rt: 0.196-0.284 min, 17 scans) 82466_BD06-142-VE.d
x10^5
819.2888
([C44H44N4O12]-H)-
Counts vs. Mass-to-Charge (m/z)

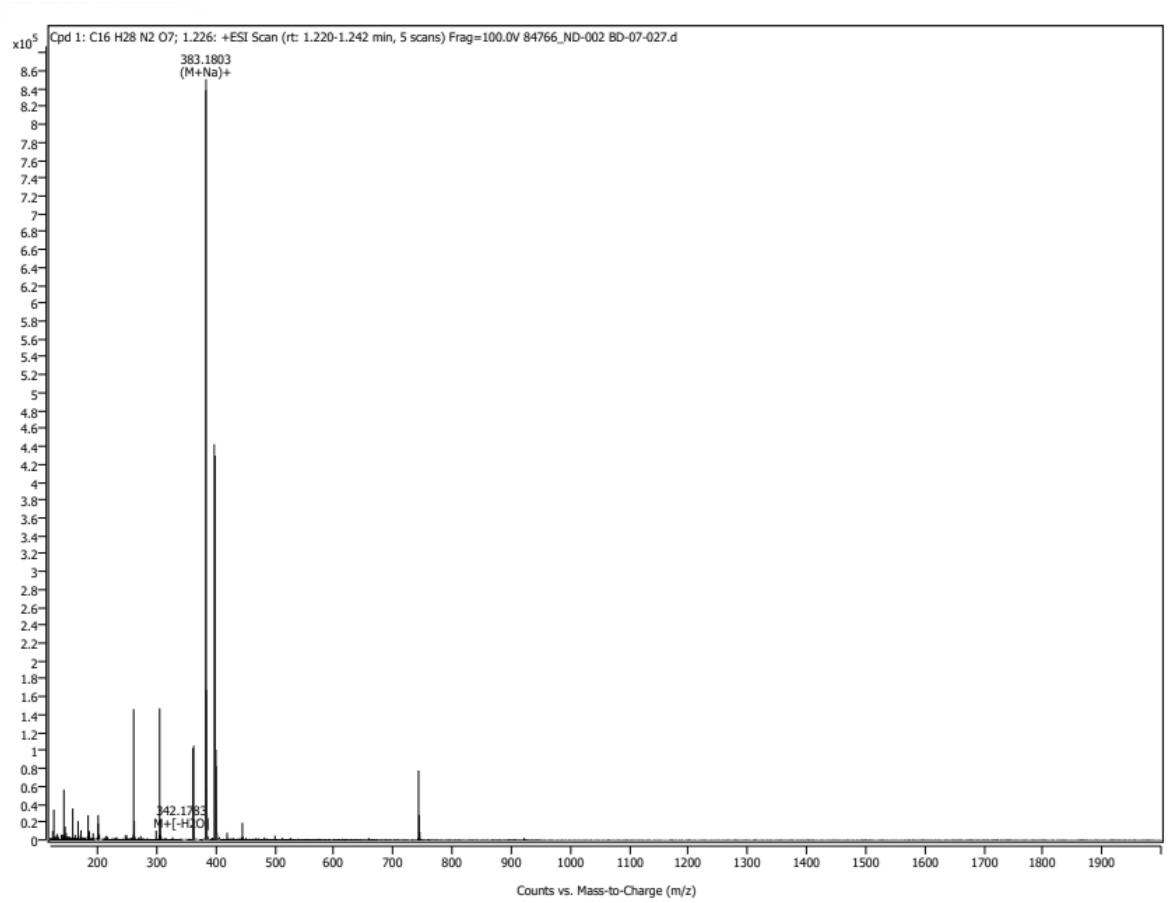

Cpd 1: C44 H44 N4 O12; 0.235 819.28878: - FBF Spectrum (rt: 0.196-0.284 min) 82466_BD06-142-VE.d Subtract
x10^5
819.2888
([C44H44N4O12]-H)-
820.2918
([C44H44N4O12]-H)-
821.2943
([C44H44N4O12]-H)-
822.2968
([C44H44N4O12]-H)-
Counts vs. Mass-to-Charge (m/z)

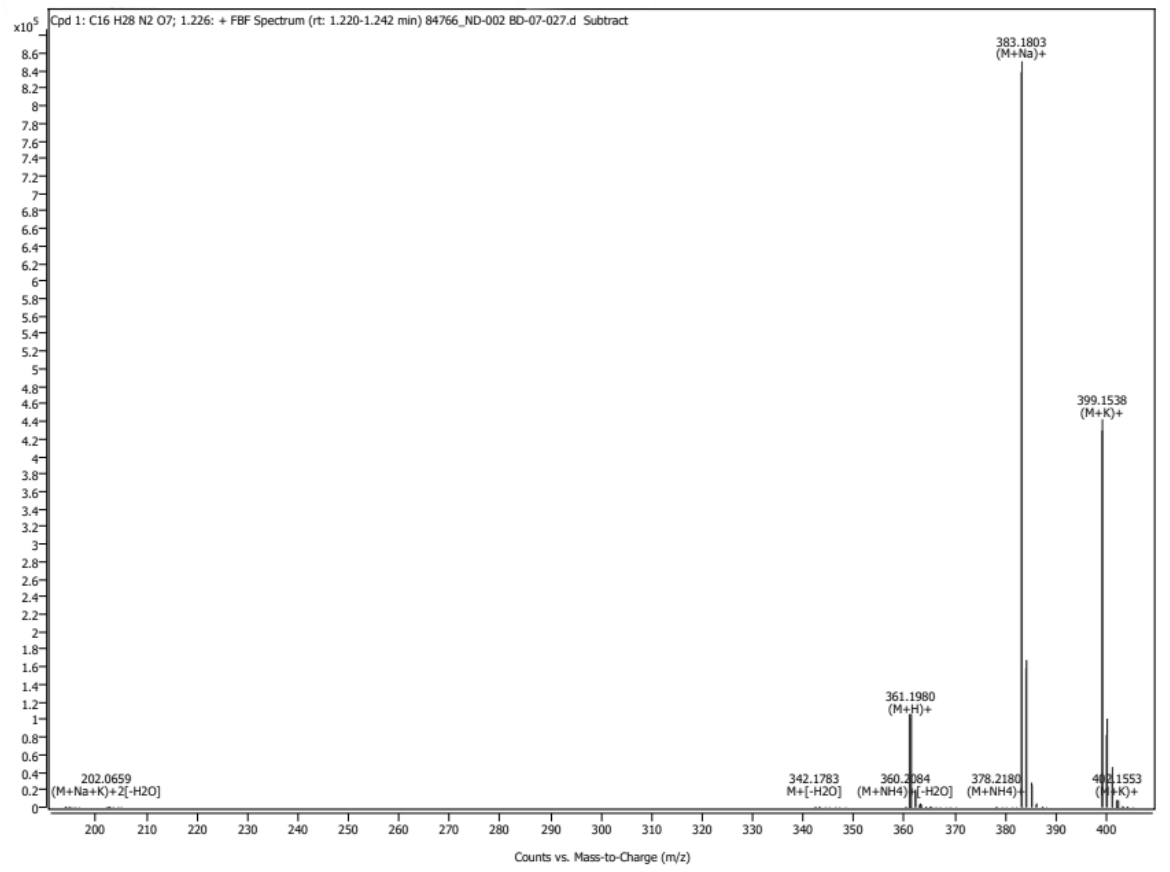
A 3.2.4 Validation Products and Precursors

Boc-VD-OMe **G1**

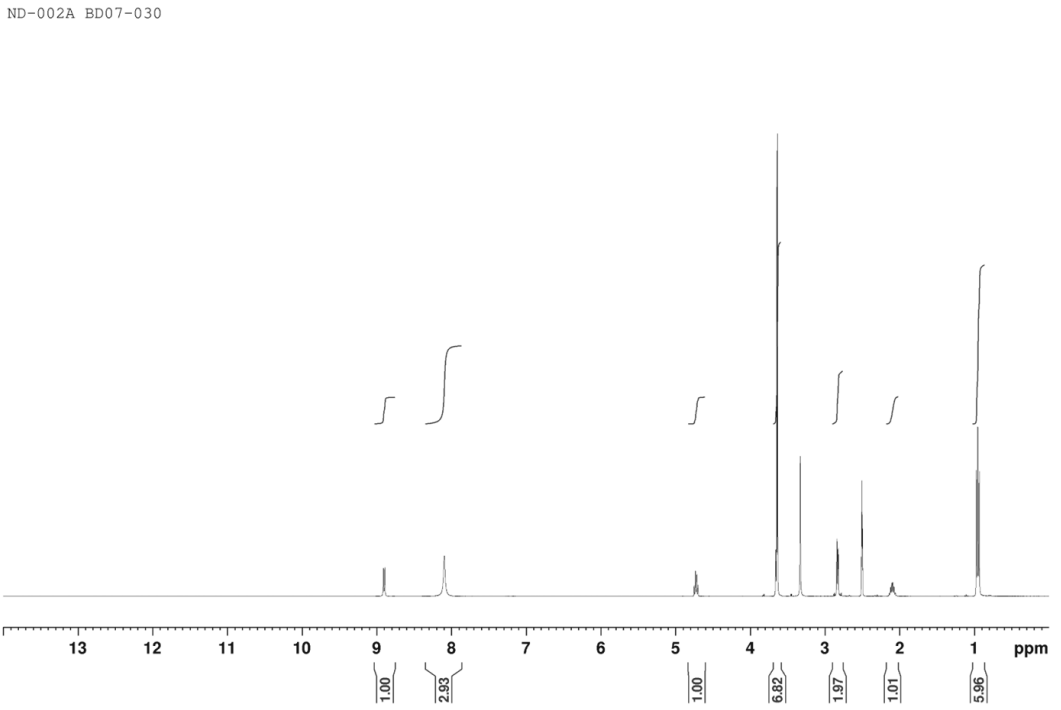


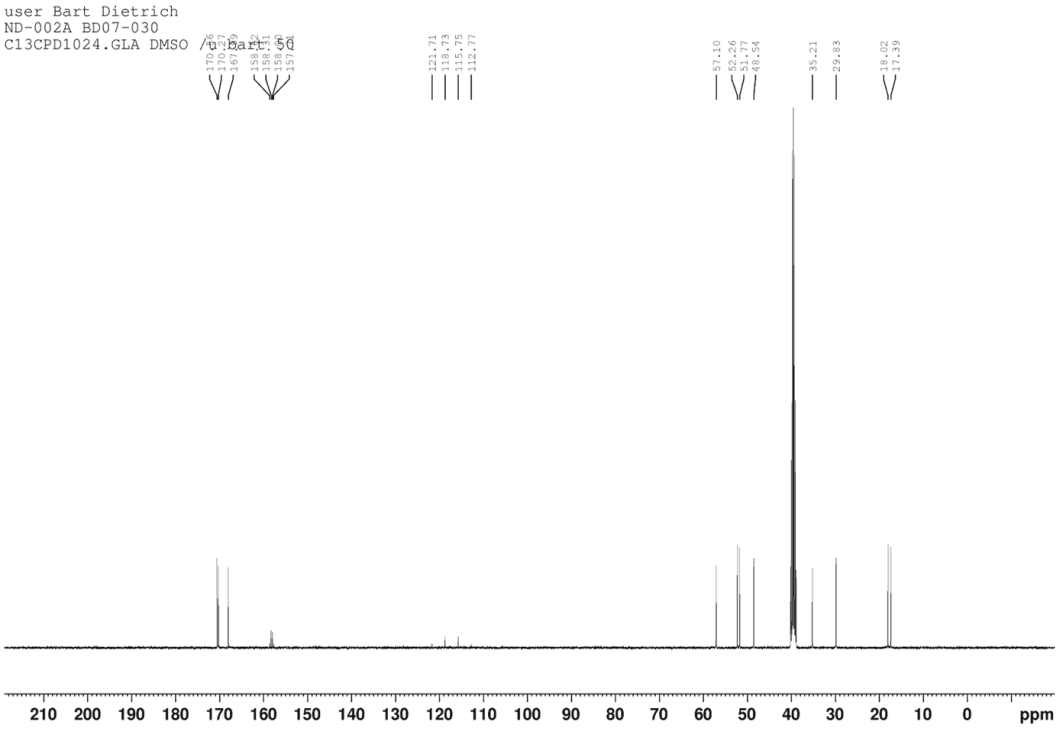


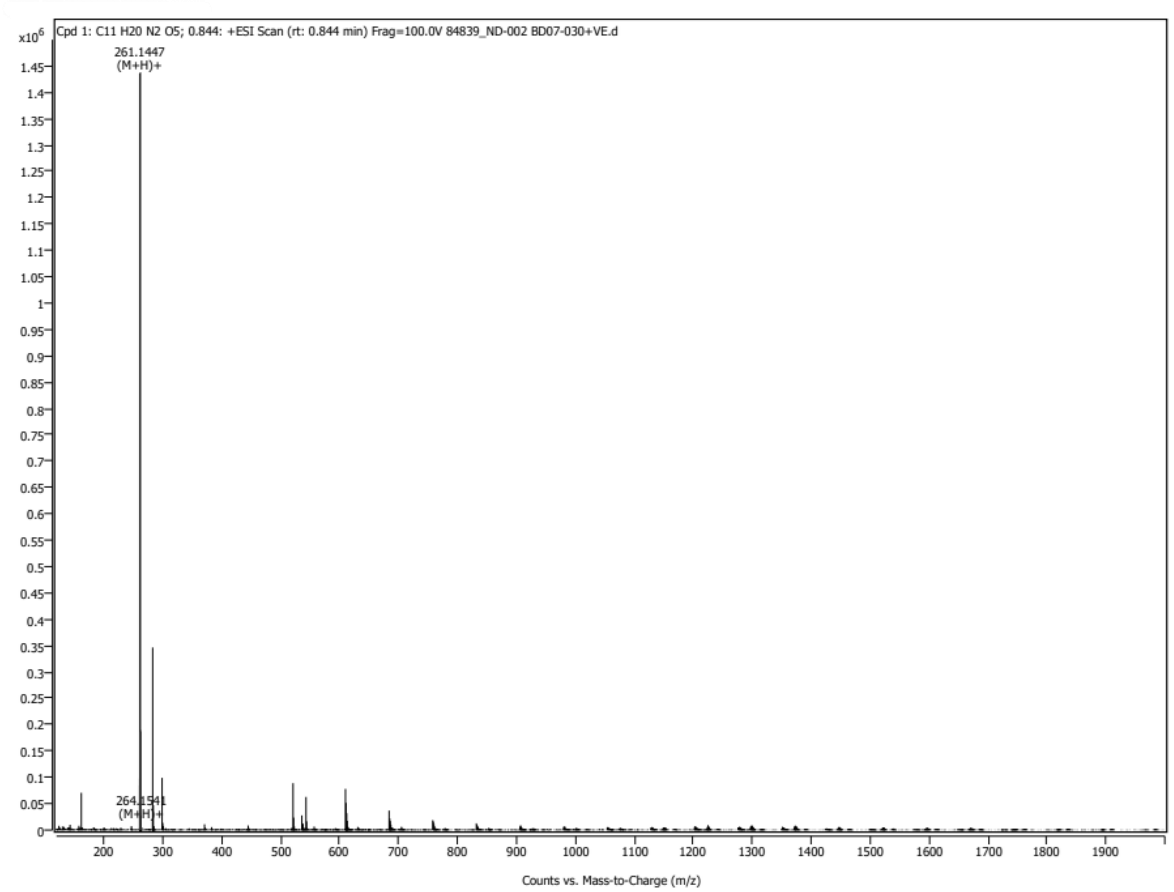

Cpd 1: C16 H28 N2 O7; 1.226: +ESI Scan (rt: 1.220-1.242 min, 5 scans) Frag=100.0V 84766_ND-002 BD-07-027.d
x10^5
383.1803
(M+Na)+
342.1783
M+[-H2O]
Counts vs. Mass-to-Charge (m/z)

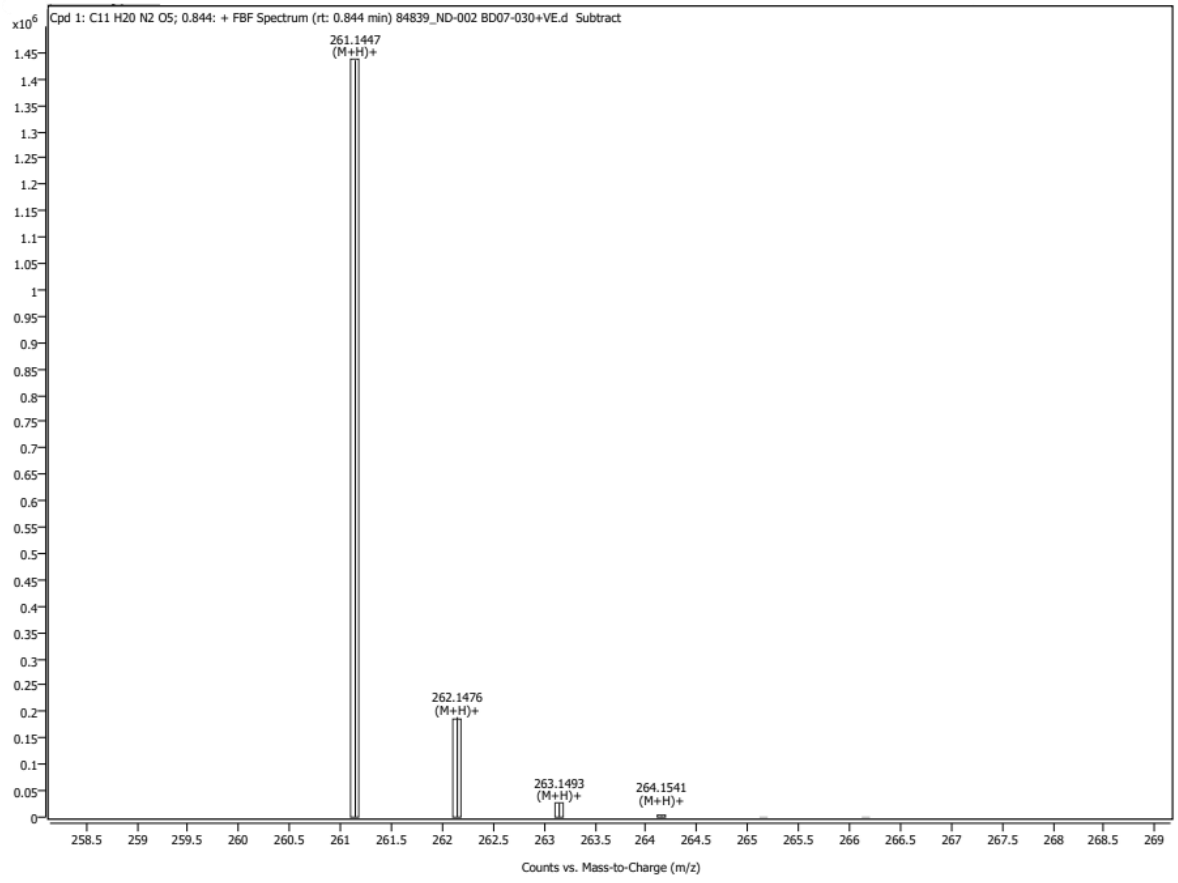

Cpd 1: C16 H28 N2 O7; 1.226: + FBF Spectrum (rt: 1.220-1.242 min) 84766_ND-002 BD-07-027.d Subtract
x10^5
383.1803
(M+Na)+
399.1538
(M+K)+
361.1980
(M+H)+
202.0659
(M+Na+K)+2[-H2O]
342.1783
M+[-H2O]
360.2084
(M+NH4)+[-H2O]
378.2180
(M+NH4)+
401.1553
(M+K)+
Counts vs. Mass-to-Charge (m/z)

VD-OMe·TFA **G2**

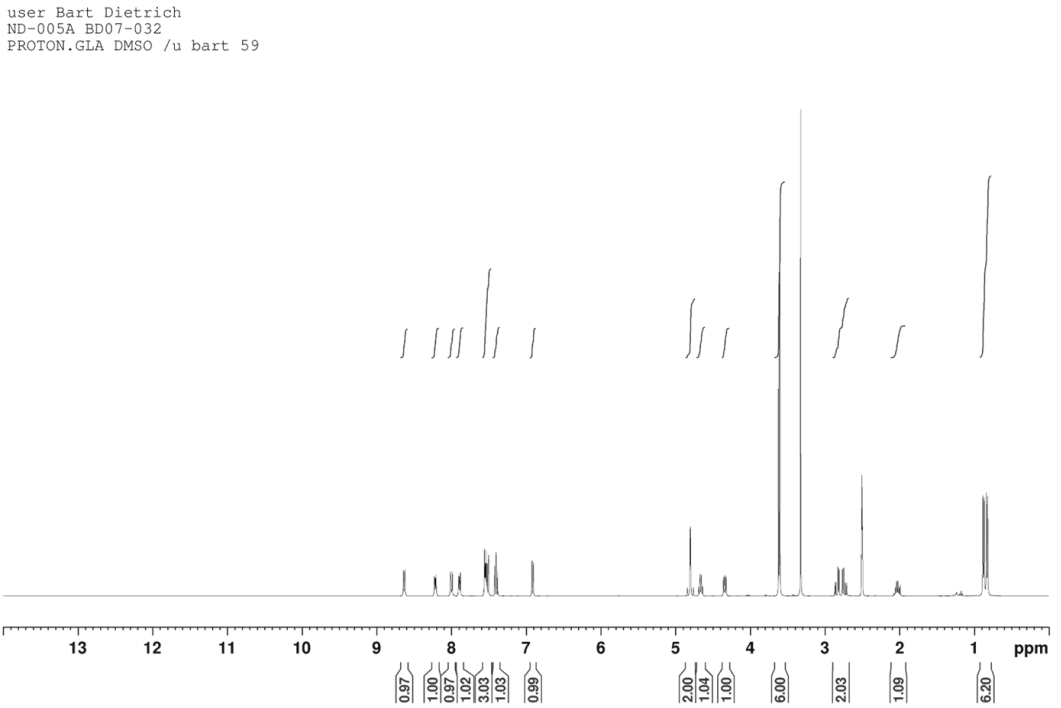

ND-002A BD07-030
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
1.00 2.93 1.00 6.82 1.97 1.01 5.96

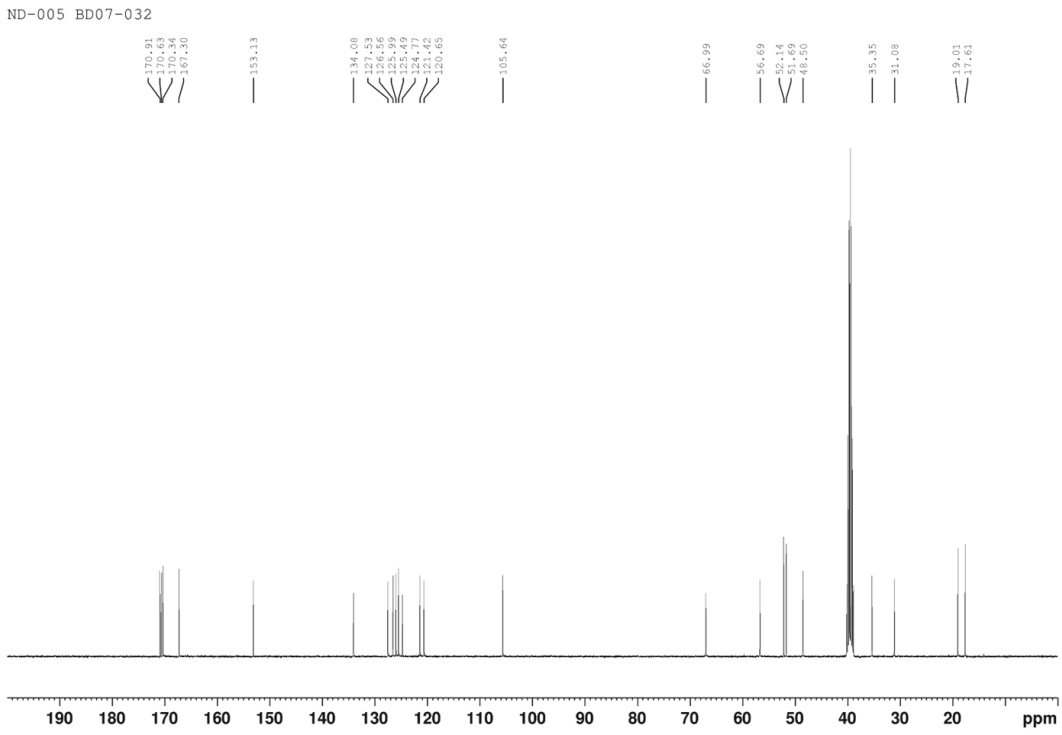

user Bart Dietrich
ND-002A BD07-030
C13CPD1024.GLA DMSO /d bart
121.71
118.73
115.75
112.77
57.10
52.26
51.77
48.54
35.21
29.83
18.02
17.39
210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 ppm

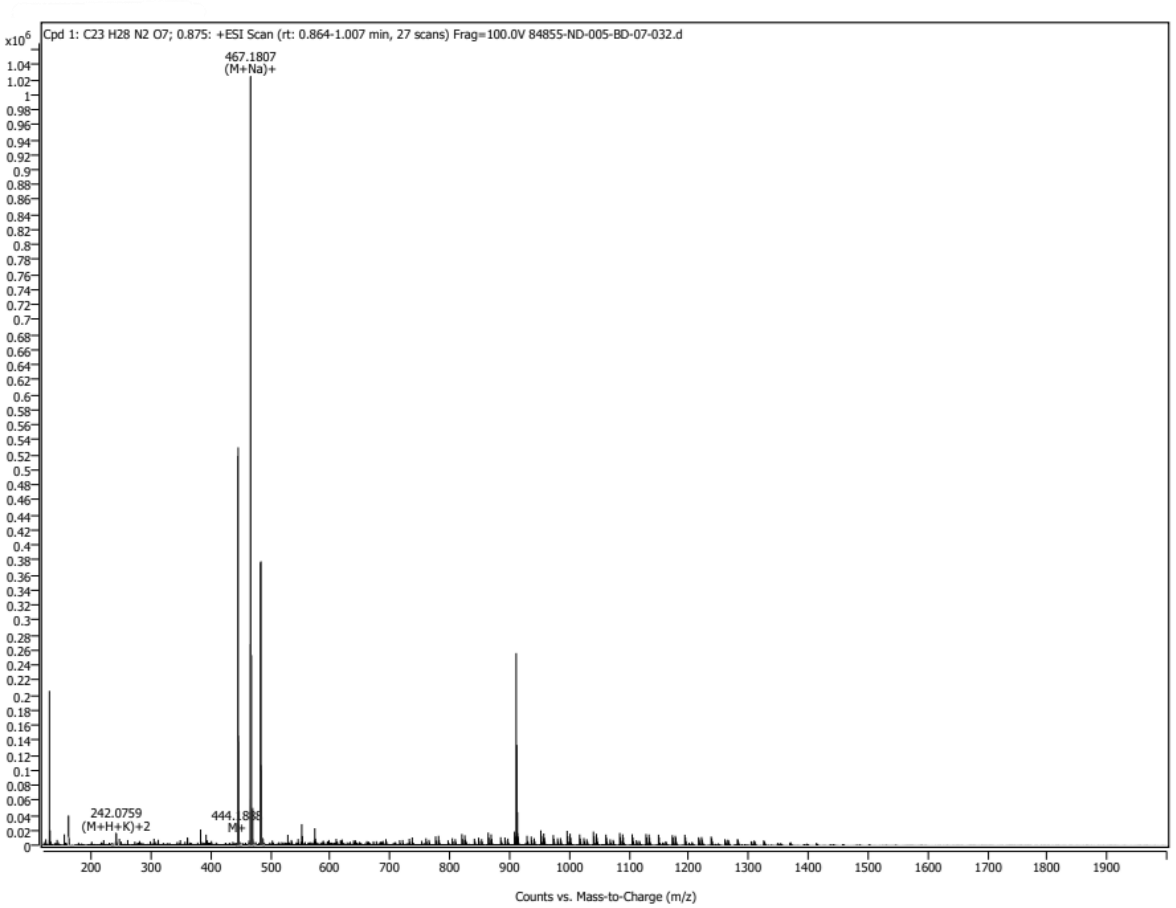

Cpd 1: C11 H20 N2 O5; 0.844: +ESI Scan (rt: 0.844 min) Frag=100.0V 84839_ND-002 BD07-030+VE.d
x10^6
261.1447
(M+H)+
264.1541
(M+H)+
Counts vs. Mass-to-Charge (m/z)

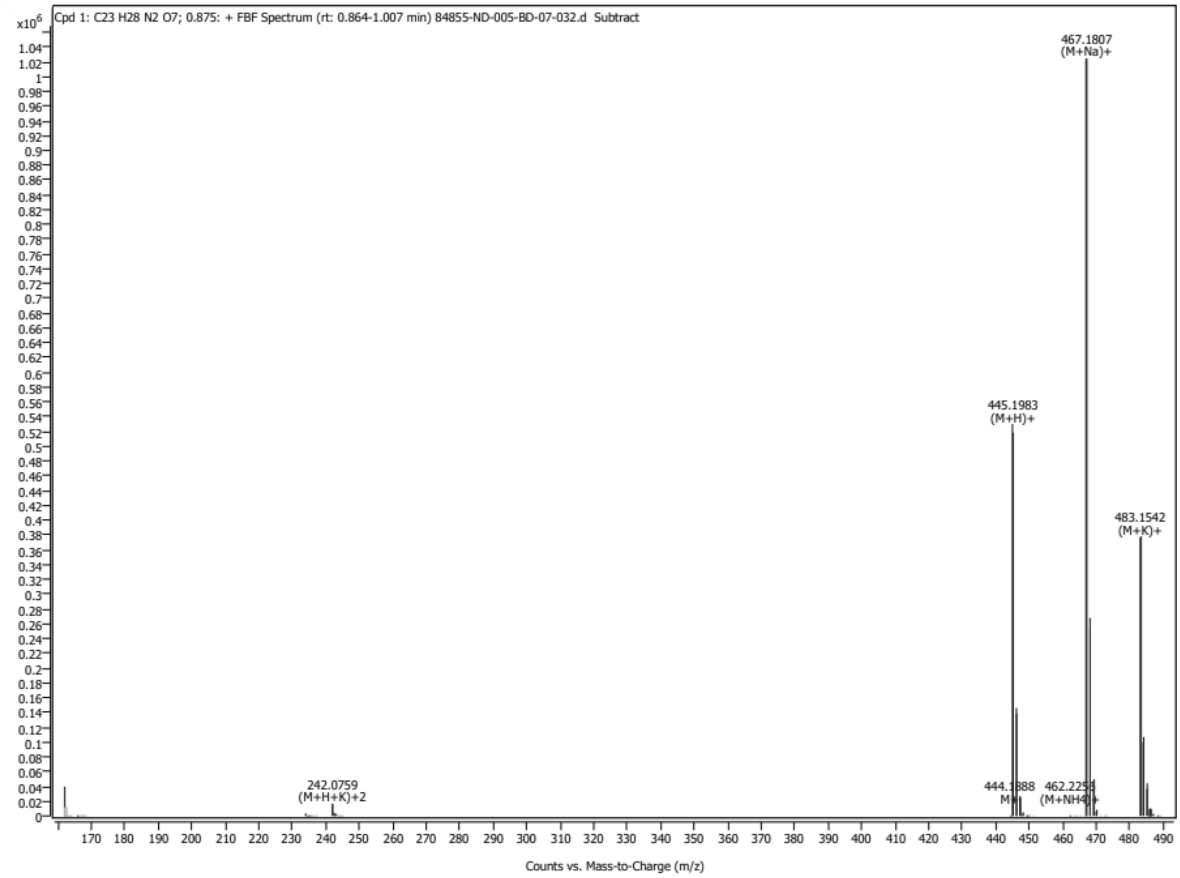

Cpd 1: C11 H20 N2 O5; 0.844: + FBF Spectrum (rt: 0.844 min) 84839_ND-002 BD07-030+VE.d Subtract
x10^6
261.1447
(M+H)+
262.1476
(M+H)+
263.1493
(M+H)+
264.1541
(M+H)+
Counts vs. Mass-to-Charge (m/z)

1Nap-VD-OMe **G3**

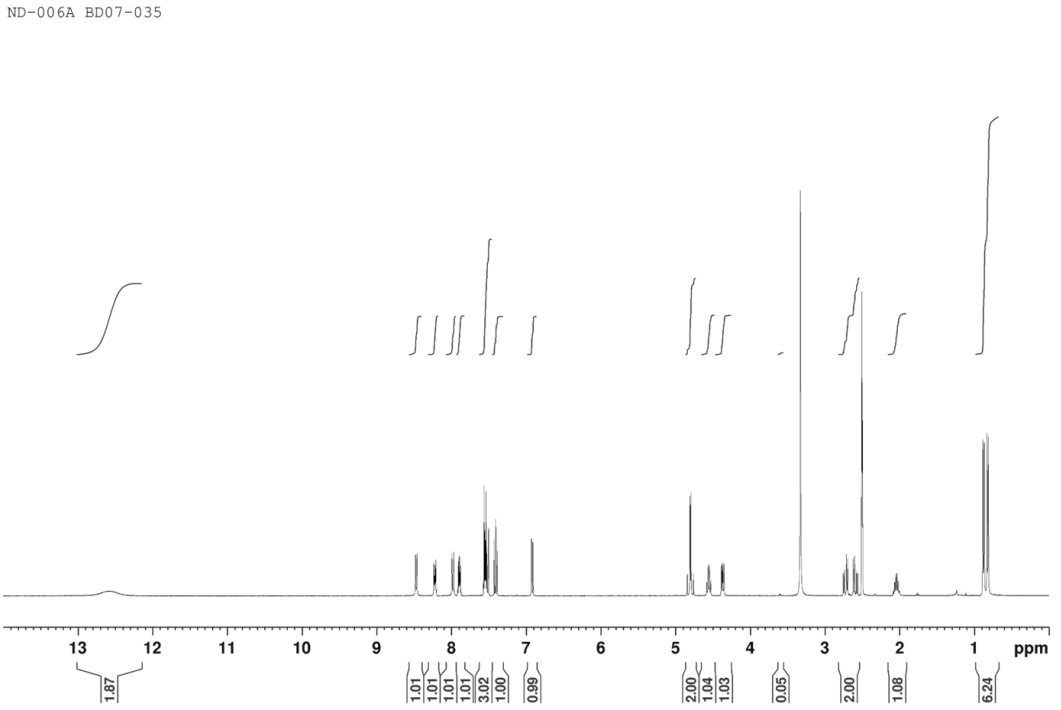


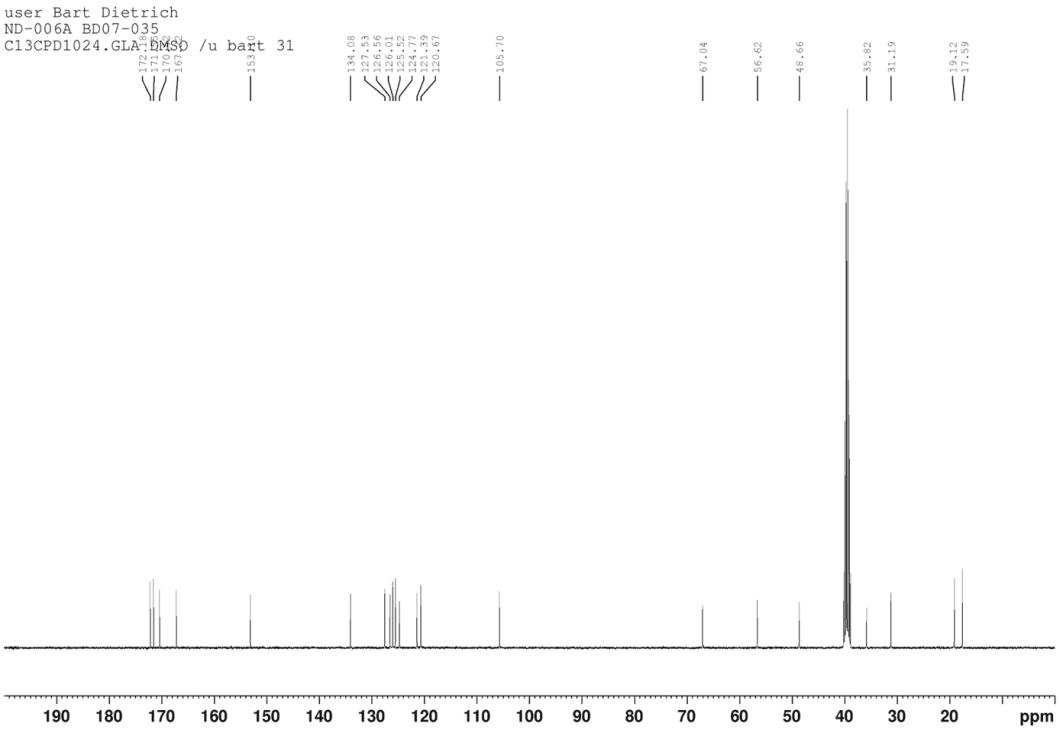


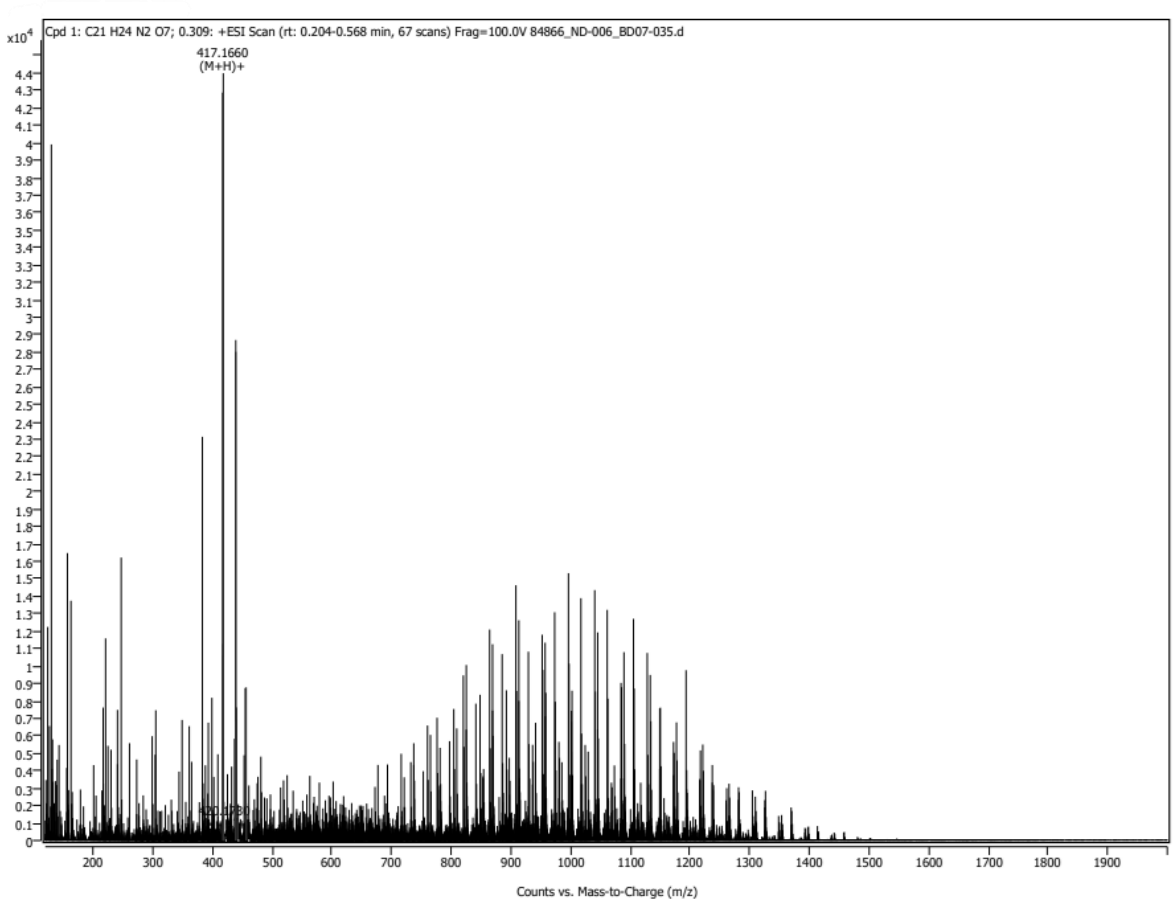

Cpd 1: C23 H28 N2 O7; 0.875: +ESI Scan (rt: 0.864-1.007 min, 27 scans) Frag=100.0V 84855-ND-005-BD-07-032.d
x10^6
467.1807
(M+Na)+
242.0759
(M+H+K)+2
Counts vs. Mass-to-Charge (m/z)

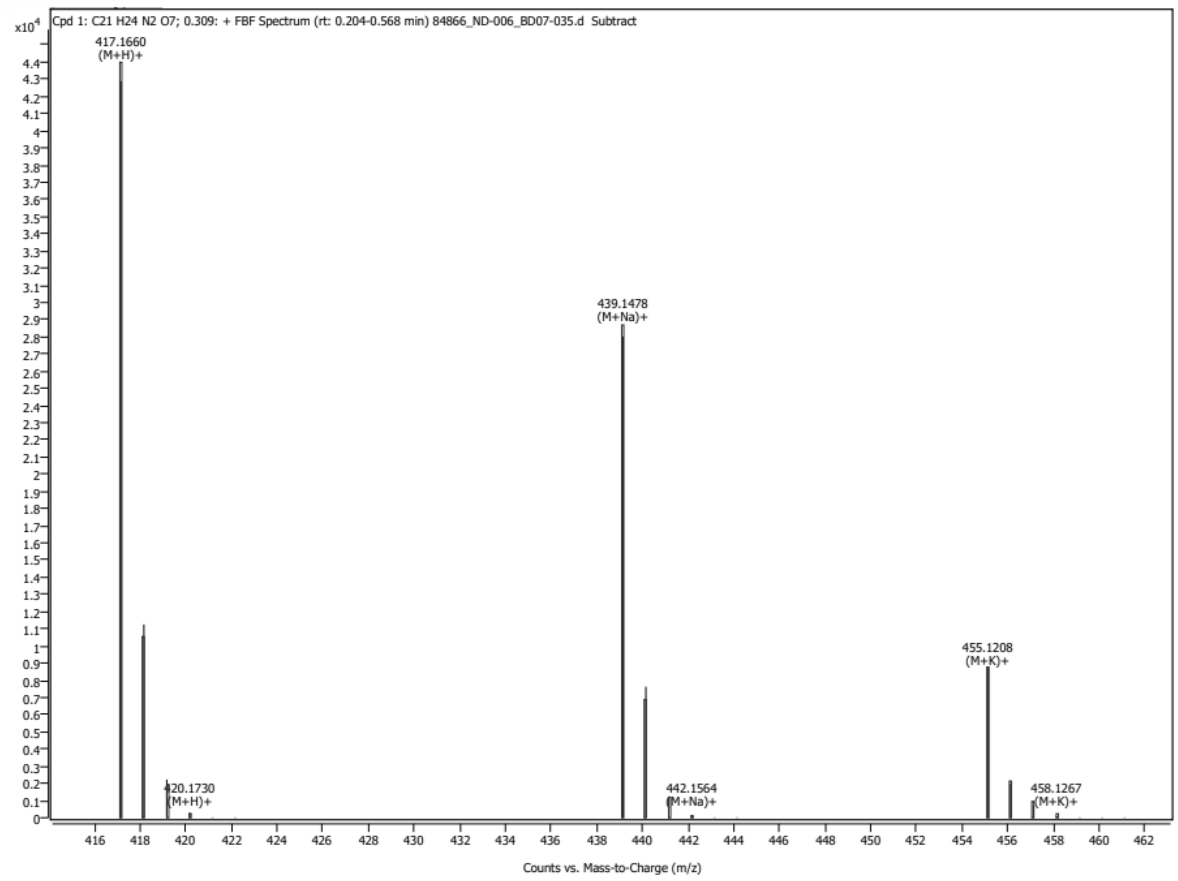

Cpd 1: C23 H28 N2 O7; 0.875: + FBF Spectrum (rt: 0.864-1.007 min) 84855-ND-005-BD-07-032.d Subtract
x10^6
467.1807
(M+Na)+
445.1983
(M+H)+
483.1542
(M+K)+
242.0759
(M+H+K)+2
Counts vs. Mass-to-Charge (m/z)

1Nap-VD **54**

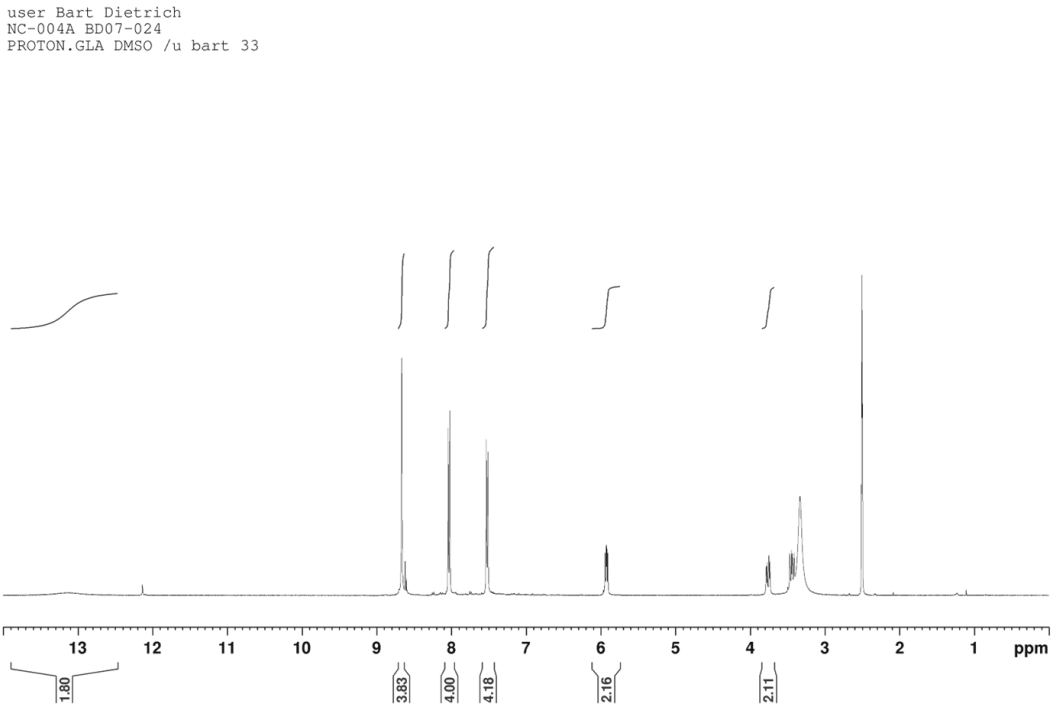

ND-006A BD07-035
13 12 11 10 9 8 7 6 5 4 3 2 1 ppm
1.87
1.01
1.01
1.01
3.02
1.00
0.99
2.00
1.04
1.03
0.05
2.00
1.08
6.24

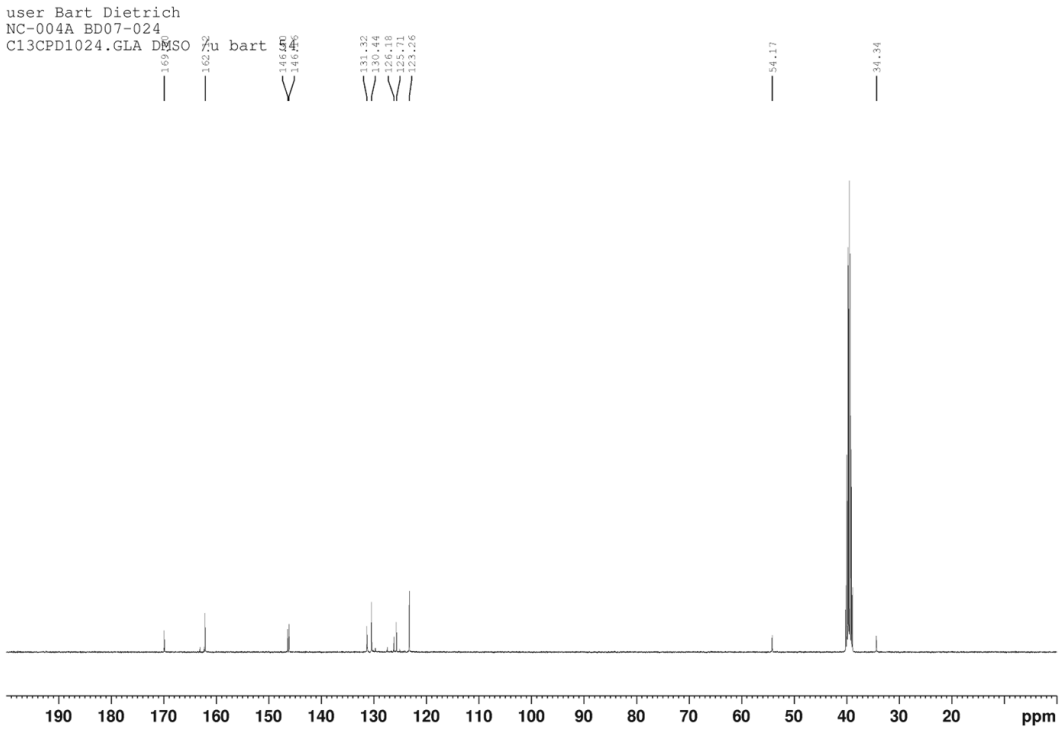

user Bart Dietrich
ND-006A BD07-035
C13CPD1024.GLA DMSO /u bart 31
172.18
171.85
171.50
170.80
167.71
153.40
134.08
127.53
126.56
126.01
125.52
124.77
121.39
120.67
105.70
67.04
56.62
48.66
35.82
31.19
19.12
17.59
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 ppm

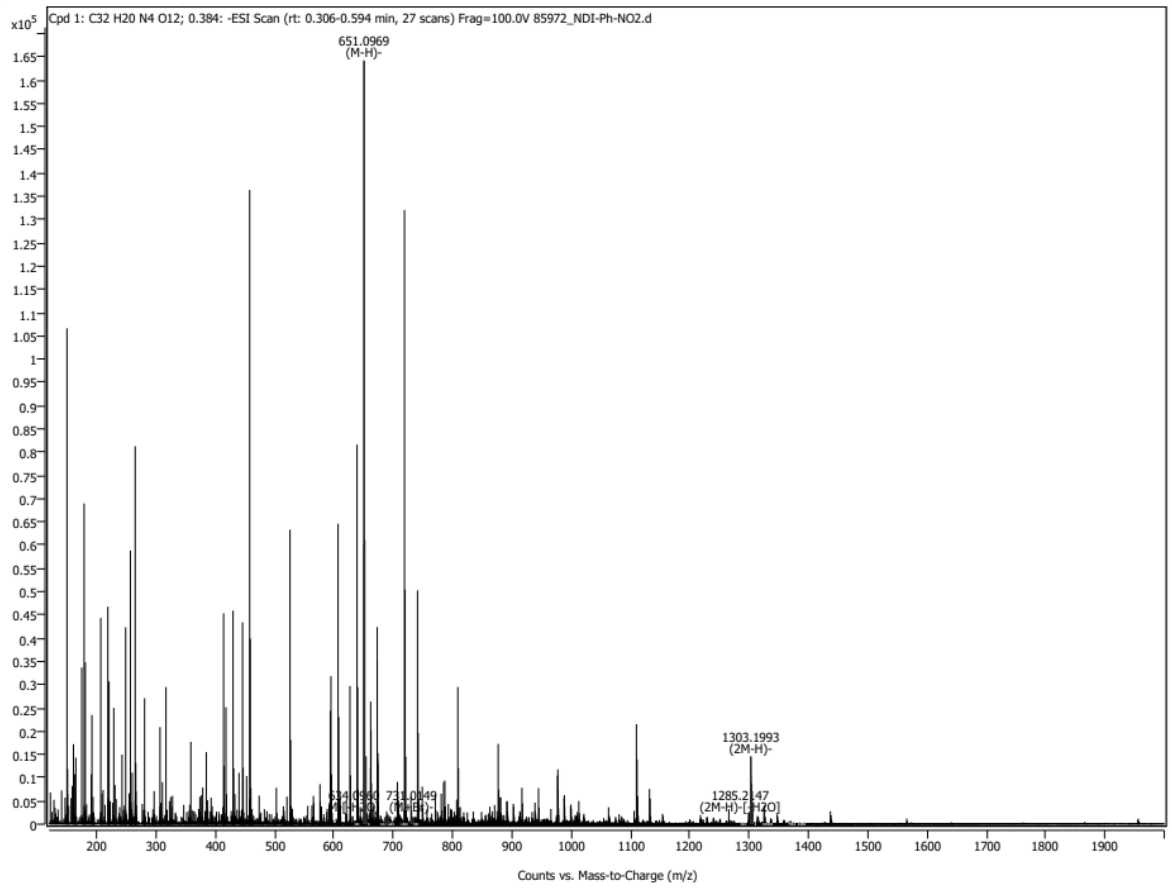

Cpd 1: C21 H24 N2 O7; 0.309: +ESI Scan (rt: 0.204-0.568 min, 67 scans) Frag=100.0V 84866_ND-006_BD07-035.d
417.1660
(M+H)+
x10^4
Counts vs. Mass-to-Charge (m/z)

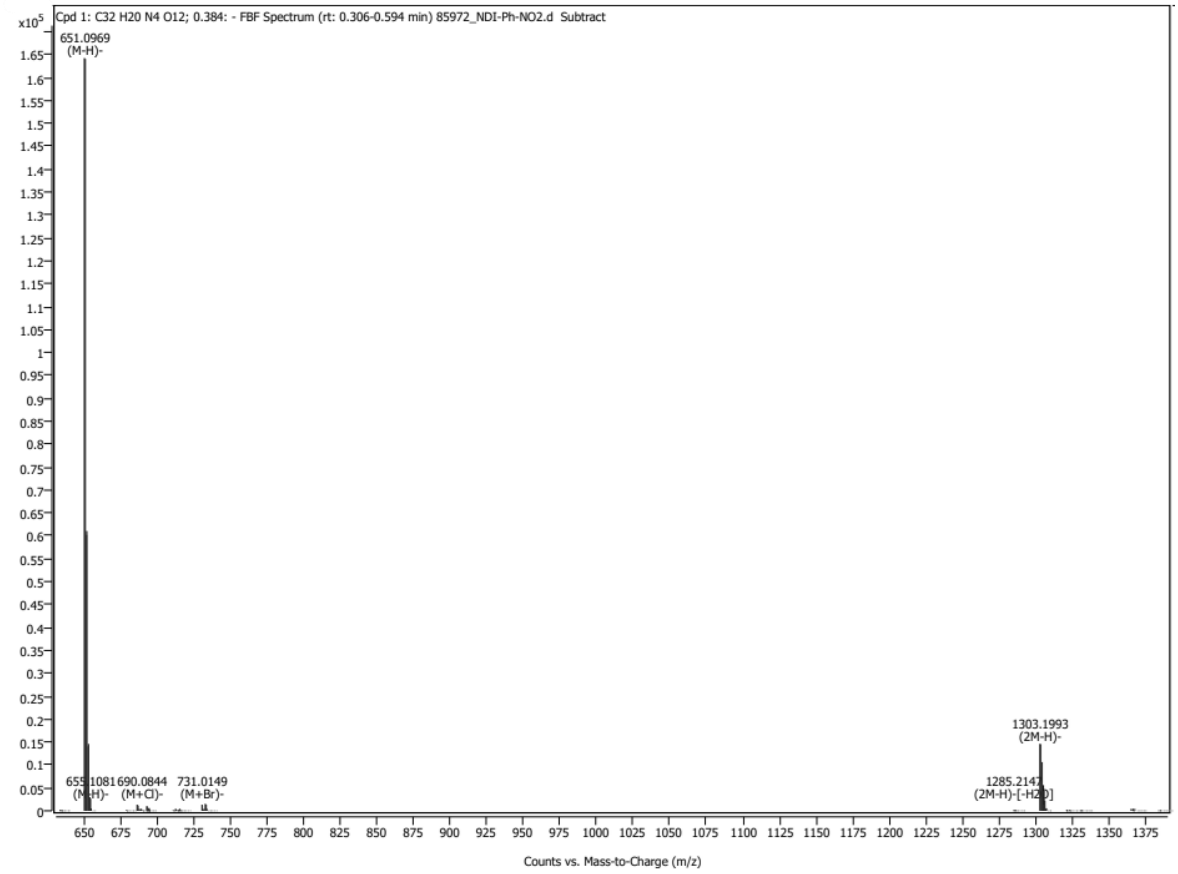

Cpd 1: C21 H24 N2 O7; 0.309: + FBF Spectrum (rt: 0.204-0.568 min) 84866_ND-006_BD07-035.d Subtract
417.1660
(M+H)+
439.1478
(M+Na)+
455.1208
(M+K)+
420.1730
(M+H)+
442.1564
(M+Na)+
458.1267
(M+K)+
x10^4
Counts vs. Mass-to-Charge (m/z)

NDI-Phe-NO_2 **55**

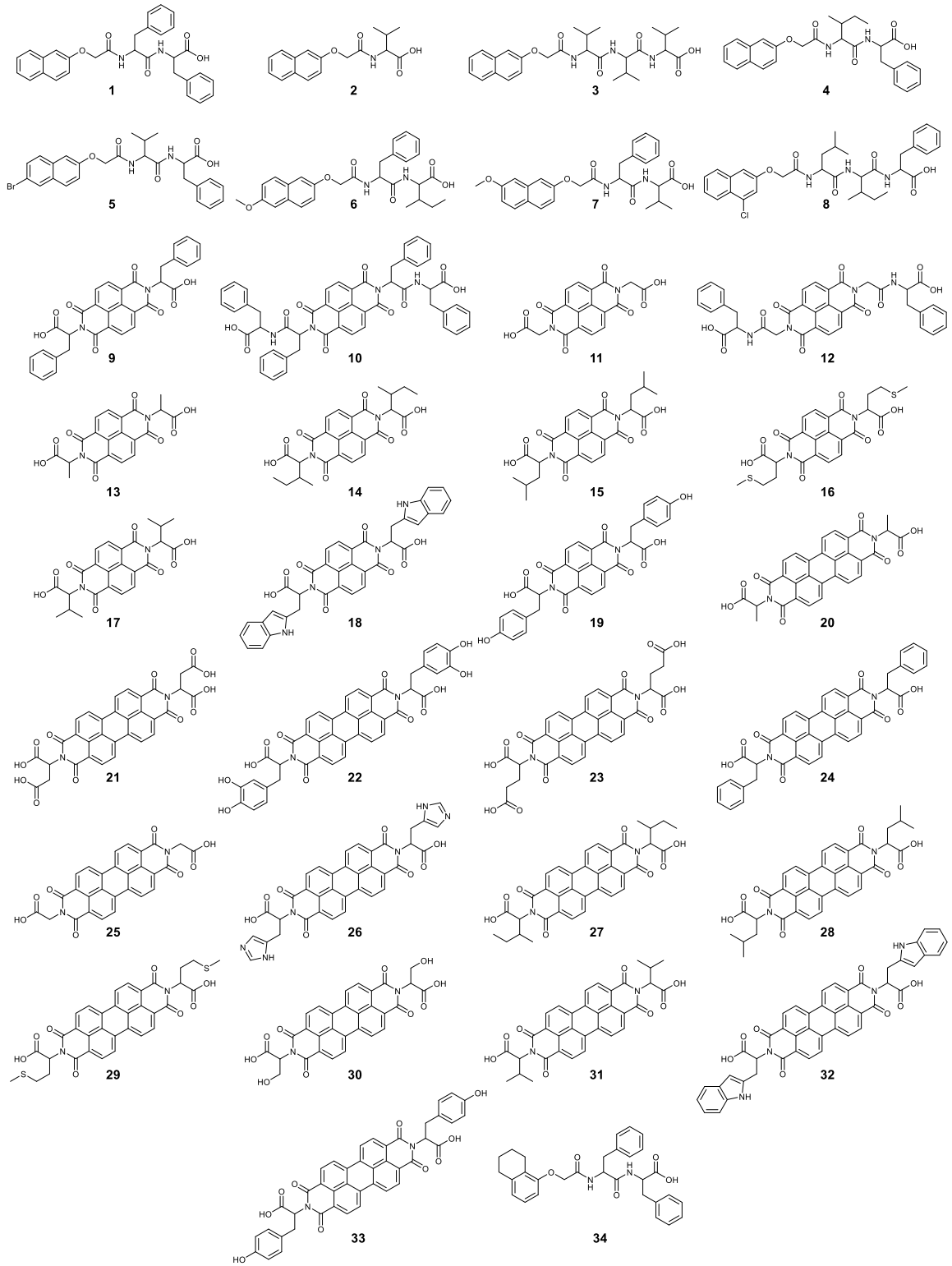


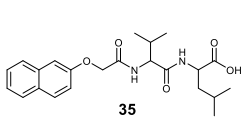


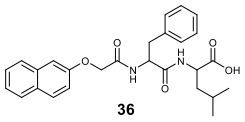

Cpd 1: C32 H20 N4 O12; 0.384: -ESI Scan (rt: 0.306-0.594 min, 27 scans) Frag=100.0V 85972_NDI-Ph-NO2.d
x10^5
651.0969
(M-H)-
1303.1993
(2M-H)-
1285.2147
(2M-H)-[-H2O]
731.0149
(M+Br)-
Counts vs. Mass-to-Charge (m/z)

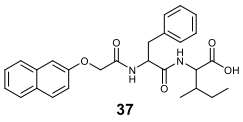

Cpd 1: C32 H20 N4 O12; 0.384: - FBF Spectrum (rt: 0.306-0.594 min) 85972_NDI-Ph-NO2.d Subtract
x10^5
651.0969
(M-H)-
1303.1993
(2M-H)-
1285.2147
(2M-H)-[-H2O]
731.0149
(M+Br)-
Counts vs. Mass-to-Charge (m/z)

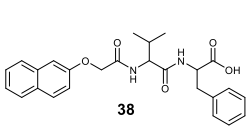
A 3.3 Model Compounds

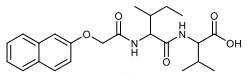


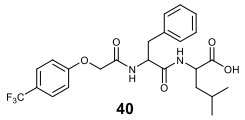

35

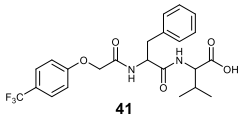

36

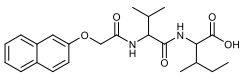

37

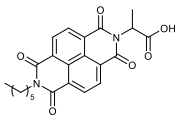

38

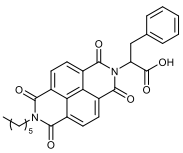

39

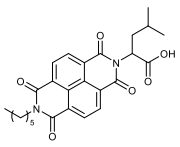

40

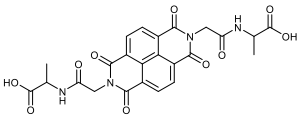

41

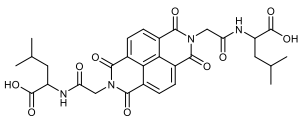

42

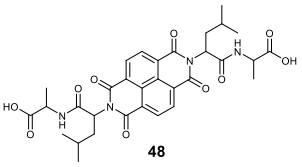

43

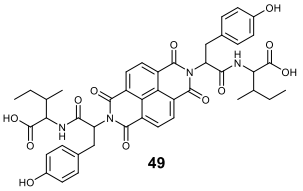

44

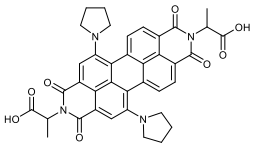

45

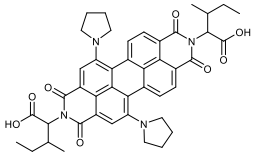

46

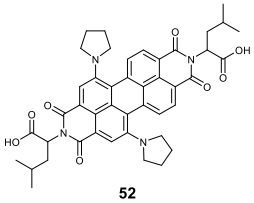

47

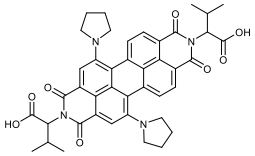

48

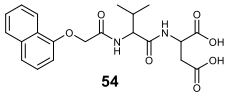

49

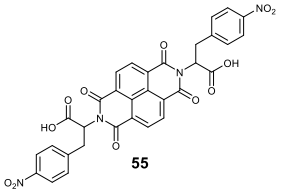

50


51


52

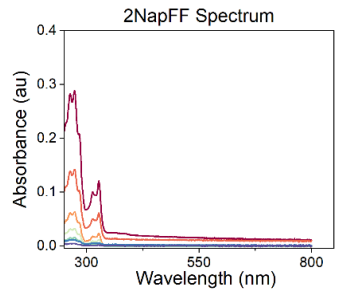
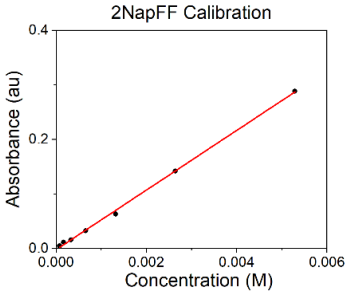
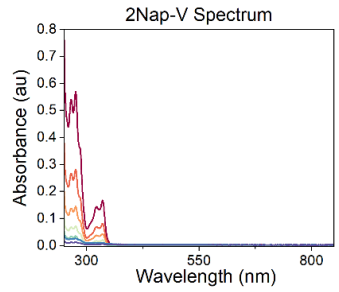
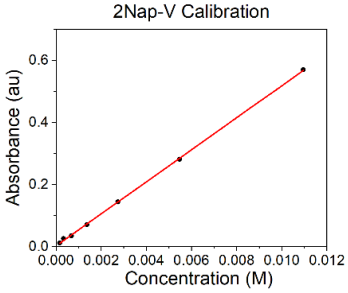
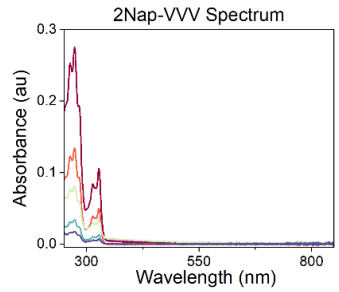
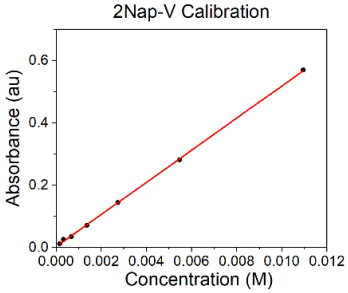
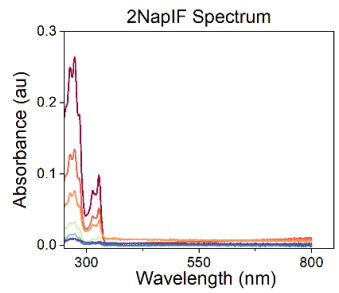
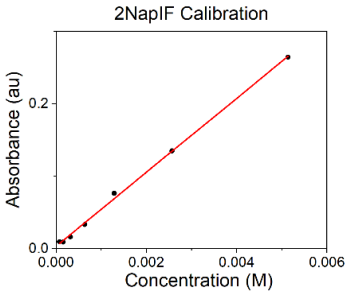

53

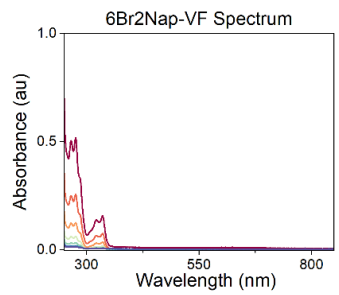
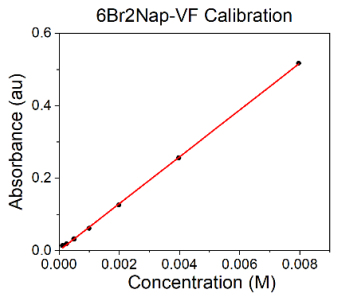
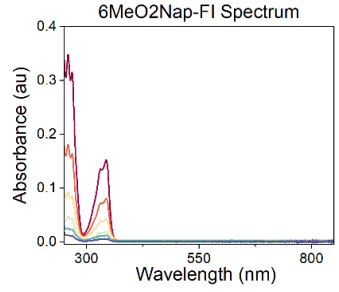
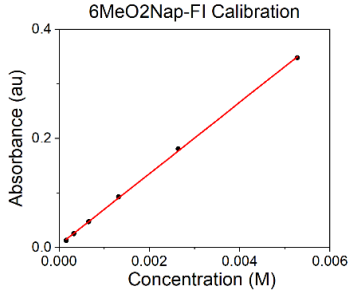
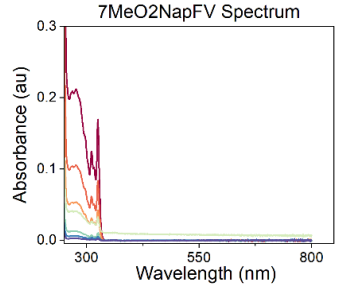
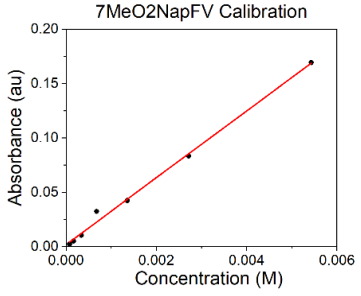
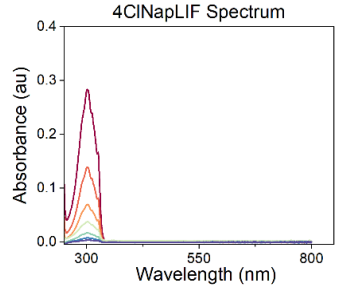
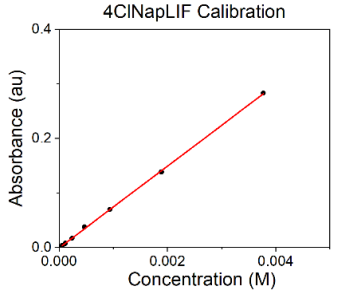
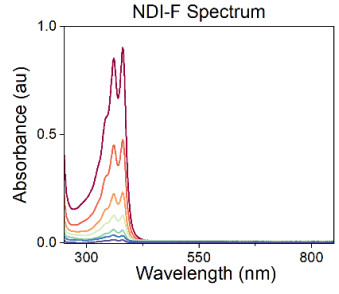
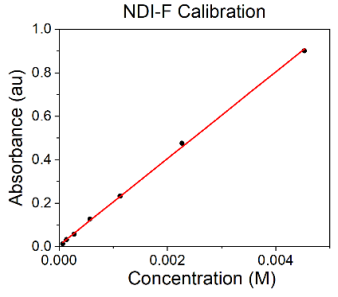
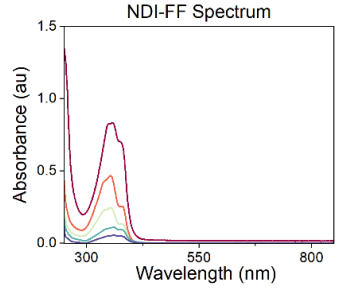
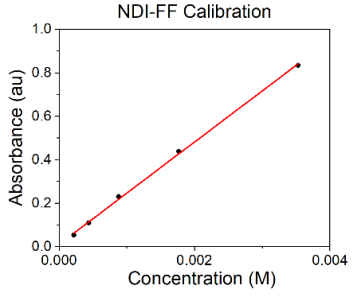

54

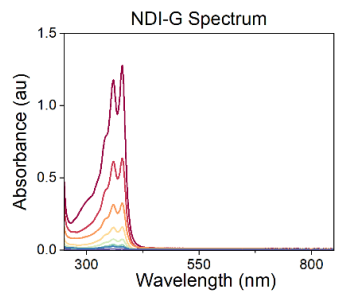
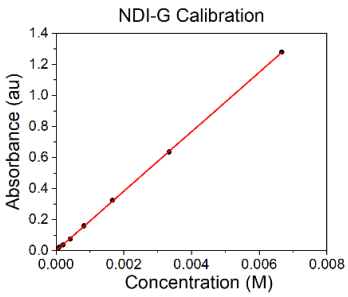
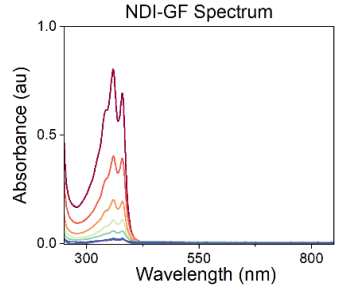
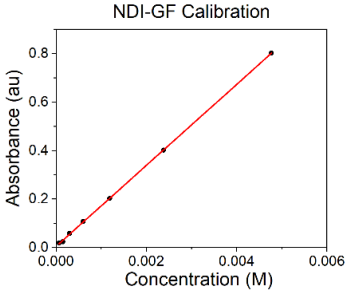
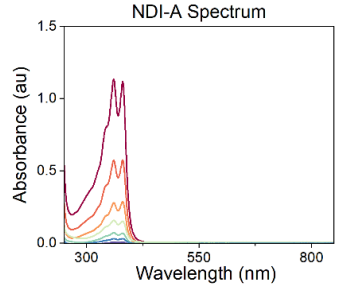
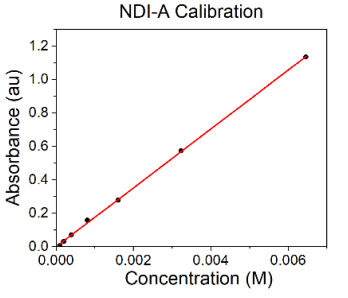
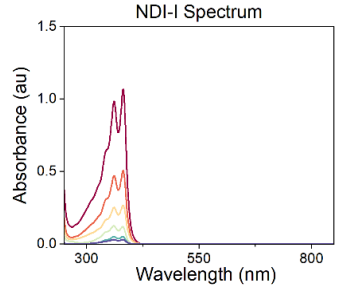
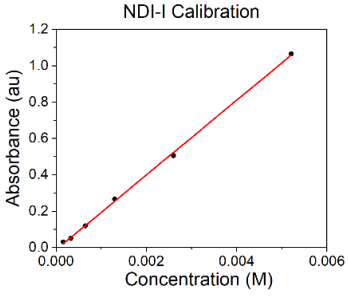
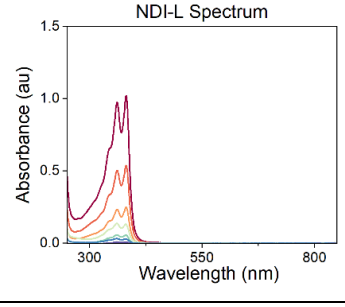
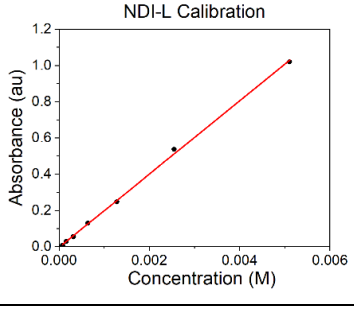
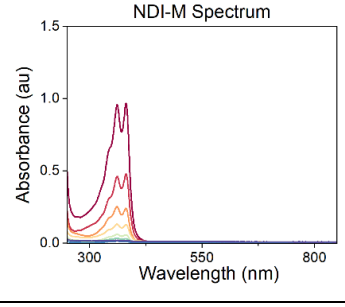
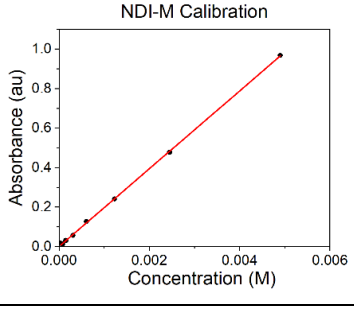

55

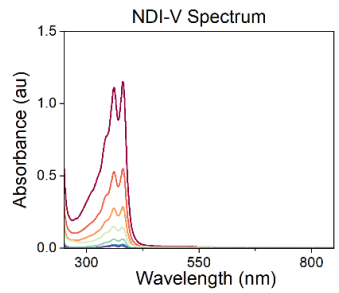
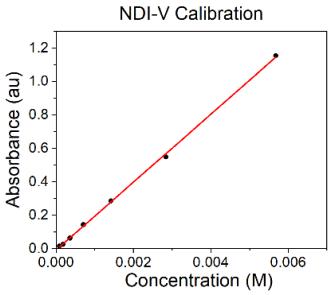
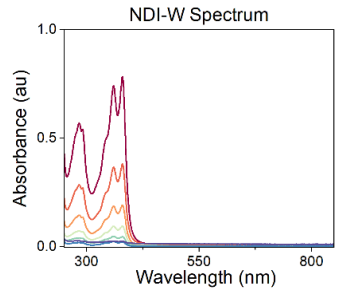
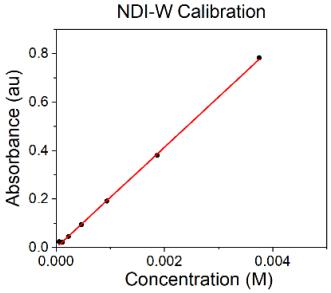
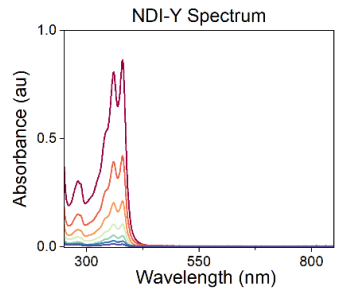
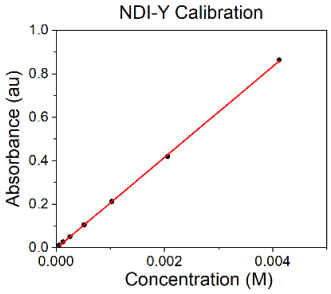
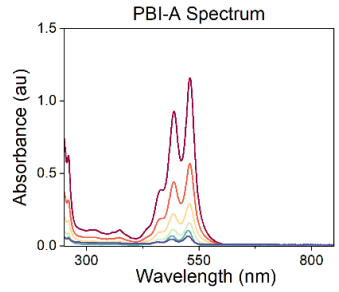
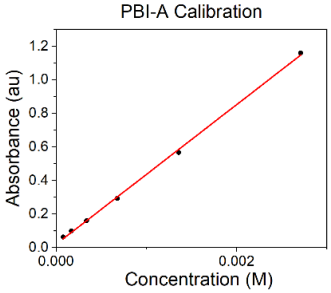
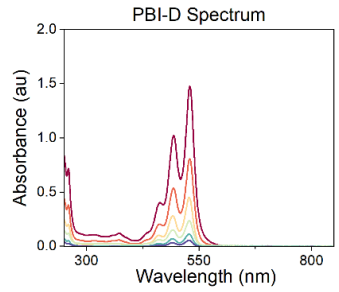
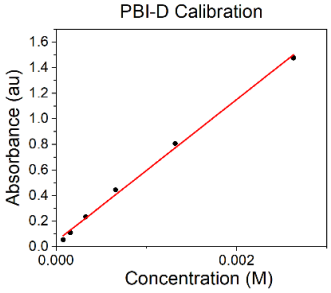
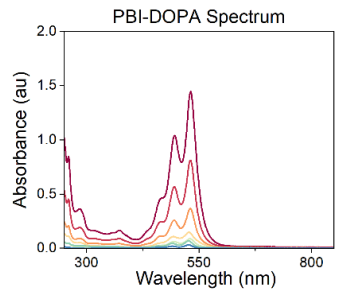
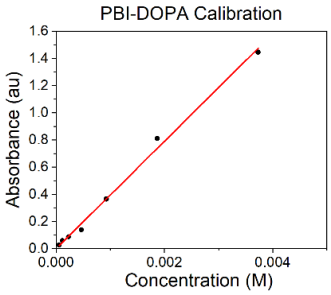
A 3.4 Calibration Data

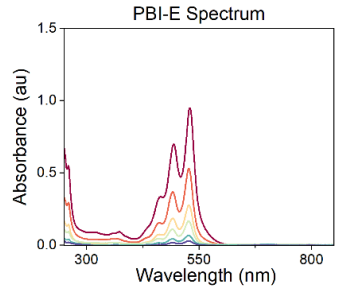
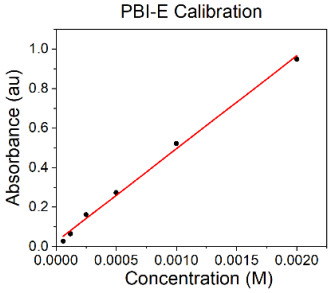
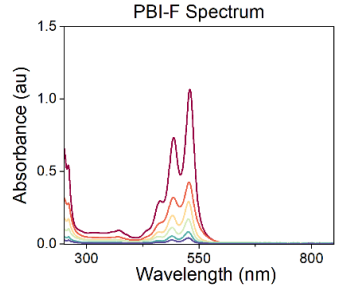
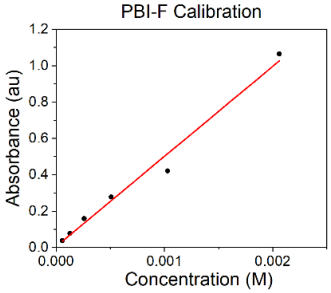
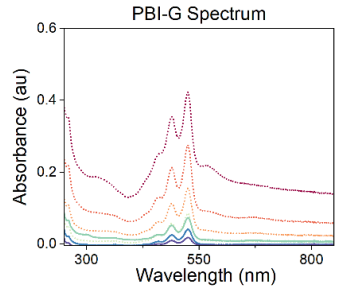
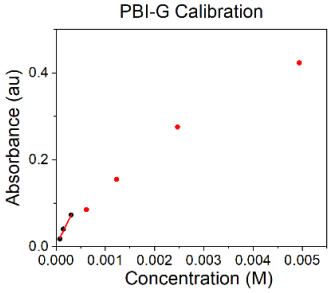
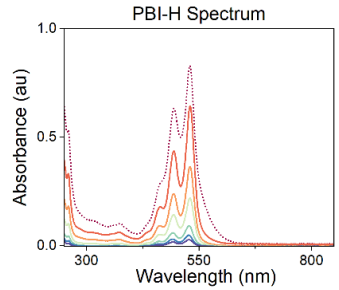
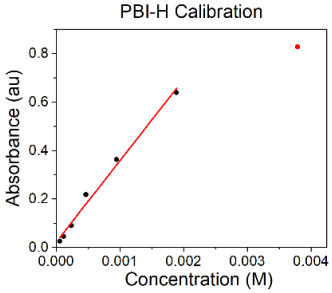
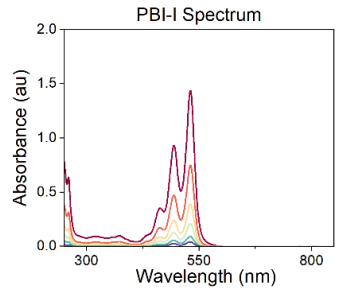
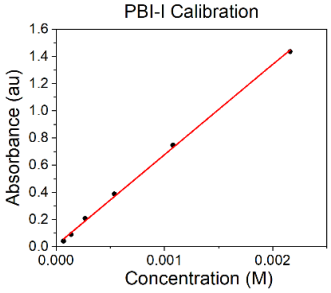
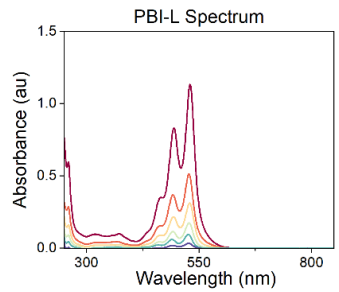
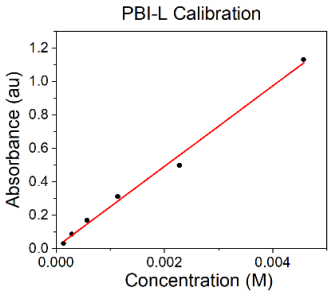
UV-vis dilution series and calibration curve in DMSO for each training, test, and validation compound. In each case the highest concentration spectra is given in red (▬) and the lowest in purple (▬). Dotted spectra show datapoints that deviate from linearity, indicating saturation at the lower calibration concentration and are subsequently rejected from the calibration curve. The calibration window for a measurement in a 0.1 mm cuvette, molar absorption coefficient, and R^2 goodness-of-fit metrics are reported.

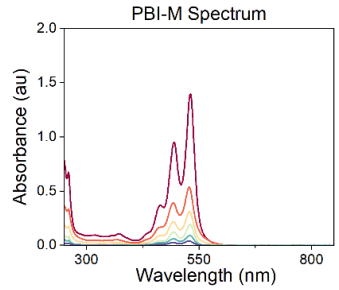
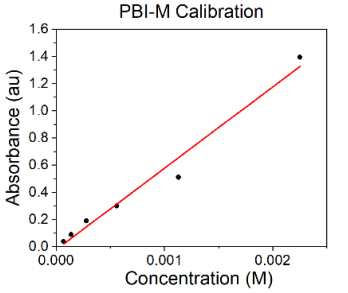
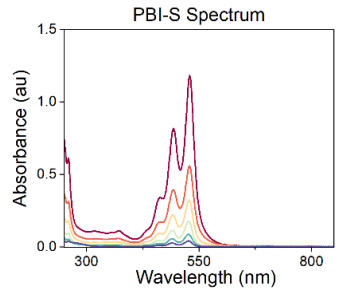
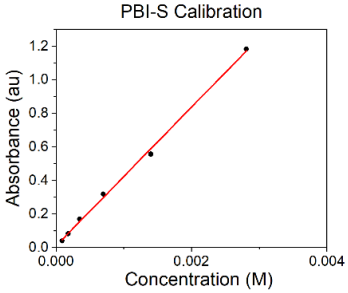
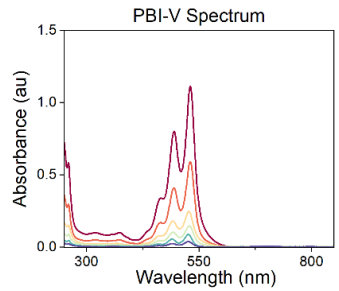
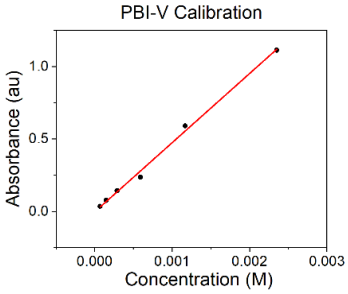
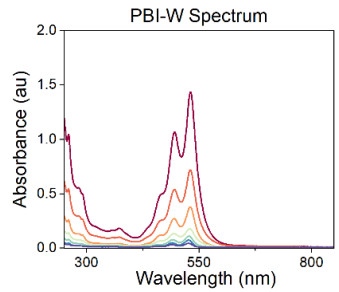
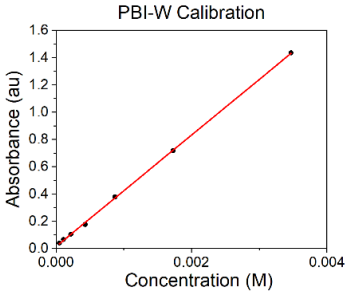
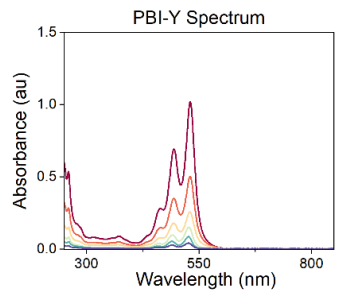
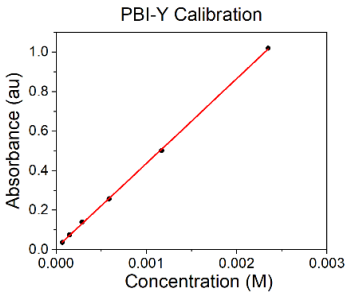
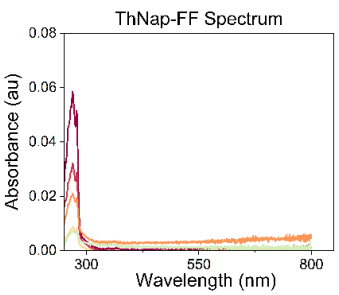
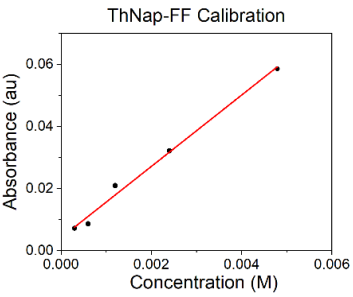
<table>
<tr>
<td>
</td>
<td>
</td>
<td colspan="2">2Nap-FF (1)</td>
</tr>
<tr>
<td rowspan="4"></td>
<td rowspan="4"></td>
<td>Min Conc. (mg mL^{-1})</td>
<td>0.041</td>
</tr>
<tr>
<td>Max Conc. (mg mL^{-1})</td>
<td>2.625</td>
</tr>
<tr>
<td>ε (M^{-1} cm^{-1})</td>
<td>5463</td>
</tr>
<tr>
<td>R^2 (au)</td>
<td>0.99</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td colspan="2">2Nap-V (2)</td>
</tr>
<tr>
<td rowspan="4"></td>
<td rowspan="4"></td>
<td>Min Conc. (mg mL^{-1})</td>
<td>0.050</td>
</tr>
<tr>
<td>Max Conc. (mg mL^{-1})</td>
<td>3.300</td>
</tr>
<tr>
<td>ε (M^{-1} cm^{-1})</td>
<td>5157</td>
</tr>
<tr>
<td>R^2 (au)</td>
<td>1.00</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td colspan="2">2Nap-VVV (3)</td>
</tr>
<tr>
<td rowspan="4"></td>
<td rowspan="4"></td>
<td>Min Conc. (mg mL^{-1})</td>
<td>0.175</td>
</tr>
<tr>
<td>Max Conc. (mg mL^{-1})</td>
<td>2.800</td>
</tr>
<tr>
<td>ε (M^{-1} cm^{-1})</td>
<td>4853</td>
</tr>
<tr>
<td>R^2 (au)</td>
<td>1.00</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td colspan="2">2Nap-IF (4)</td>
</tr>
<tr>
<td rowspan="4"></td>
<td rowspan="4"></td>
<td>Min Conc. (mg mL^{-1})</td>
<td>0.038</td>
</tr>
<tr>
<td>Max Conc. (mg mL^{-1})</td>
<td>2.430</td>
</tr>
<tr>
<td>ε (M^{-1} cm^{-1})</td>
<td>5104</td>
</tr>
<tr>
<td>R^2 (au)</td>
<td>1.00</td>
</tr>
</table>

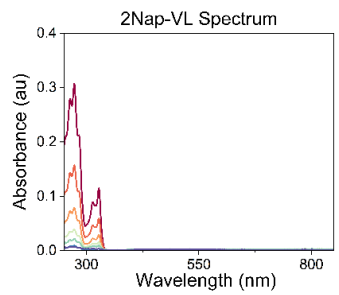
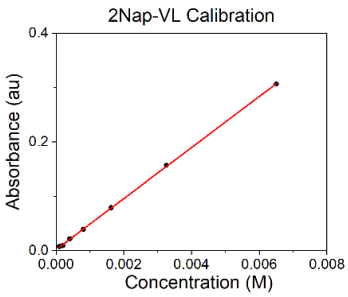
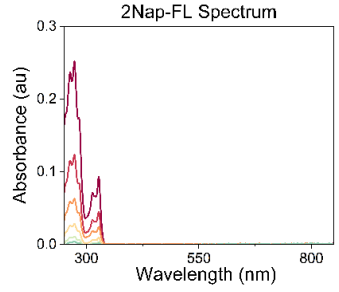
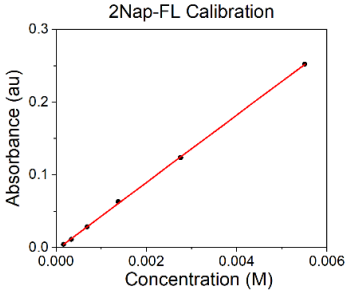
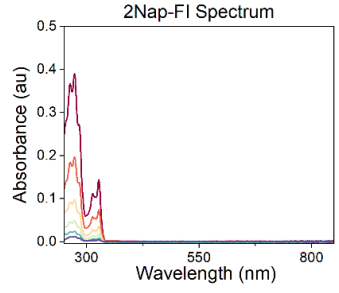
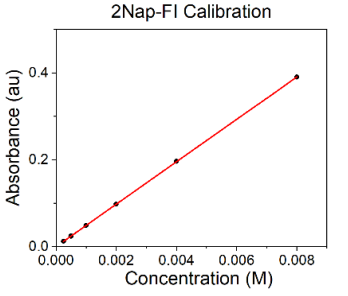
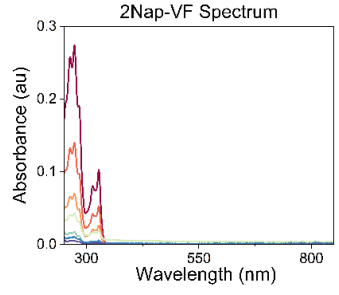
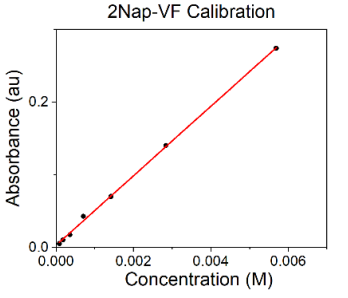
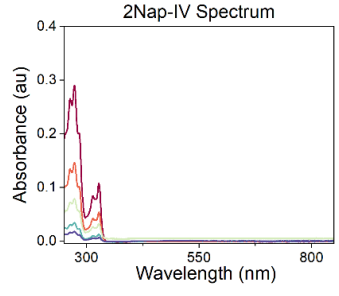
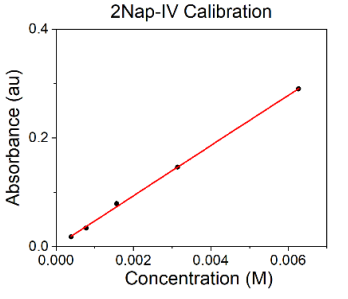
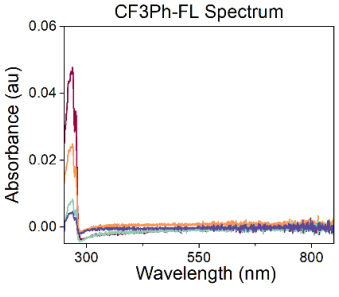
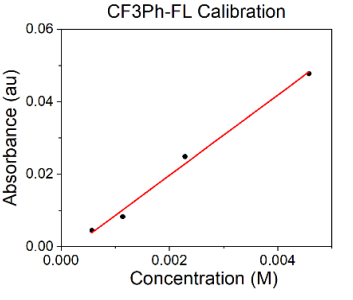
		**6Br2Nap-VF (5)**	
		Min Conc. (mg mL^{-1})	0.065
		Max Conc. (mg mL^{-1})	4.200
		ε (M^{-1} cm^{-1})	6461
		R^2 (au)	1.00
		**6MeO2Nap-FI (6)**	
		Min Conc. (mg mL^{-1})	0.081
		Max Conc. (mg mL^{-1})	2.600
		ε (M^{-1} cm^{-1})	6540
		R^2 (au)	1.00
		**7MeO2Nap-FV (7)**	
		Min Conc. (mg mL^{-1})	0.041
		Max Conc. (mg mL^{-1})	2.600
		ε (M^{-1} cm^{-1})	3071
		R^2 (au)	0.99
		**4ClNap-LIF (8)**	
		Min Conc. (mg mL^{-1})	0.035
		Max Conc. (mg mL^{-1})	2.250
		ε (M^{-1} cm^{-1})	7498
		R^2 (au)	1.00
		**NDI-F (9)**	
		Min Conc. (mg mL^{-1})	0.040
		Max Conc. (mg mL^{-1})	2.550
		ε (M^{-1} cm^{-1})	19931
		R^2 (au)	1.00
		**NDI-FF (10)**	
		Min Conc. (mg mL^{-1})	0.189
		Max Conc. (mg mL^{-1})	3.030
		ε (M^{-1} cm^{-1})	23486
		R^2 (au)	1.00

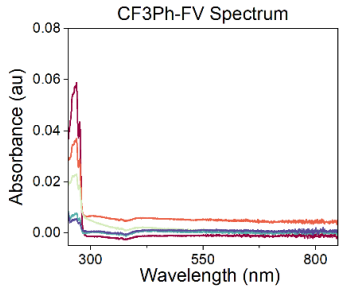
		**NDI-G (11)** Min Conc. (mg mL^{-1}) 0.020 Max Conc. (mg mL^{-1}) 2.550 ε (M^{-1} cm^{-1}) 19152 R^2 (au) 1.00
		**NDI-GF (12)** Min Conc. (mg mL^{-1}) 0.050 Max Conc. (mg mL^{-1}) 3.230 ε (M^{-1} cm^{-1}) 16711 R^2 (au) 1.00
		**NDI-A (13)** Min Conc. (mg mL^{-1}) 0.041 Max Conc. (mg mL^{-1}) 2.650 ε (M^{-1} cm^{-1}) 17616 R^2 (au) 1.00
		**NDI-I (14)** Min Conc. (mg mL^{-1}) 0.080 Max Conc. (mg mL^{-1}) 2.580 ε (M^{-1} cm^{-1}) 20552 R^2 (au) 1.00
		**NDI-L (15)** Min Conc. (mg mL^{-1}) 0.039 Max Conc. (mg mL^{-1}) 2.530 ε (M^{-1} cm^{-1}) 20230 R^2 (au) 1.00
		**NDI-M (16)** Min Conc. (mg mL^{-1}) 0.020 Max Conc. (mg mL^{-1}) 2.600 ε (M^{-1} cm^{-1}) 19660 R^2 (au) 1.00

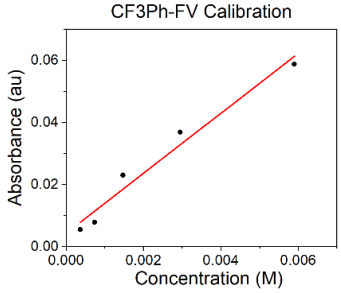
		**NDI-V (17)** Min Conc. (mg mL^{-1}) 0.041 Max Conc. (mg mL^{-1}) 2.650 ε (M^{-1} cm^{-1}) 20309 R^2 (au) 1.00
		**NDI-W (18)** Min Conc. (mg mL^{-1}) 0.038 Max Conc. (mg mL^{-1}) 2.400 ε (M^{-1} cm^{-1}) 20751 R^2 (au) 1.00
		**NDI-Y (19)** Min Conc. (mg mL^{-1}) 0.038 Max Conc. (mg mL^{-1}) 2.450 ε (M^{-1} cm^{-1}) 20930 R^2 (au) 1.00
		**PBI-A (20)** Min Conc. (mg mL^{-1}) 0.045 Max Conc. (mg mL^{-1}) 1.450 ε (M^{-1} cm^{-1}) 41631 R^2 (au) 1.00
		**PBI-D (21)** Min Conc. (mg mL^{-1}) 0.013 Max Conc. (mg mL^{-1}) 1.640 ε (M^{-1} cm^{-1}) 55454 R^2 (au) 1.00
		**PBI-DOPA (22)** Min Conc. (mg mL^{-1}) 0.044 Max Conc. (mg mL^{-1}) 2.800 ε (M^{-1} cm^{-1}) 39500 R^2 (au) 0.99

		**PBI-E (23)** Min Conc. (mg mL^{-1}) 0.041 Max Conc. (mg mL^{-1}) 1.300 ε (M^{-1} cm^{-1}) 47055 R^2 (au) 0.99
		**PBI-F (24)** Min Conc. (mg mL^{-1}) 0.044 Max Conc. (mg mL^{-1}) 1.410 ε (M^{-1} cm^{-1}) 49572 R^2 (au) 0.98
		**PBI-G (25)** Min Conc. (mg mL^{-1}) 0.039 Max Conc. (mg mL^{-1}) 0.150 ε (M^{-1} cm^{-1}) 23222 R^2 (au) 0.97
		**PBI-H (26)** Min Conc. (mg mL^{-1}) 0.039 Max Conc. (mg mL^{-1}) 1.260 ε (M^{-1} cm^{-1}) 33619 R^2 (au) 0.99
		**PBI-I (27)** Min Conc. (mg mL^{-1}) 0.042 Max Conc. (mg mL^{-1}) 1.340 ε (M^{-1} cm^{-1}) 66426 R^2 (au) 1.00
		**PBI-L (28)** Min Conc. (mg mL^{-1}) 0.088 Max Conc. (mg mL^{-1}) 2.830 ε (M^{-1} cm^{-1}) 24057 R^2 (au) 0.99

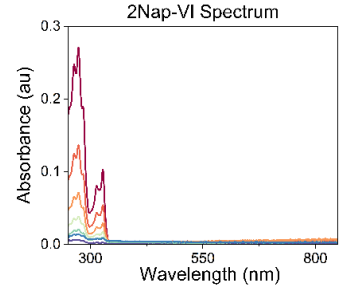
		**PBI-M (29)**	
		Min Conc. (mg mL^{-1})	0.046
		Max Conc. (mg mL^{-1})	1.480
		ε (M^{-1} cm^{-1})	59885
		R^2 (au)	0.97
		**PBI-S (30)**	
		Min Conc. (mg mL^{-1})	0.047
		Max Conc. (mg mL^{-1})	1.500
		ε (M^{-1} cm^{-1})	41343
		R^2 (au)	1.00
		**PBI-V (31)**	
		Min Conc. (mg mL^{-1})	0.043
		Max Conc. (mg mL^{-1})	1.390
		ε (M^{-1} cm^{-1})	47714
		R^2 (au)	1.00
		**PBI-W (32)**	
		Min Conc. (mg mL^{-1})	0.041
		Max Conc. (mg mL^{-1})	2.650
		ε (M^{-1} cm^{-1})	40852
		R^2 (au)	1.00
		**PBI-Y (33)**	
		Min Conc. (mg mL^{-1})	0.043
		Max Conc. (mg mL^{-1})	1.390
		ε (M^{-1} cm^{-1})	42831
		R^2 (au)	1.00
		**ThNap-FF (34)**	
		Min Conc. (mg mL^{-1})	0.019
		Max Conc. (mg mL^{-1})	2.400
		ε (M^{-1} cm^{-1})	1149
		R^2 (au)	0.99

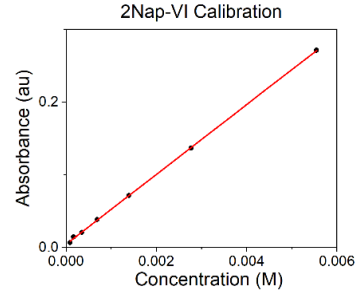
		**2Nap-VL (35)** Min Conc. (mg mL^{-1}) 0.042 Max Conc. (mg mL^{-1}) 2.700 ε (M^{-1} cm^{-1}) 4698 R^2 (au) 1.00
		**2Nap-FL (36)** Min Conc. (mg mL^{-1}) 0.080 Max Conc. (mg mL^{-1}) 2.550 ε (M^{-1} cm^{-1}) 4636 R^2 (au) 1.00
		**2Nap-FI (37)** Min Conc. (mg mL^{-1}) 0.116 Max Conc. (mg mL^{-1}) 3.700 ε (M^{-1} cm^{-1}) 4880 R^2 (au) 1.00
		**2Nap-VF (38)** Min Conc. (mg mL^{-1}) 0.040 Max Conc. (mg mL^{-1}) 2.550 ε (M^{-1} cm^{-1}) 4781 R^2 (au) 1.00
		**2Nap-IV (39)** Min Conc. (mg mL^{-1}) 0.163 Max Conc. (mg mL^{-1}) 2.600 ε (M^{-1} cm^{-1}) 4624 R^2 (au) 1.00
		**CF_3Ph-FL (40)** Min Conc. (mg mL^{-1}) 0.275 Max Conc. (mg mL^{-1}) 2.200 ε (M^{-1} cm^{-1}) 1107 R^2 (au) 0.99



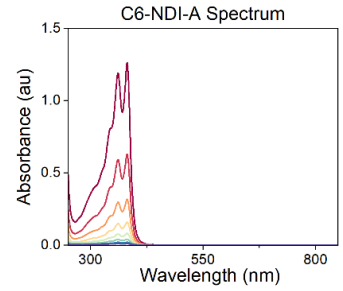


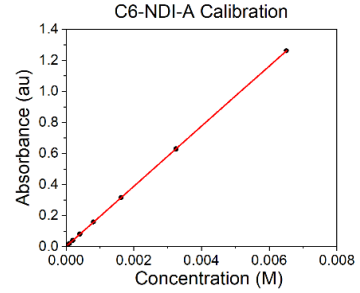
CF_3Ph-FV (41)	
Min Conc. (mg mL^{-1})	0.172
Max Conc. (mg mL^{-1})	2.750
ε (M^{-1} cm^{-1})	965
R^2 (au)	0.96



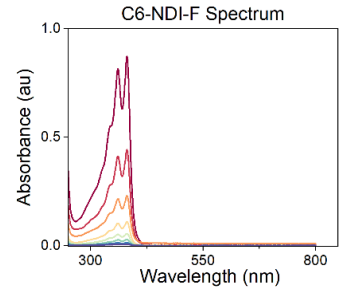


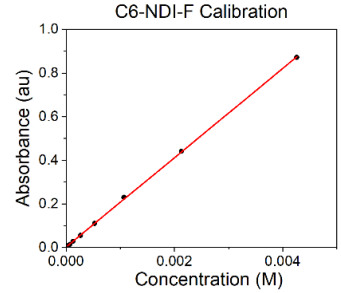
2Nap-VI (42)	
Min Conc. (mg mL^{-1})	0.036
Max Conc. (mg mL^{-1})	2.300
ε (M^{-1} cm^{-1})	4804
R^2 (au)	1.00



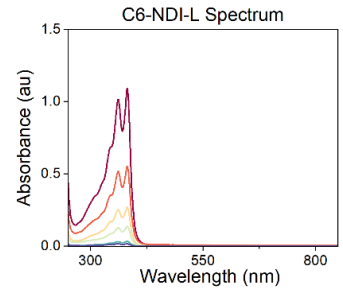


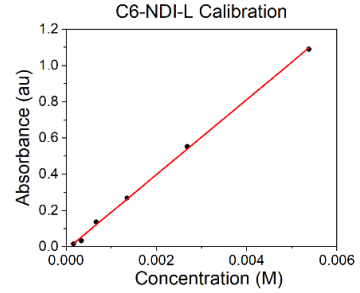
C_6-NDI-A (43)	
Min Conc. (mg mL^{-1})	0.021
Max Conc. (mg mL^{-1})	2.750
ε (M^{-1} cm^{-1})	19357
R^2 (au)	1.00



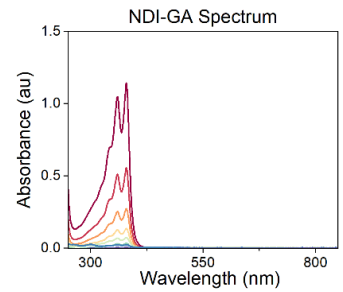


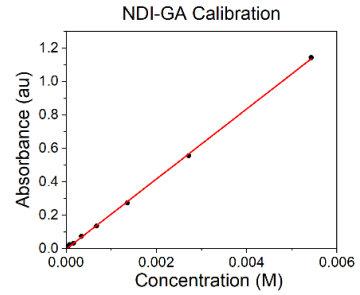
C_6-NDI-F (44)	
Min Conc. (mg mL^{-1})	0.017
Max Conc. (mg mL^{-1})	2.130
ε (M^{-1} cm^{-1})	20482
R^2 (au)	1.00



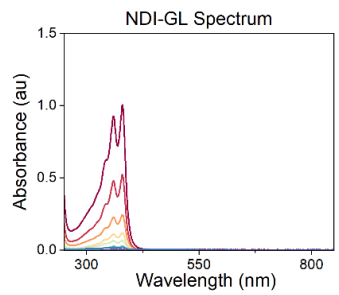
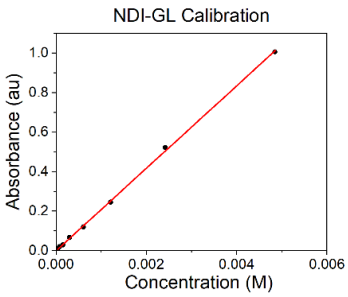
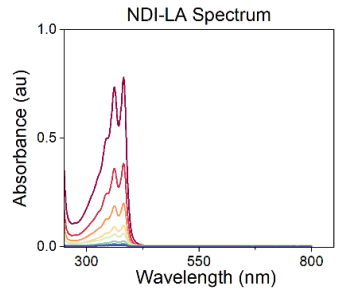
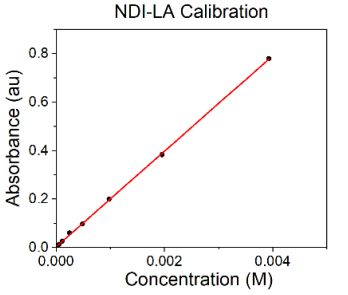
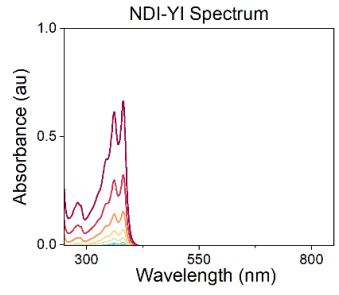
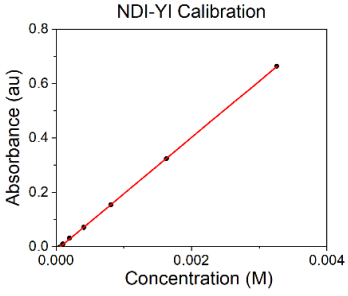
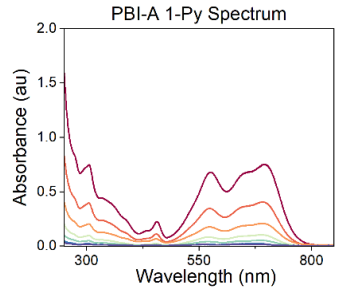
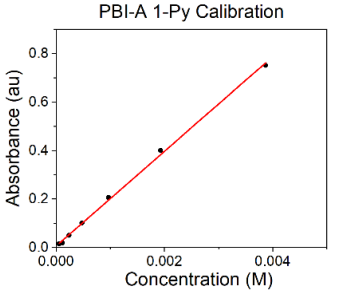
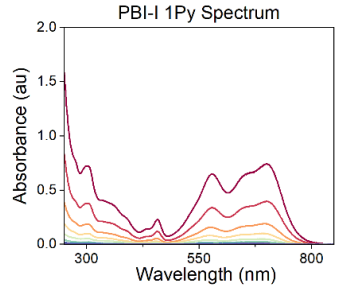
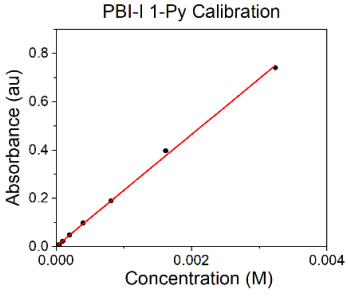
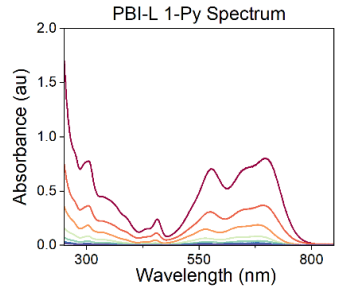
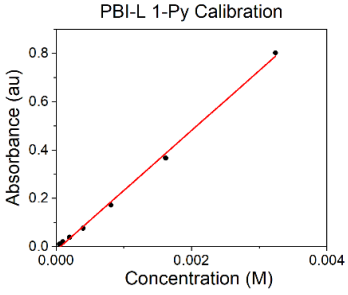


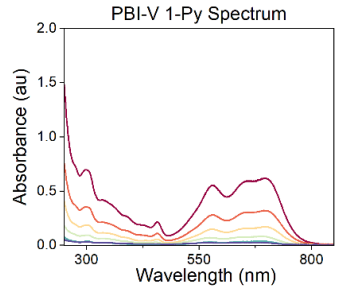
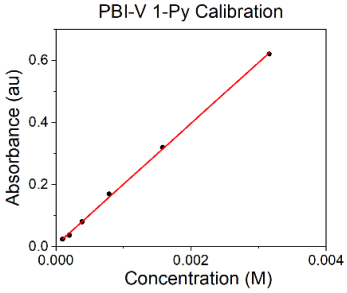
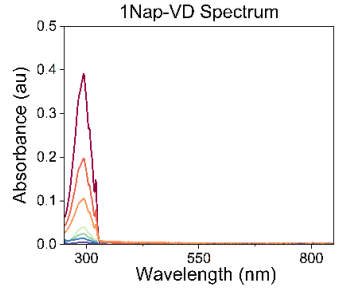
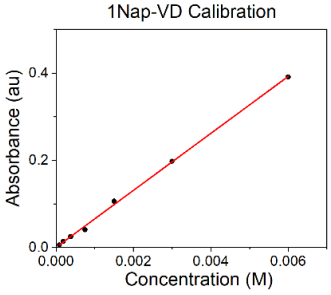
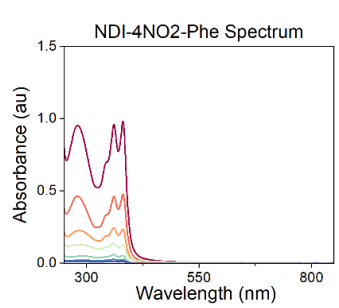
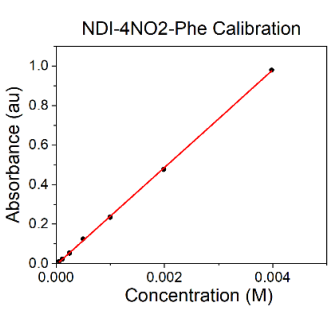
C_6-NDI-L (45)	
Min Conc. (mg mL^{-1})	0.078
Max Conc. (mg mL^{-1})	2.500
ε (M^{-1} cm^{-1})	20710
R^2 (au)	1.00





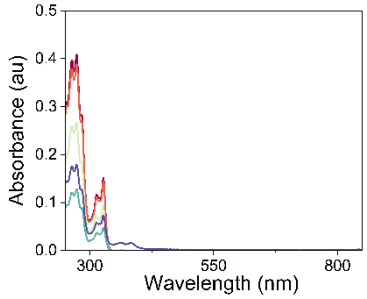
NDI-GA (46)	
Min Conc. (mg mL^{-1})	0.022
Max Conc. (mg mL^{-1})	2.850
ε (M^{-1} cm^{-1})	20959
R^2 (au)	1.00

		**NDI-GL (47)**	
		Min Conc. (mg mL^{-1})	0.023
		Max Conc. (mg mL^{-1})	2.950
		ε (M^{-1} cm^{-1})	20864
		R^2 (au)	1.00
		**NDI-LA (48)**	
		Min Conc. (mg mL^{-1})	0.020
		Max Conc. (mg mL^{-1})	2.500
		ε (M^{-1} cm^{-1})	19712
		R^2 (au)	1.00
		**NDI-YI (49)**	
		Min Conc. (mg mL^{-1})	0.021
		Max Conc. (mg mL^{-1})	2.680
		ε (M^{-1} cm^{-1})	20715
		R^2 (au)	1.00
		**PBI-A 1-Py (50)**	
		Min Conc. (mg mL^{-1})	0.020
		Max Conc. (mg mL^{-1})	2.600
		ε (M^{-1} cm^{-1})	19524
		R^2 (au)	1.00
		**PBI-I 1-Py (51)**	
		Min Conc. (mg mL^{-1})	0.019
		Max Conc. (mg mL^{-1})	2.450
		ε (M^{-1} cm^{-1})	23113
		R^2 (au)	1.00
		**PBI-L 1-Py (52)**	
		Min Conc. (mg mL^{-1})	0.038
		Max Conc. (mg mL^{-1})	2.450
		ε (M^{-1} cm^{-1})	24806
		R^2 (au)	1.00

		**PBI-V 1Py (53)**
		Min Conc. (mg mL^{-1}) 0.018
		Max Conc. (mg mL^{-1}) 2.300
		ε (M^{-1} cm^{-1}) 19589
		R^2 (au) 1.00
		**1Nap-VD (54)**
		Min Conc. (mg mL^{-1}) 0.040
		Max Conc. (mg mL^{-1}) 2.500
		ε (M^{-1} cm^{-1}) 6536
		R^2 (au) 1.00
		**NDI-4-Nitro-Phenylalanine (55)**
		Min Conc. (mg mL^{-1}) 0.041
		Max Conc. (mg mL^{-1}) 2.600
		ε (M^{-1} cm^{-1}) 24648
		R^2 (au) 1.00

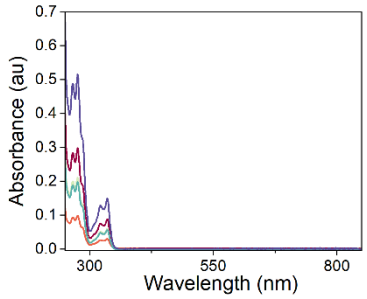
A 3.5 Solubility Data

Solubility calculation for each training, test, and validation compound. In each case spectra of the DMSO reconstituted sample is given for each solvent: acetone (▬), DMF (▬), DMSO (▬), ethanol (▬), and THF (▬). Using the calibration curve $\lambda_{max,}$ its associated absorbance, the dilution factor, and a known path length of 0.01 mm, the concentration of each sample can be determine using the Beer-Lambert relationship.



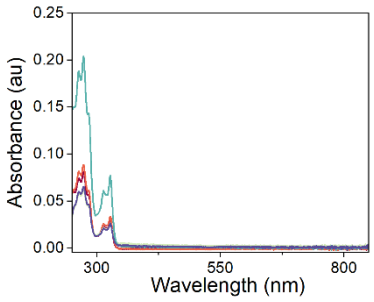
2Nap-FF (1)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	273.5	0.4089	2	0.01	7.4334
DMF		0.38898	1	0.001	35.35638
DMSO		0.26686	1	0.001	24.25627
Ethanol		0.12824	1	0.001	11.65639
THF		0.1792	1	0.001	16.2884



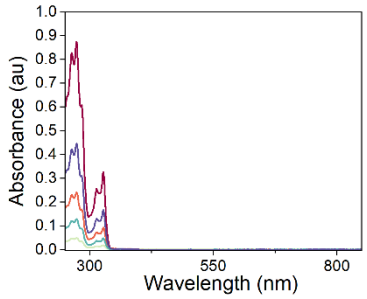
2Nap-V (2)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	276.0	0.29718	1	0.001	17.36518
DMF		0.09855	4	0.001	23.03437
DMSO		0.20996	4	0.001	49.07454
Ethanol		0.19817	1	0.001	11.57971
THF		0.51424	1	0.001	30.04869



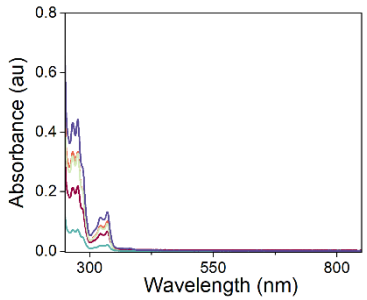
2Nap-VVV (3)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	273.5	0.08093	1	0.001	8.33164
DMF		0.08851	4	0.001	36.44795
DMSO		0.06275	4	0.001	25.84012
Ethanol		0.20368	1	0.001	20.96859
THF		0.06536	4	0.001	26.91491



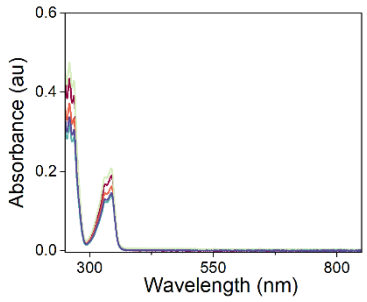
2Nap-IF (4)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	273.5	0.87422	2	0.01	16.15389
DMF		0.24068	1	0.001	22.23649
DMSO		0.04924	4	0.001	18.19719
Ethanol		0.12824	1	0.001	11.84813
THF		0.44533	1	0.001	41.14416



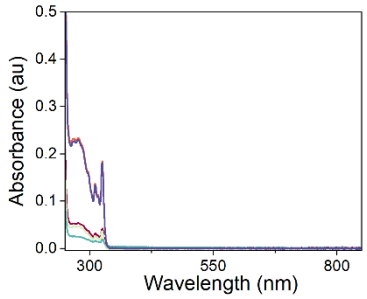
6Br2Nap-VF (5)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	275.5	0.21927	1	0.001	17.8993
DMF		0.33442	1	0.001	27.29915
DMSO		0.32799	1	0.001	26.77426
Ethanol		0.07298	1	0.001	5.95745
THF		0.4418	1	0.001	36.06472



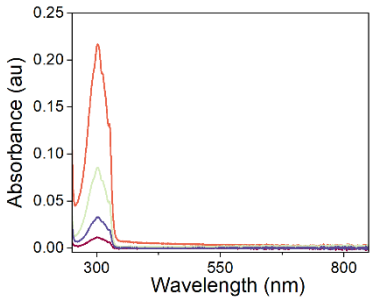
6MeO2Nap-FI (6)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	259.0	0.43373	1	0.001	32.66703
DMF		0.37156	1	0.001	27.9846
DMSO		0.47426	1	0.001	35.71961
Ethanol		0.31768	1	0.001	23.92655
THF		0.33669	1	0.001	25.35832



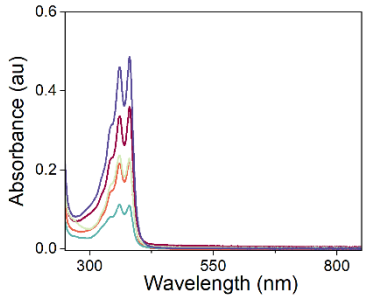
7MeO2Nap-FV (7)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	325.5	0.04202	1	0.001	6.54779
DMF		0.18443	1	0.001	28.73889
DMSO		0.03611	4	0.001	22.50743
Ethanol		0.01844	1	0.001	2.87342
THF		0.18033	1	0.001	28.10001



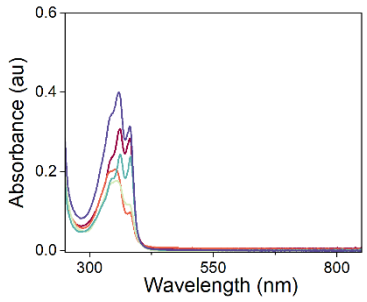
4ClNap-LIF (8)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	302.0	0.01119	1	0.001	0.88965
DMF		0.21661	1	0.001	17.22133
DMSO		0.08527	4	0.001	27.11718
Ethanol		--	--	--	--
THF		0.03271	1	0.001	2.60057



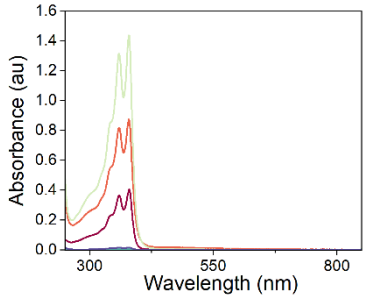
NDI-F (9)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	380.5	0.35954	4	0.01	4.05904
DMF		0.22339	40	0.01	25.21972
DMSO		0.22847	40	0.01	25.79323
Ethanol		0.10986	10	0.01	3.10067
THF		0.4858	20	0.01	27.42231



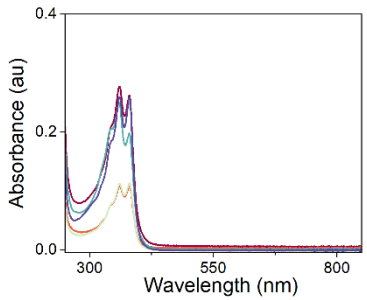
NDI-FF (10)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	358.5	0.3002	10	0.01	10.95284
DMF		0.19048	40	0.01	27.79876
DMSO		0.17268	40	0.01	25.20102
Ethanol		0.22891	10	0.01	8.35181
THF		0.39918	10	0.01	14.56414



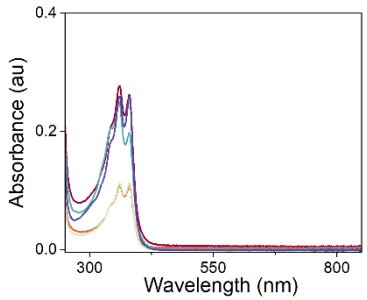
NDI-G (11)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	379.5	0.40382	1	0.01	8.06038
DMF		0.87163	1	0.001	17.39801
DMSO		1.43124	1	0.01	28.568
Ethanol		0.00382	1	0.001	0.07625
THF		0.01597	1	0.001	0.31877



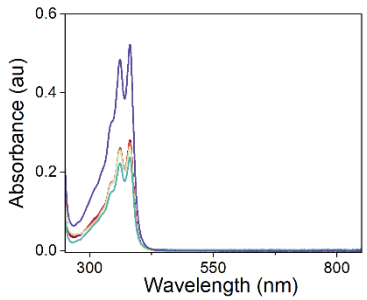
NDI-GF (12)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	379.5	0.11218	0.01	5	2.27112
DMF		0.11426	0.01	40	18.50587
DMSO		0.31172	0.01	40	50.48704
Ethanol		0.09815	0.001	1	3.97416
THF		0.29539	0.001	1	11.96055



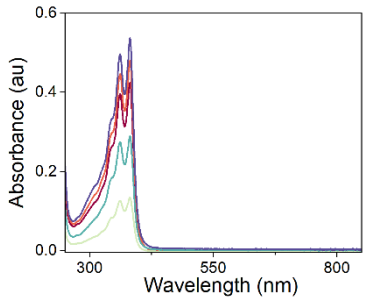
NDI-A (13)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	360.0	0.27605	10	0.01	6.4302
DMF		0.10703	80	0.01	19.94491
DMSO		0.11179	80	0.01	20.83193
Ethanol		0.24874	10	0.01	5.79405
THF		0.25932	40	0.01	24.16198



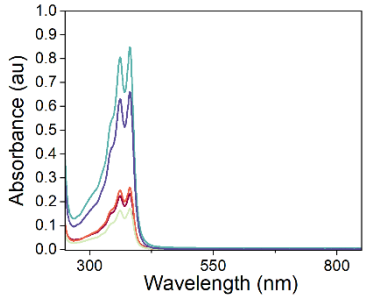
NDI-I (14)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone		0.27921	40	0.01	26.8722
DMF		0.27352	40	0.01	26.32457
DMSO		0.26272	40	0.01	25.28514
Ethanol		0.23613	40	0.01	22.72602
THF		0.52167	20	0.01	25.10372



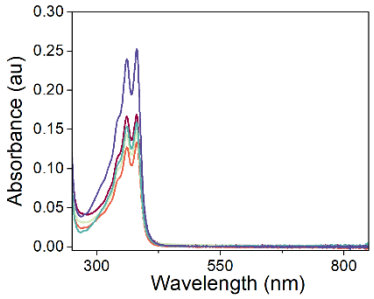
NDI-L (15)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	381.0	0.42369	20	0.01	20.71327
DMF		0.47957	20	0.01	23.44512
DMSO		0.13437	80	0.01	26.27621
Ethanol		0.28852	40	0.01	28.21021
THF		0.53643	20	0.01	26.22488



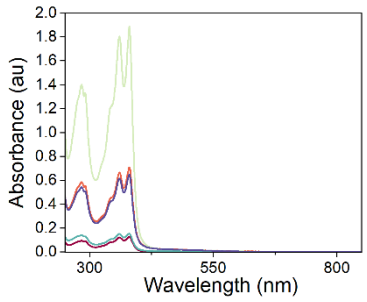
NDI-M (16)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	381.0	0.23232	0.001	4	25.07874
DMF		0.25811	0.001	4	27.86275
DMSO		0.17051	0.001	4	18.40641
Ethanol		0.84867	0.001	1	22.9033
THF		0.65951	0.001	4	71.19353



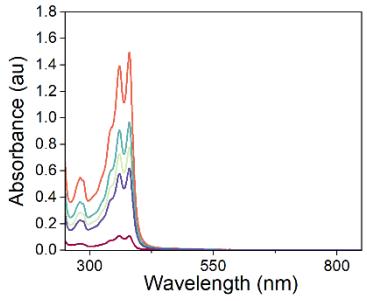
NDI-V (17)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone		0.16797	40	0.01	15.4315
DMF		0.13299	80	0.01	24.43574
DMSO		0.14583	80	0.01	26.79498
Ethanol		0.15862	40	0.01	14.57251
THF		0.25217	40	0.01	23.16701



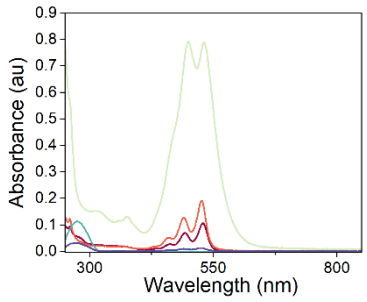
NDI-W (18)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	381.0	0.12717	50	0.01	19.62951
DMF		0.70819	1	0.001	21.86273
DMSO		1.88711	4	0.01	23.303
Ethanol		0.15445	1	0.001	4.76807
THF		0.64836	1	0.001	20.01571



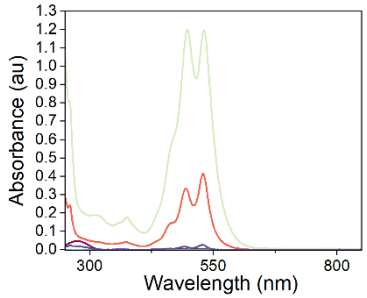
NDI-Y (19)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	380.5	0.10804	50	0.01	15.34473
DMF		1.48566	1	0.001	42.20112
DMSO		0.7766	1	0.001	22.05982
Ethanol		0.96227	1	0.001	27.33389
THF		0.61741	1	0.001	17.53792



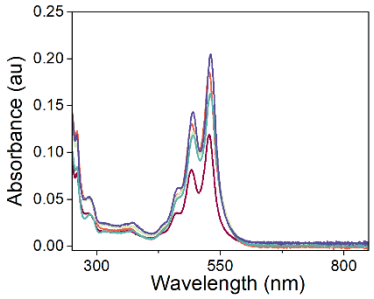
PBI-A (20)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	380.5	0.10633	1	0.01	0.13651
DMF		0.17882	10	0.01	2.29578
DMSO		0.78319	1	0.001	10.05499
Ethanol		-0.00126	1	0.001	0.00
THF		0.012	1	0.001	0.15406



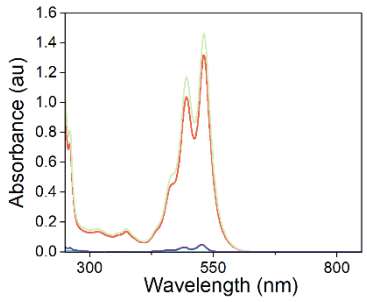
PBI-D (21)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	529.5	0.00306	1	0.001	0.03435
DMF		0.41261	1	0.001	4.63176
DMSO		1.18889	1	0.001	13.34591
Ethanol		0.00101	1	0.001	0.01134
THF		0.02604	1	0.001	0.29231



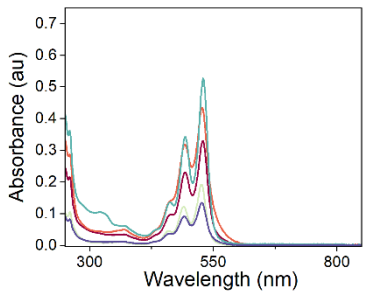
PBI-DOPA (22)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	531.5	0.10851	80	0.01	16.49726
DMF		0.17119	80	0.01	26.02677
DMSO		0.16052	80	0.01	24.40457
Ethanol		0.16079	80	0.01	24.44562
THF		0.20307	80	0.01	30.87363



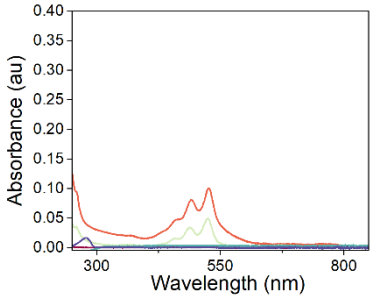
PBI-E (23)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	529.0	0.0024	1	0.001	0.03318
DMF		1.29903	1	0.001	17.95949
DMSO		1.43923	1	0.001	19.8978
Ethanol		0.00158	1	0.001	0.02184
THF		0.04742	1	0.001	0.6556



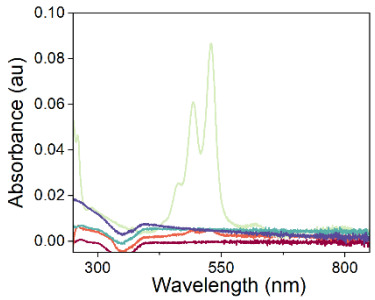
PBI-F (24)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	529.5	0.3283	4	0.01	1.81907
DMF		0.42565	20	0.01	11.79236
DMSO		0.17662	80	0.01	19.57257
Ethanol		0.52657	1	0.01	0.72941
THF		0.12866	40	0.01	7.12888



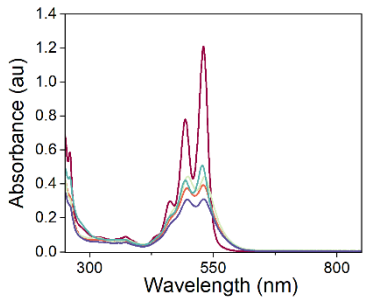
PBI-G (25)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	524.5	0.001	1	0.001	0.06168
DMF		0.09863	1	0.01	0.60832
DMSO		0.04939	4	0.01	1.21849
Ethanol		0.00362	1	0.01	0.02233
THF		0.0011	1	0.01	0.00678



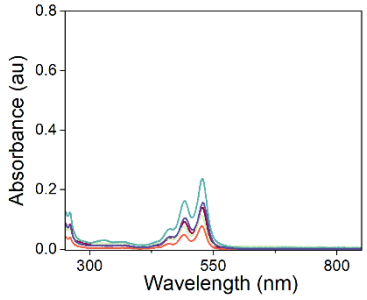
PBI-H (26)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	524.5	0.10633	1	0.001	3.20234
DMF		0.17882	1	0.001	5.38552
DMSO		0.78319	5	0.01	11.79367
Ethanol		-0.00126	1	0.001	0.00
THF		0.012	1	0.001	0.3614



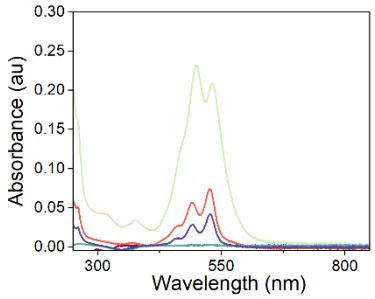
PBI-I (27)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	531.0	1.19857	1	0.01	1.11625
DMF		0.39181	1	0.001	3.64901
DMSO		0.43931	1	0.001	4.09139
Ethanol		0.48875	1	0.01	0.45518
THF		0.30992	1	0.001	2.88635



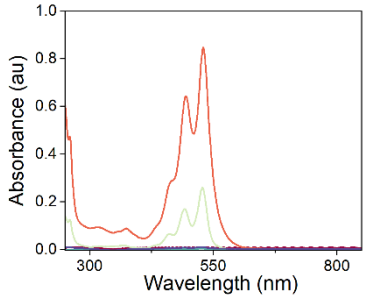
PBI-L (28)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	530.0	0.13594	20	0.01	6.99156
DMF		0.07203	80	0.01	14.81835
DMSO		0.10301	80	0.01	21.19171
Ethanol		0.22969	10	0.01	5.90661
THF		0.15465	80	0.01	31.81533



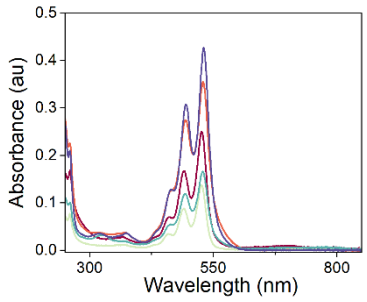
PBI-M (29)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	531.0	0.00258	1	0.001	0.02821
DMF		0.06778	1	0.001	0.74102
DMSO		0.20913	1	0.001	2.28637
Ethanol		0.00293	1	0.001	0.03203
THF		0.03869	1	0.01	0.0423



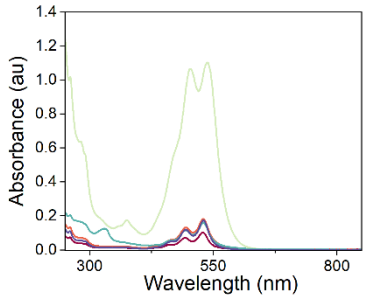
PBI-S (30)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	529.5	0.01017	1	0.001	0.13148
DMF		0.84519	1	0.001	10.92657
DMSO		0.25515	80	0.01	26.38852
Ethanol		0.00328	1	0.01	0.00424
THF		0.00975	1	0.001	0.12605



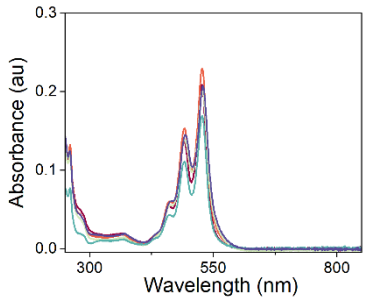
PBI-V (31)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	530.5	0.22859	1	0.01	0.28294
DMF		0.35244	20	0.01	8.7248
DMSO		0.11977	80	0.01	11.85983
Ethanol		0.1625	1	0.001	2.01138
THF		0.4263	1	0.001	5.27662



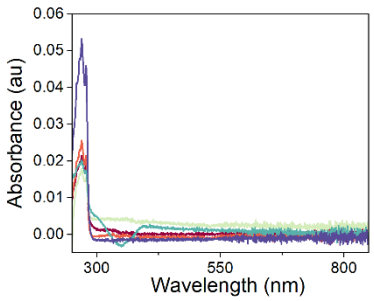
PBI-W (32)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	531.0	0.09988	80	0.01	14.95804
DMF		0.18056	80	0.01	27.04069
DMSO		1.04907	1	0.001	19.6386
Ethanol		0.15921	1	0.01	0.29804
THF		0.17264	80	0.01	25.85459



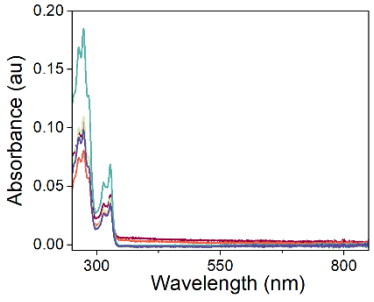
PBI-Y (33)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	530.5	0.19109	10	0.01	2.63491
DMF		0.21629	80	0.01	23.85911
DMSO		0.19356	80	0.01	21.35175
Ethanol		0.15814	80	0.01	17.44454
THF		0.20435	80	0.01	22.542



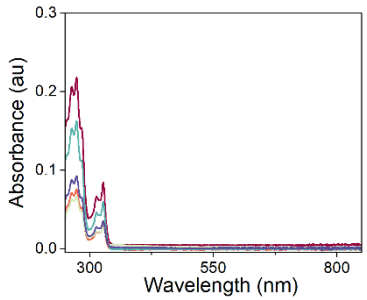
ThNap-FF (34)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	269.5	0.02136	1	0.001	9.30619
DMF		0.0255	4	0.001	44.43969
DMSO		0.01781	4	0.001	31.03807
Ethanol		0.02013	1	0.001	8.7703
THF		0.05312	1	0.001	23.14349



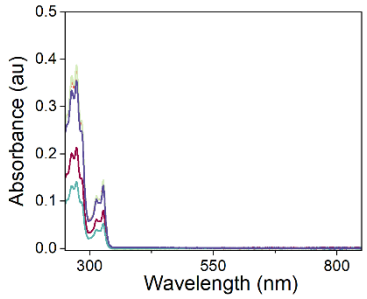
2Nap-VL (35)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	273.0	0.10434	1	0.001	9.20582
DMF		0.08019	4	0.001	28.30034
DMSO		0.10946	4	0.001	38.6302
Ethanol		0.18429	1	0.001	16.25973
THF		0.09733	4	0.001	34.34933



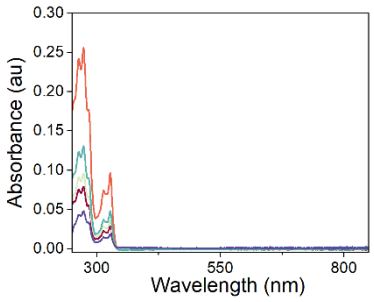
2Nap-FL (36)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	273.5	0.21779	1	0.001	21.72967
DMF		0.07556	4	0.001	30.15555
DMSO		0.06666	4	0.001	26.60361
Ethanol		0.16235	1	0.001	16.19823
THF		0.09253	4	0.001	36.92817



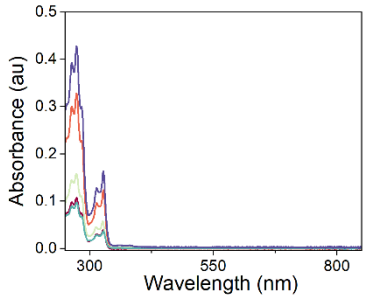
2Nap-FI (37)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	273.5	0.2131	1	0.001	20.19865
DMF		0.3806	1	0.001	36.07511
DMSO		0.38721	1	0.001	36.70164
Ethanol		0.14058	1	0.001	13.32485
THF		0.35427	1	0.001	33.57942



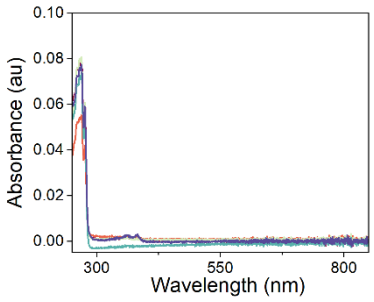
2Nap-VF (38)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	273.5	0.07893	1	0.001	7.40466
DMF		0.25567	1	0.001	23.98517
DMSO		0.09589	4	0.001	35.98292
Ethanol		0.13061	1	0.001	12.25292
THF		0.04772	4	0.001	17.90703



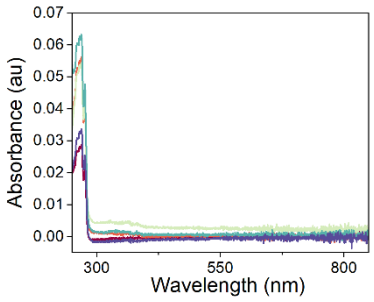
2Nap-IV (39)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	273.0	0.1066	1	0.01	0.95557
DMF		0.32592	1	0.001	29.2158
DMSO		0.15794	1	0.001	14.1579
Ethanol		0.09659	1	0.01	0.86584
THF		0.42525	1	0.001	38.11984

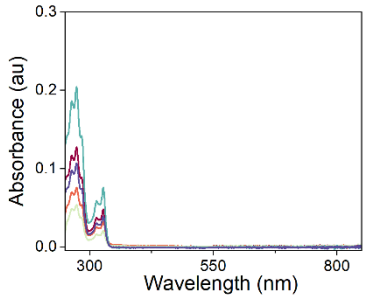


CF_3Ph-FL (40)

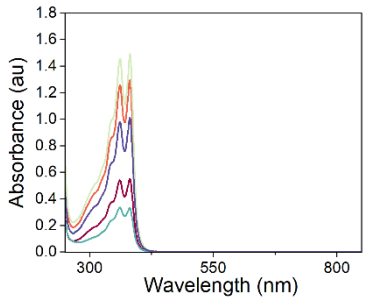
Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	269.0	0.07979	1	0.001	34.63189
DMF		0.05514	1	0.001	23.93285
DMSO		0.08057	1	0.001	34.97044
Ethanol		0.07357	1	0.001	31.93217
THF		0.07744	1	0.001	33.6119



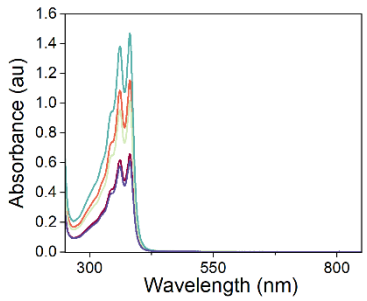
CF_3Ph-FV (41)					
Solvent	**λ_{max} (nm)**	**λ_{max} Absorbance (au)**	**Dilution Factor (au)**	**Path Length (cm)**	**Concentration (mg mL^{-1})**
Acetone	268.5	0.02859	1	0.001	13.81978
DMF		0.05605	1	0.001	27.09335
DMSO		0.05435	1	0.001	26.27161
Ethanol		0.06319	1	0.001	30.54467
THF		0.03359	1	0.001	16.23668



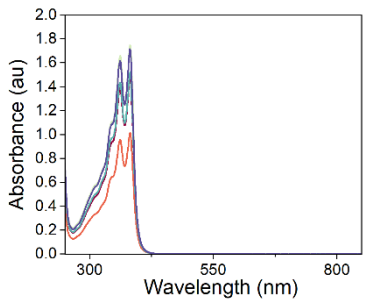
2Nap-VI (42)					
Solvent	**λ_{max} (nm)**	**λ_{max} Absorbance (au)**	**Dilution Factor (au)**	**Path Length (cm)**	**Concentration (mg mL^{-1})**
Acetone	273.5	0.12725	1	0.001	10.97942
DMF		0.0757	4	0.001	26.12627
DMSO		0.05366	4	0.001	18.51963
Ethanol		0.20396	1	0.001	17.59813
THF		0.10679	4	0.001	36.85633



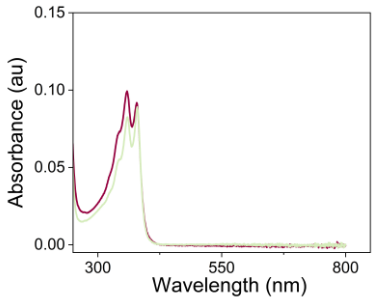
C_6-NDI-A (43)					
Solvent	**λ_{max} (nm)**	**λ_{max} Absorbance (au)**	**Dilution Factor (au)**	**Path Length (cm)**	**Concentration (mg mL^{-1})**
Acetone	381.5	0.54923	1	0.001	11.98619
DMF		1.29405	1	0.001	28.24087
DMSO		1.49149	1	0.001	32.54973
Ethanol		0.33045	1	0.001	7.21162
THF		1.00809	1	0.001	22.00018



C_6-NDI-F (44)					
Solvent	**λ_{max} (nm)**	**λ_{max} Absorbance (au)**	**Dilution Factor (au)**	**Path Length (cm)**	**Concentration (mg mL^{-1})**
Acetone	381.0	0.65783	1	0.001	16.01184
DMF		1.15244	1	0.001	28.05085
DMSO		1.01673	1	0.001	24.74761
Ethanol		1.46768	1	0.001	35.72391
THF		0.60874	1	0.001	14.81697

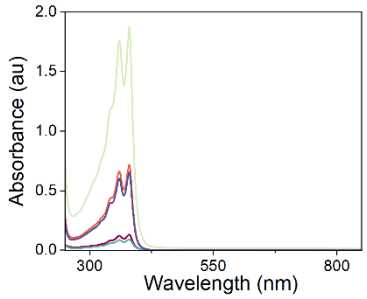


C_6-NDI-L (45)					
Solvent	**λ_{max} (nm)**	**λ_{max} Absorbance (au)**	**Dilution Factor (au)**	**Path Length (cm)**	**Concentration (mg mL^{-1})**
Acetone	381.5	1.47885	1	0.001	33.17023
DMF		1.01199	1	0.001	22.69868
DMSO		1.74551	1	0.001	39.15134
Ethanol		1.52155	1	0.001	34.12798
THF		1.71097	1	0.001	38.37662



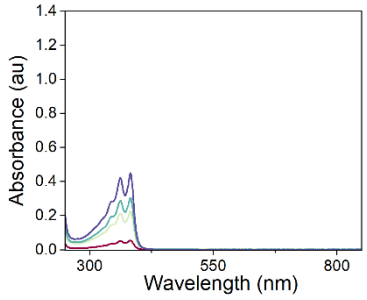
NDI-GA (46)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	379.5	0.09086	2	0.01	0.4547
DMF		-	-	-	-
DMSO		0.08906	40	0.01	8.9139
Ethanol		-	-	-	-
THF		-	-	-	-



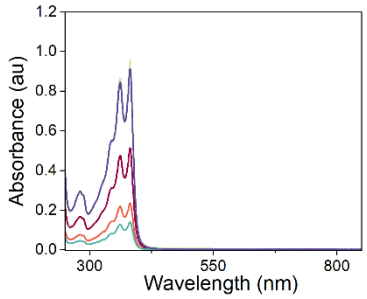
NDI-GL (47)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	379.5	0.13151	5	0.01	1.91806
DMF		0.71605	1	0.001	20.88708
DMSO		1.86981	1	0.001	54.5421
Ethanol		0.09127	1	0.001	2.66233
THF		0.6518	1	0.001	19.01292



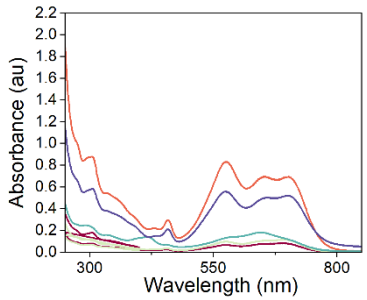
NDI-LA (48)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	382.5	0.05477	200	0.01	35.37933
DMF		0.22527	4	0.001	29.10317
DMSO		0.22527	4	0.001	29.10317
Ethanol		0.30539	4	0.001	39.45406
THF		0.44931	4	0.001	58.04742



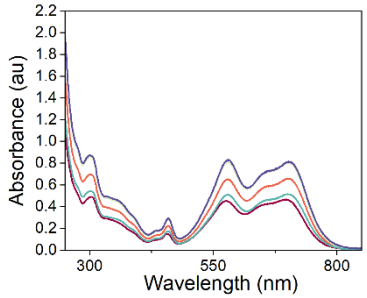
NDI-YI (49)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	381.5	0.51179	2	0.01	4.05603
DMF		0.23526	4	0.001	37.28953
DMSO		0.9567	1	0.001	37.91007
Ethanol		0.13995	1	0.001	5.54564
THF		0.91158	1	0.001	36.12216



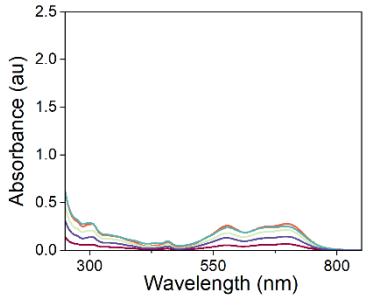
PBI-A 1-Py (50)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	697.5	0.08082	8	0.01	2.22769
DMF		0.69011	1	0.001	23.77741
DMSO		0.11427	40	0.01	15.74847
Ethanol		0.12081	1	0.001	4.16245
THF		0.51869	1	0.001	17.87121



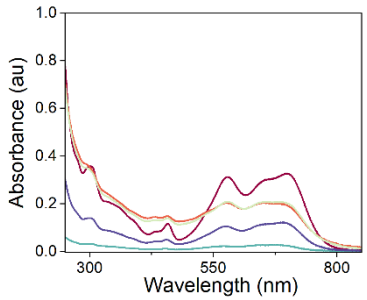
PBI-I 1-Py (51)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	700.0	0.4605	1	0.001	15.07957
DMF		0.65633	1	0.001	21.49223
DMSO		0.79951	1	0.001	26.18081
Ethanol		0.51332	1	0.001	16.80921
THF		0.80684	1	0.001	26.42084



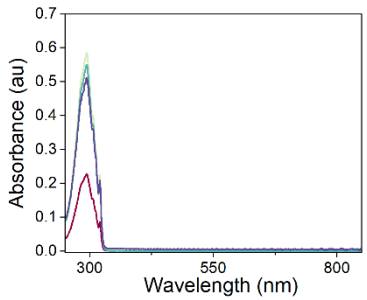
PBI-L 1-Py (52)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	695.5	0.06549	100	0.01	19.98176
DMF		0.27796	4	0.001	33.92354
DMSO		0.2166	40	0.01	26.43487
Ethanol		0.25273	1	0.001	7.71109
THF		0.14479	4	0.001	17.67085



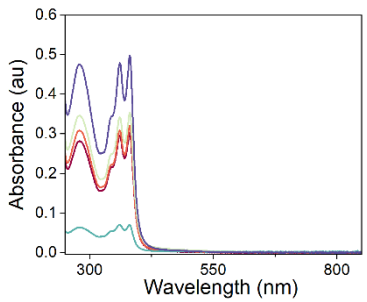
PBI-V 1Py (53)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	697.5	0.35716	1	0.001	12.10153
DMF		0.35237	1	0.001	7.24038
DMSO		0.34067	1	0.001	7.61614
Ethanol		0.03168	1	0.001	0.91598
THF		0.13972	1	0.001	4.44408



1Nap-VD (54)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	294.0	0.22655	1	0.001	14.43424
DMF		0.54808	1	0.001	34.91997
DMSO		0.58404	1	0.001	37.2111
Ethanol		0.54934	1	0.001	35.00025
THF		0.51001	1	0.001	32.49441



NDI-4-Nitro-Phenylalanine (55)

Solvent	λ_{max} (nm)	λ_{max} Absorbance (au)	Dilution Factor (au)	Path Length (cm)	Concentration (mg mL^{-1})
Acetone	381.0	0.30184	1	0.001	7.9909
DMF		0.32096	4	0.001	33.98832
DMSO		0.35334	4	0.001	37.41723
Ethanol		0.06964	1	0.001	1.84365
THF		0.49769	1	0.001	13.17582

A 3.6 Classification Model Data Summary

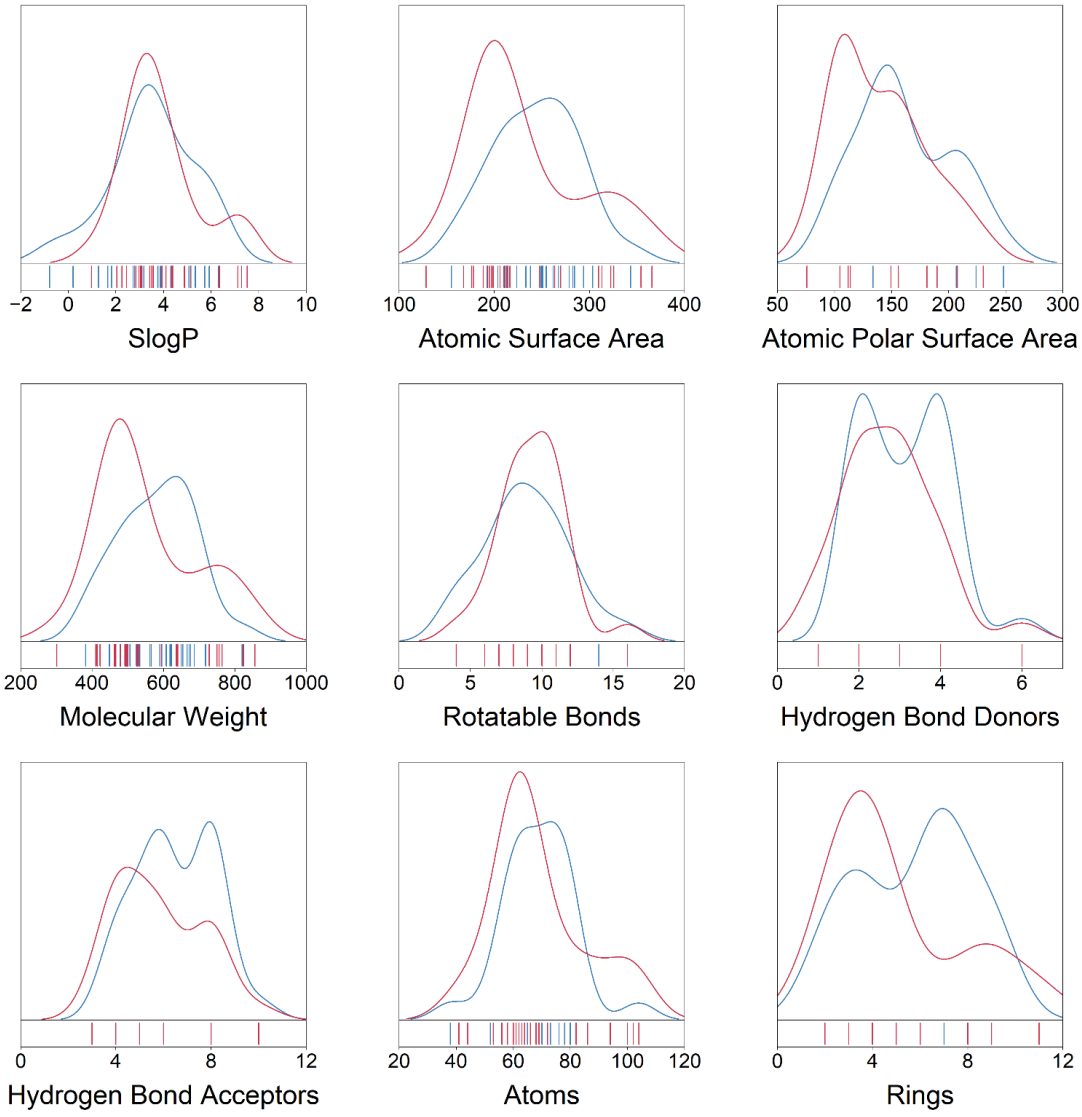
Solubility classification of training, test, and validation compounds used for model development. Compounds with a solubility greater than or equal to 10 mg mL^{-1} are deemed to be soluble (S), while those with a solubility less than 10 mg mL^{-1} are considered insoluble (I).

Compound	Acetone	DMF	DMSO	Ethanol	THF	Compound	Acetone	DMF	DMSO	Ethanol	THF	Compound	Acetone	DMF	DMSO	Ethanol	THF
1	I	S	S	S	S	**19**	S	S	S	S	S	**37**	S	S	S	S	S
2	S	S	S	S	S	**20**	I	I	S	I	I	**38**	I	S	S	S	S
3	I	S	S	S	S	**21**	I	I	S	I	I	**39**	I	S	S	I	S
4	S	S	S	S	S	**22**	S	S	S	S	S	**40**	S	S	S	S	S
5	S	S	S	I	S	**23**	I	S	S	I	I	**41**	S	S	S	S	S
6	S	S	S	S	S	**24**	I	S	S	I	I	**42**	S	S	S	S	S
7	I	S	S	I	S	**25**	I	I	I	I	I	**43**	S	S	S	I	S
8	I	S	S	[a]	I	**26**	I	I	S	I	I	**44**	S	S	S	S	S
9	I	S	S	I	S	**27**	I	I	I	I	I	**45**	S	S	S	S	S
10	S	S	S	I	S	**28**	I	S	S	I	S	**46**	I	[a]	I	[a]	[a]
11	I	S	S	I	I	**29**	I	I	I	I	I	**47**	I	S	S	I	S
12	I	S	S	I	S	**30**	I	S	S	I	I	**48**	S	S	S	S	S
13	I	S	S	I	S	**31**	I	I	S	I	I	**49**	I	S	S	I	S
14	S	S	S	S	S	**32**	S	S	S	I	S	**50**	I	S	S	I	S
15	S	S	S	S	S	**33**	I	S	S	S	S	**51**	S	S	S	S	S
16	S	S	S	S	S	**34**	I	S	S	I	S	**52**	S	S	S	I	S
17	S	S	S	S	S	**35**	I	S	S	S	S	**53**	S	I	I	I	I
18	S	S	S	I	S	**36**	S	S	S	S	S	**54**	S	S	S	S	S
												55	I	S	S	I	S

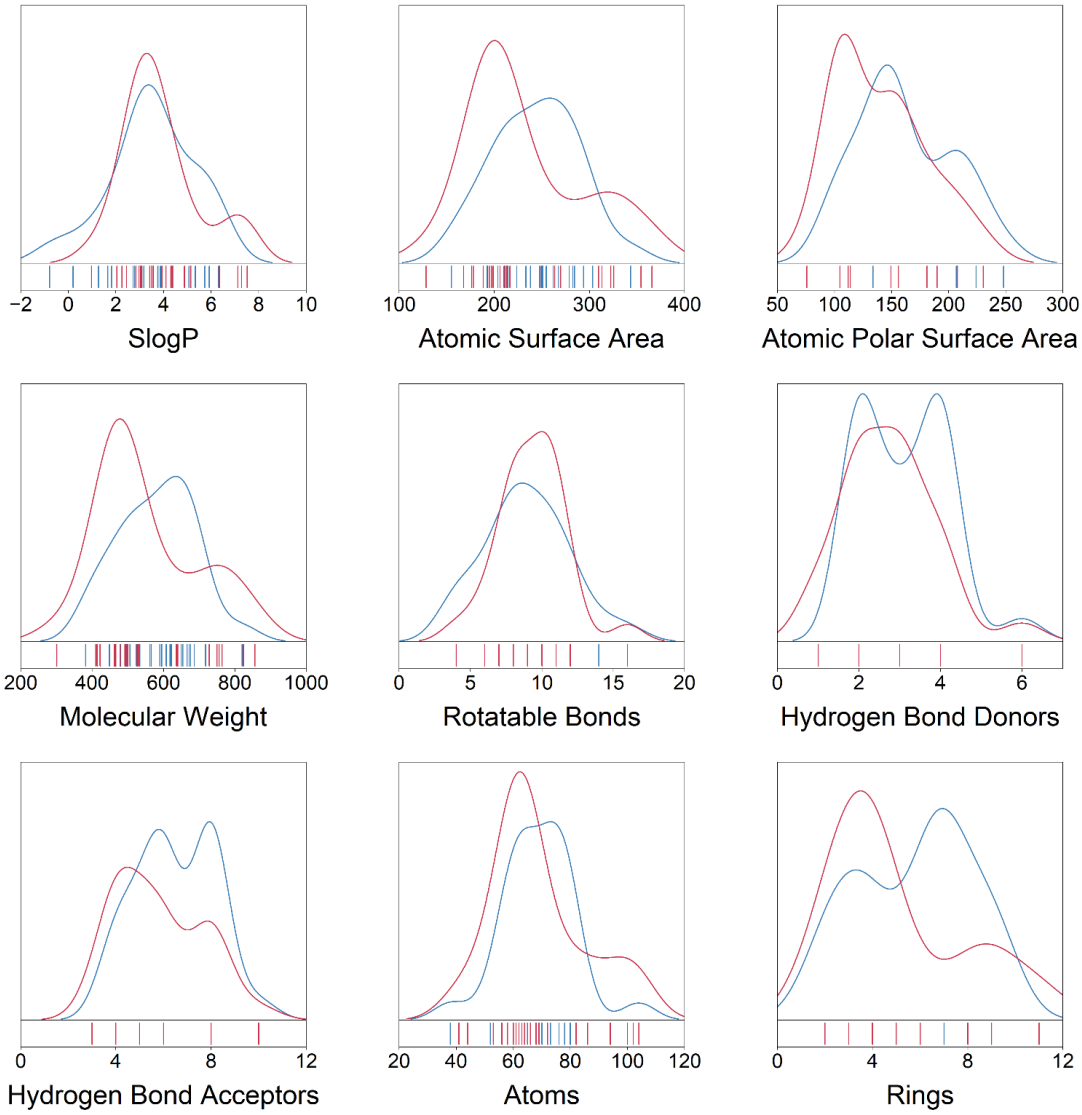
a Insufficient material for solubility measurement. Compound is not included in the model training or testing for this solvent.

A 3.7 Physicochemical Descriptor Distribution

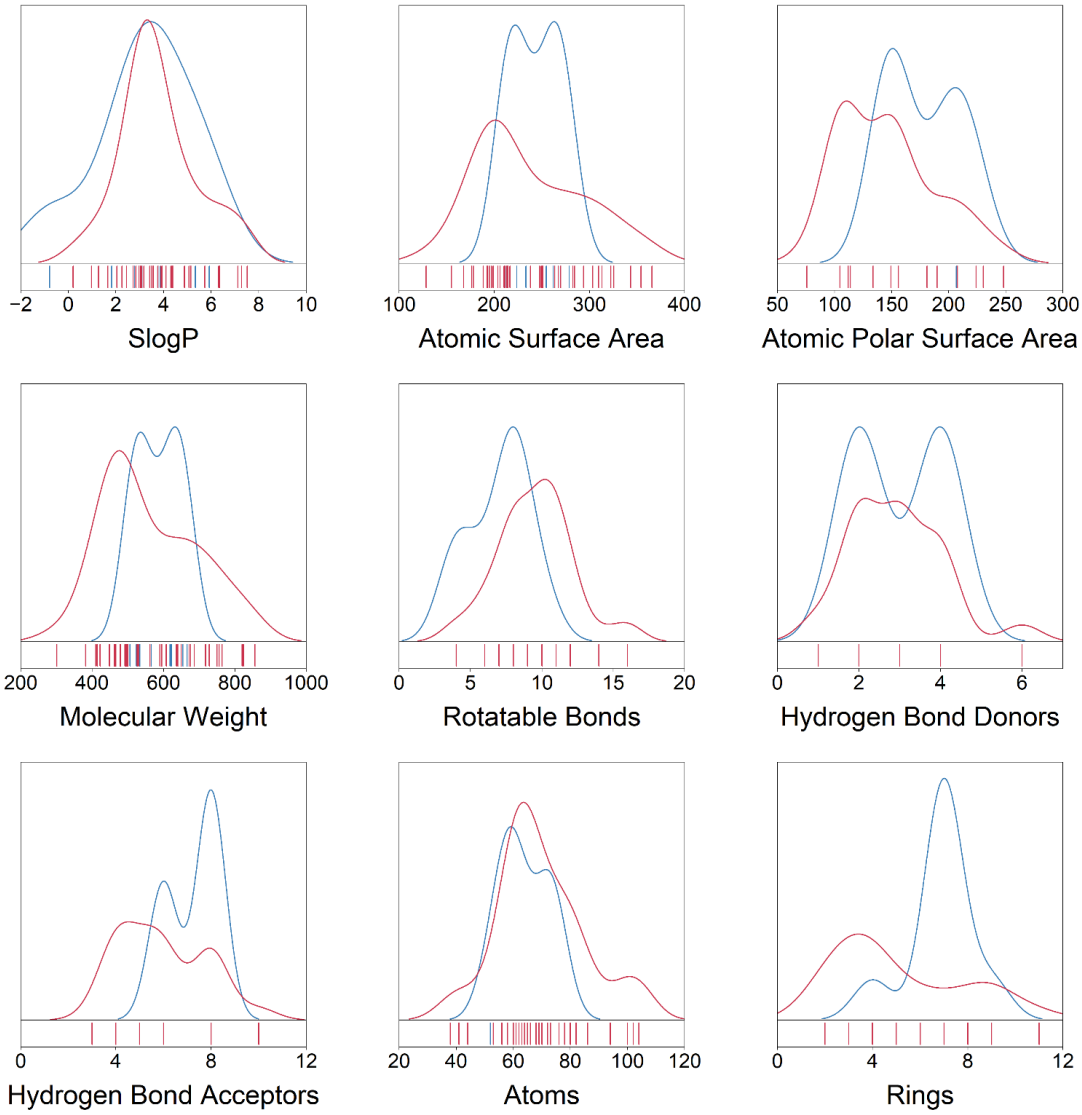
Plot of 9 physicochemical descriptors calculated and their distribution, separated by acetone solubility classification (soluble, red; insoluble, blue) for the training set. The value of each molecule's descriptor is depicted by | marker.



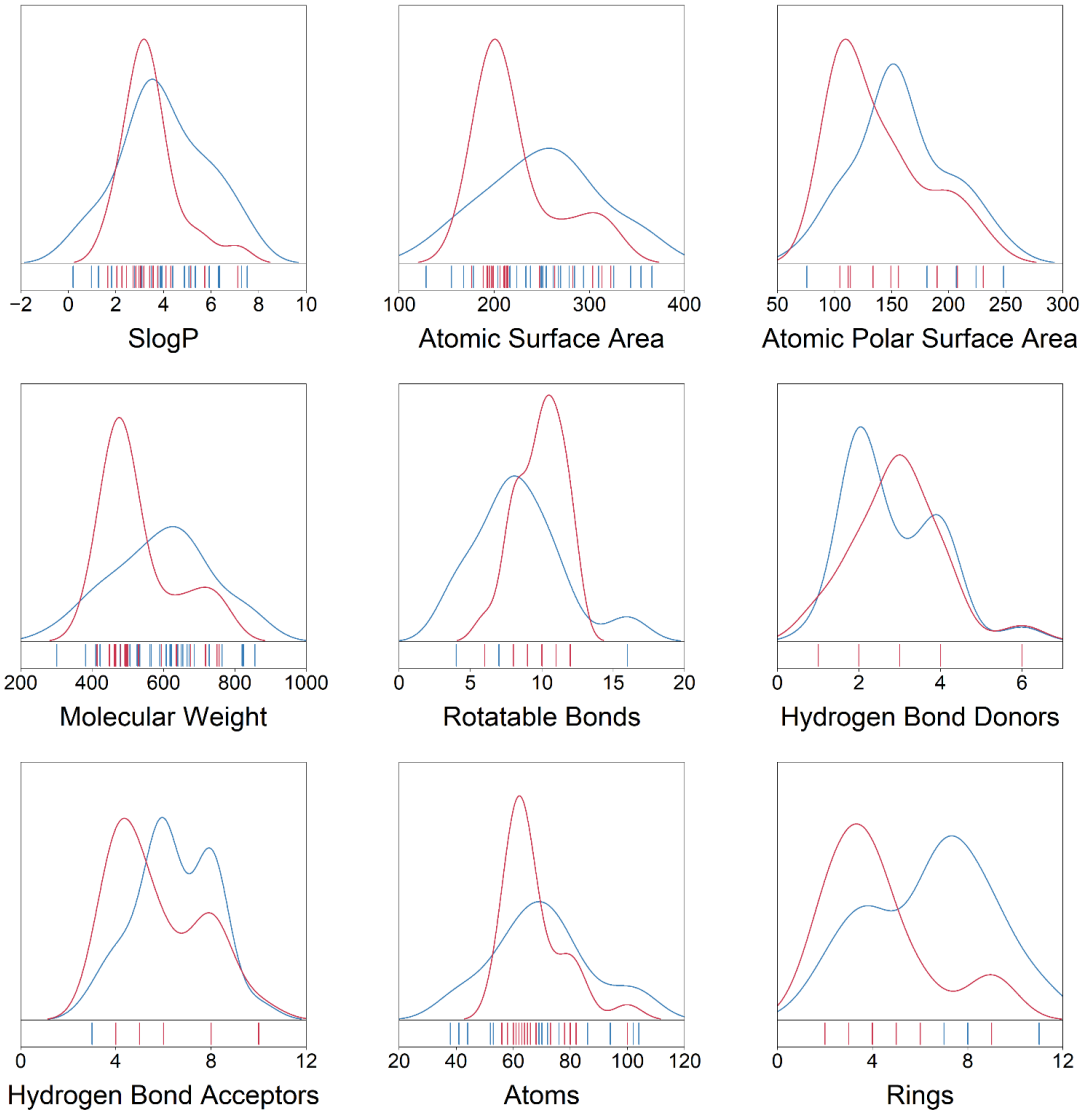
Plot of 9 physicochemical descriptors calculated and their distribution, separated by DMF solubility classification (soluble, red; insoluble, blue) for the training set. The value of each molecule's descriptor is depicted by | marker.



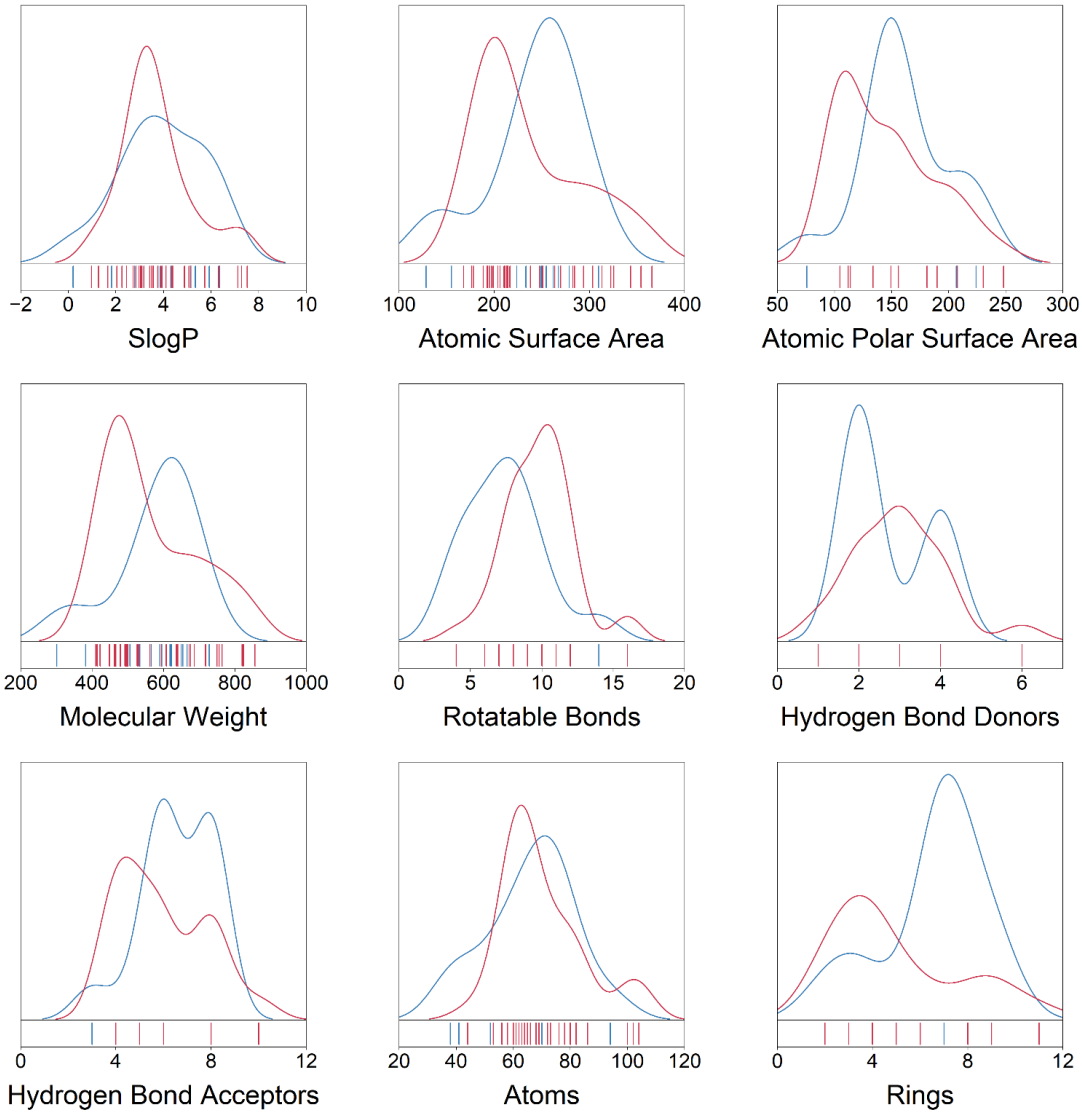
Plot of 9 physicochemical descriptors calculated and their distribution, separated by DMSO solubility classification (soluble, red; insoluble, blue) for the training set. The value of each molecule's descriptor is depicted by | marker.



Plot of 9 physicochemical descriptors calculated and their distribution, separated by ethanol solubility classification (soluble, red; insoluble, blue) for the training set. The value of each molecule's descriptor is depicted by | marker.

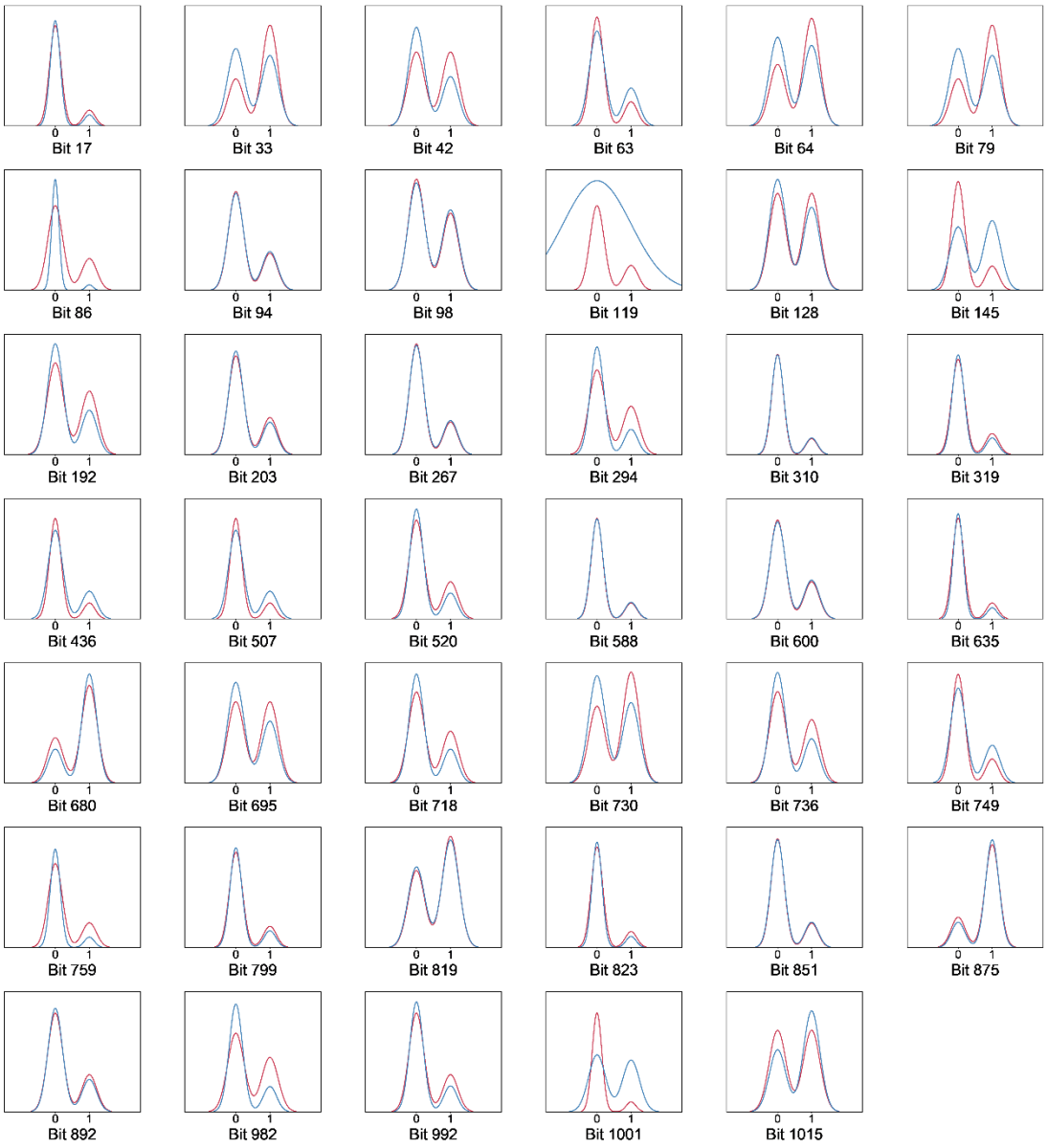


Plot of 9 physicochemical descriptors calculated and their distribution, separated by THF solubility classification (soluble, red; insoluble, blue) for the training set. The value of each molecule's descriptor is depicted by | marker.

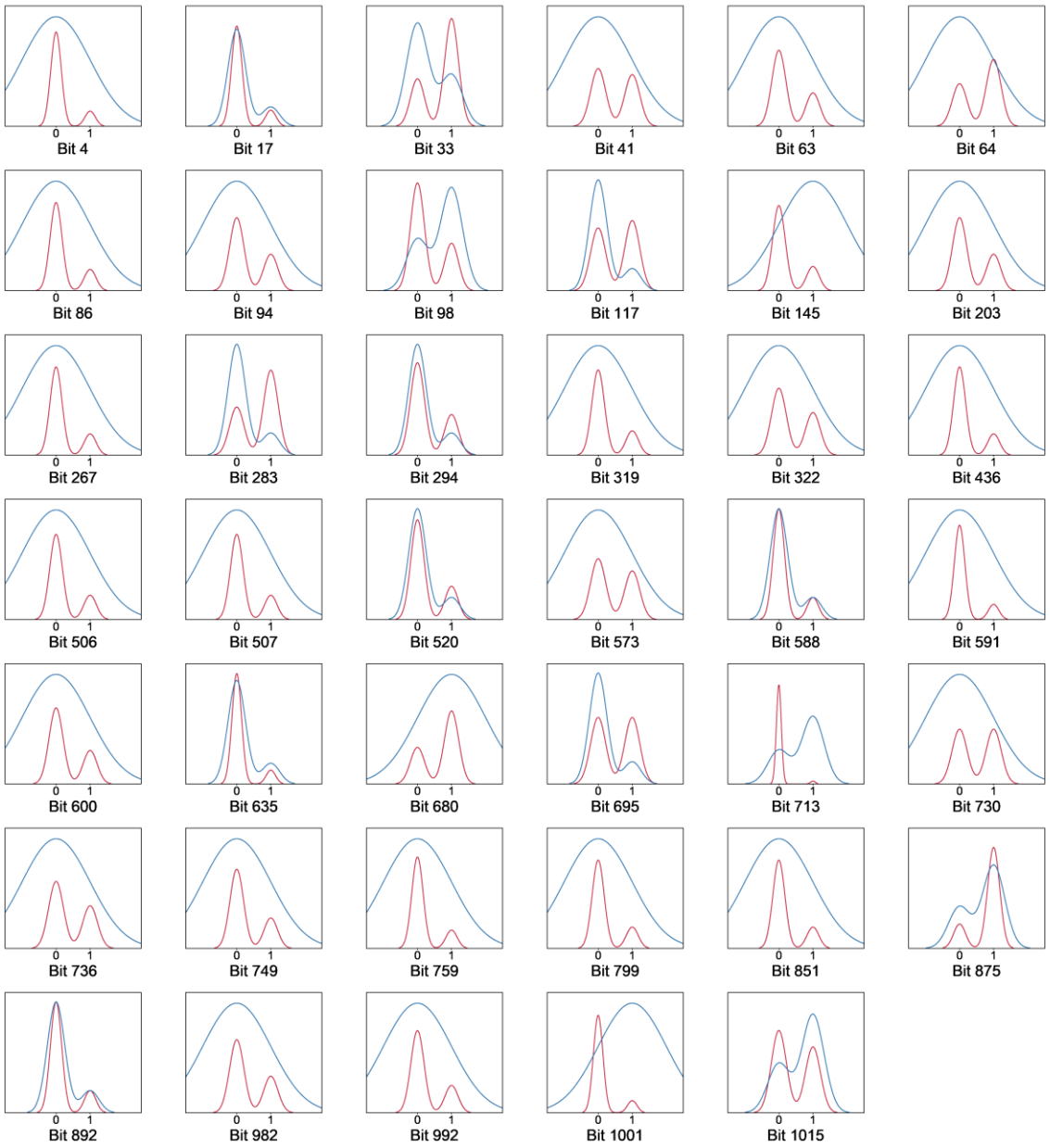


A 3.8 Fingerprint Descriptor Distribution

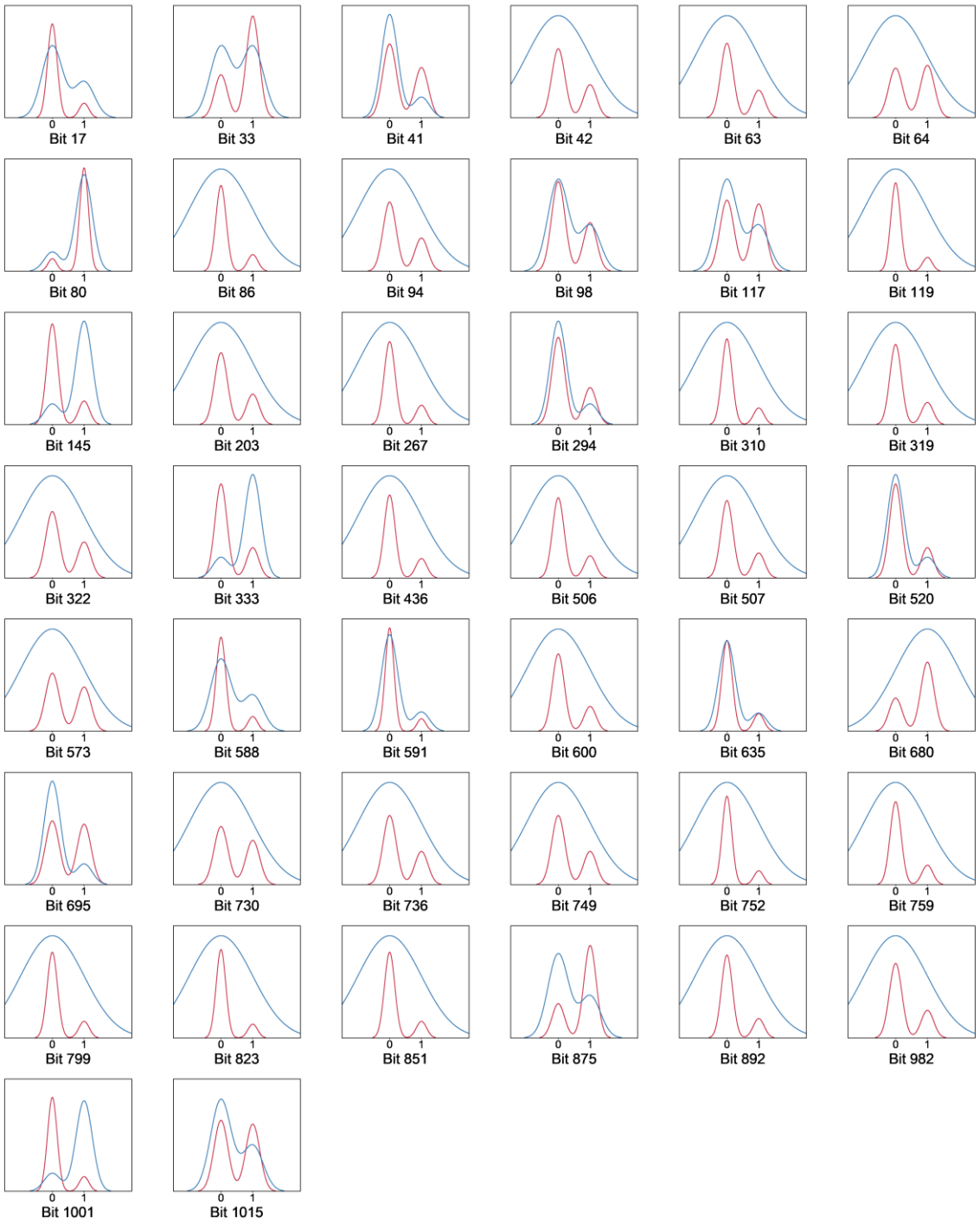
Plot of the 41 Morgan fingerprint descriptors and their distributions, separated by acetone solubility classification (soluble, red; insoluble, blue) for the training set.



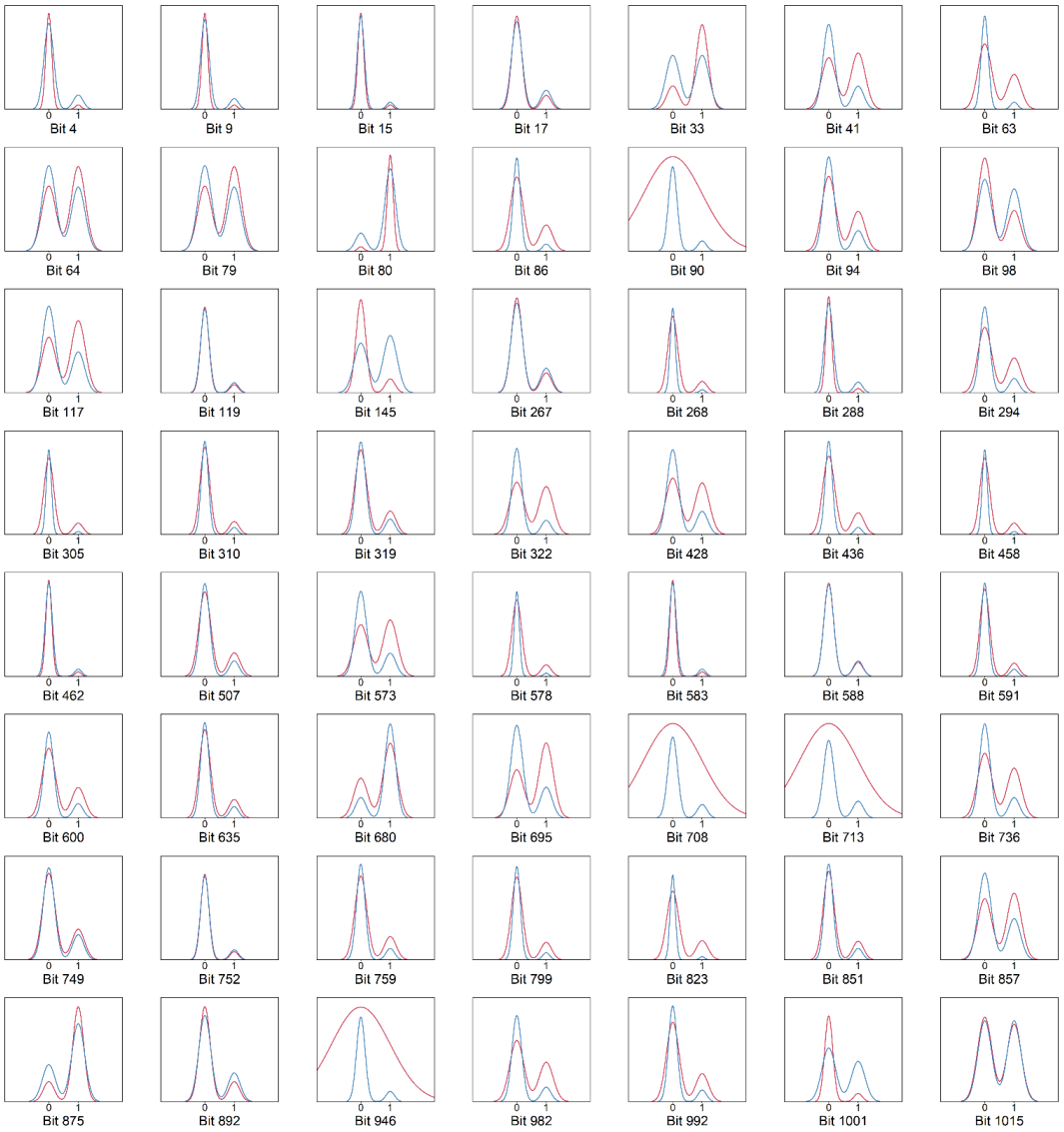
Plot of the 41 Morgan fingerprint descriptors and their distributions, separated by DMF solubility classification (soluble, red; insoluble, blue) for the training set.



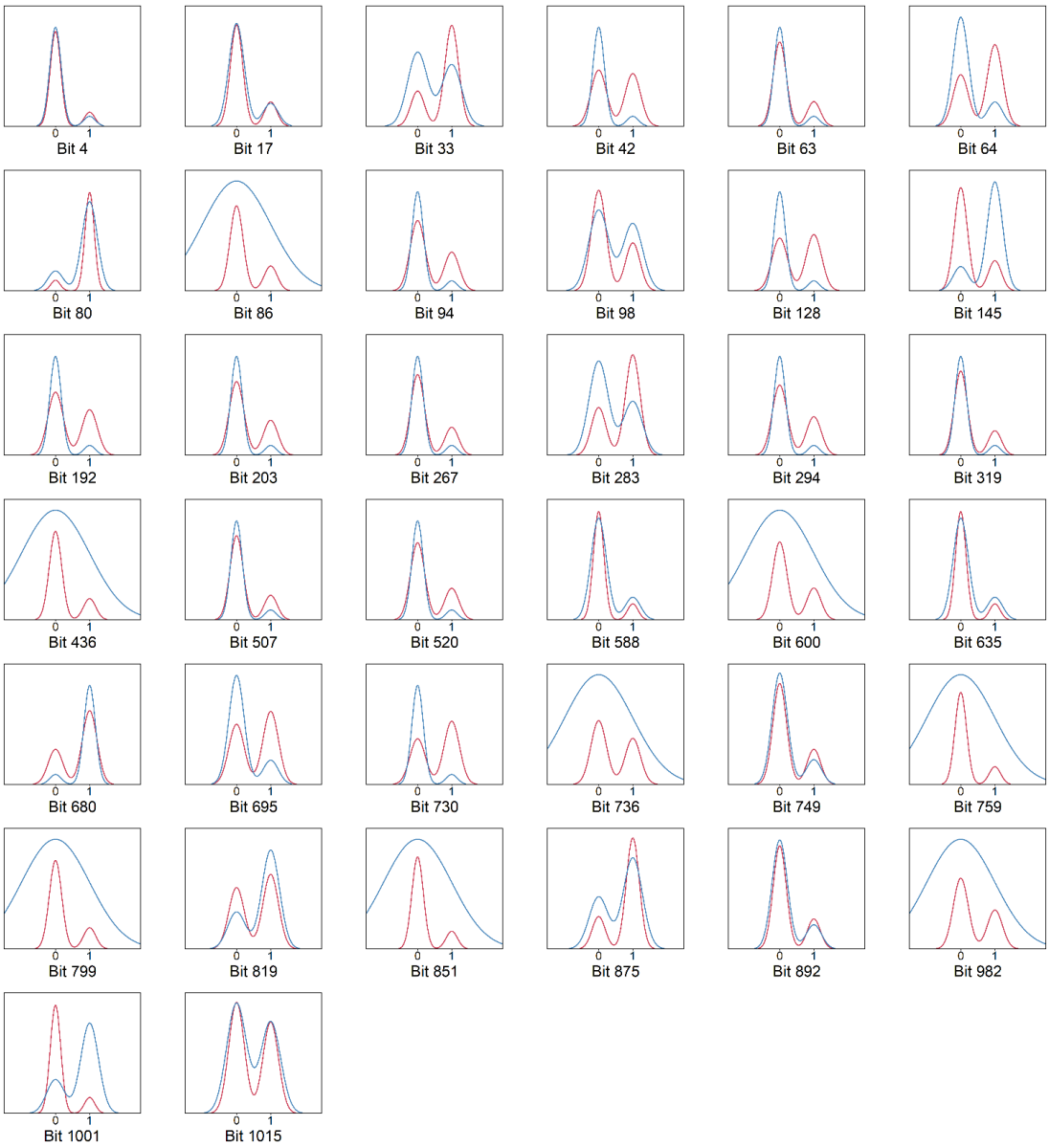
Plot of the 44 Morgan fingerprint descriptors and their distributions, separated by DMSO solubility classification (soluble, red; insoluble, blue) for the training set.



Plot of the 56 Morgan fingerprint descriptors and their distributions, separated by ethanol solubility classification (soluble, red; insoluble, blue) for the training set.



Plot of the 38 Morgan fingerprint descriptors and their distributions, separated by THF solubility classification (soluble, red; insoluble, blue) for the training set.



A 3.9 Partitioning Impact Assessment

Summary of the number of models which provided good predictive statistics for their respective training set and in-domain (maxdist and dsz) test set compounds over 10 random stratified partitions. The number of out-of-domain compounds for each partition is also tabulated.

		Solvent				
Partition	Partition Number	Acetone	DMF	DMSO	Ethanol	THF
Training	1	5	5	5	5	5
	2	6	5	4	4	4
	3	3	4	4	6	3
	4	4	6	5	5	5
	5	5	4	4	6	5
	6	6	5	5	4	5
	7	6	5	4	7	5
	8	6	5	4	7	5
	9	6	5	5	3	4
	10	6	5	5	6	5
Maxdist	1	4	5	4	1	5
	2	2	5	4	0	4
	3	0	5	4	0	1
	4	0	6	5	0	1
	5	1	4	4	3	4
	6	4	5	5	5	5
	7	0	5	5	0	6
	8	0	5	5	0	1
	9	0	5	3	6	4
	10	0	5	5	0	0
Dsz	1	4	5	4	1	5
	2	0	5	0	0	3
	3	3	5	4	0	0
	4	0	5	5	1	0
	5	4	4	4	5	4
	6	4	5	5	3	3
	7	5	5	5	0	5
	8	2	4	5	0	1
	9	0	5	0	2	5
	10	5	5	5	2	5
Dsz Out-of-Domain Compounds	1	4	2	4	1	4
	2	5	5	1	1	5
	3	6	5	6	2	3
	4	2	5	6	1	5
	5	2	3	6	1	1
	6	4	3	6	5	4
	7	4	3	4	4	3
	8	6	5	4	4	5
	9	2	6	3	3	4
	10	5	1	5	3	4

A 3.10 Model Performance Metrics

Optimisation and performance statistics of the QSPR models developed for the training set.

		Resampling results of optimal model	Performance of optimal model on training set				
Model	Method	MCC	ROC AUC	Balanced Accuracy	MCC	K	Overall quality of model
	RF	0.183	0.973	0.932	0.869	0.861	Good
	C4.5	0.420	0.938	0.861	0.725	0.721	Good
	kNN	0.229	0.705	0.652	0.303	0.303	Bad
Acetone	SVM	0.358	0.763	0.500	ND	0.000	Bad
	XG	0.410	0.971	0.907	0.814	0.814	Good
	NB		0.806	0.726	0.504	0.448	Good
	LR	0.228	0.884	0.883	0.768	0.767	Good
	RF	0.946	1.000	0.986	0.913	0.909	Good
	C4.5	0.976	1.000	0.986	0.913	0.909	Good
	kNN	0.798	1.000	0.889	0.577	0.500	Good
DMF	SVM	0.821	0.317	0.500	ND	0.000	Bad
	XG	0.973	1.000	1.000	1.000	1.000	Good
	NB		0.991	0.681	0.273	0.139	Bad
	LR	0.894	1.000	0.986	0.913	0.909	Good
	RF	0.897	0.982	0.986	0.913	0.909	Good
	C4.5	0.947	0.989	0.986	0.913	0.909	Good
	kNN	0.609	0.998	0.851	0.498	0.397	Good
DMSO	SVM	0.679	0.236	0.500	ND	0.000	Bad
	XG	0.897	0.993	0.973	0.842	0.830	Good
	NB		0.975	0.608	0.193	0.071	Bad
	LR	0.849	0.989	0.973	0.842	0.830	Good
	RF	0.535	1.000	1.000	1.000	1.000	Good
	C4.5	0.585	0.998	0.978	0.952	0.951	Good
	kNN	0.480	0.849	0.752	0.505	0.505	Good
Ethanol	SVM	0.495	1.000	1.000	1.000	1.000	Good
	XG	0.522	1.000	1.000	1.000	1.000	Good
	NB		0.810	0.609	0.330	0.196	Bad
	LR	0.486	0.944	0.818	0.655	0.647	Good
	RF	0.873	0.984	0.955	0.938	0.937	Good
	C4.5	0.873	0.987	0.955	0.938	0.937	Good
	kNN	0.588	0.916	0.826	0.578	0.540	Good
THF	SVM	0.701	0.535	0.500	ND	0.000	Bad
	XG	0.871	0.990	0.955	0.938	0.937	Good
	NB		0.905	0.726	0.421	0.301	Bad
	LR	0.683	0.969	0.903	0.773	0.767	Good

Optimisation and performance statistics of the QSPR models developed for the in-domain external test set compounds using the maxdist applicability domain.

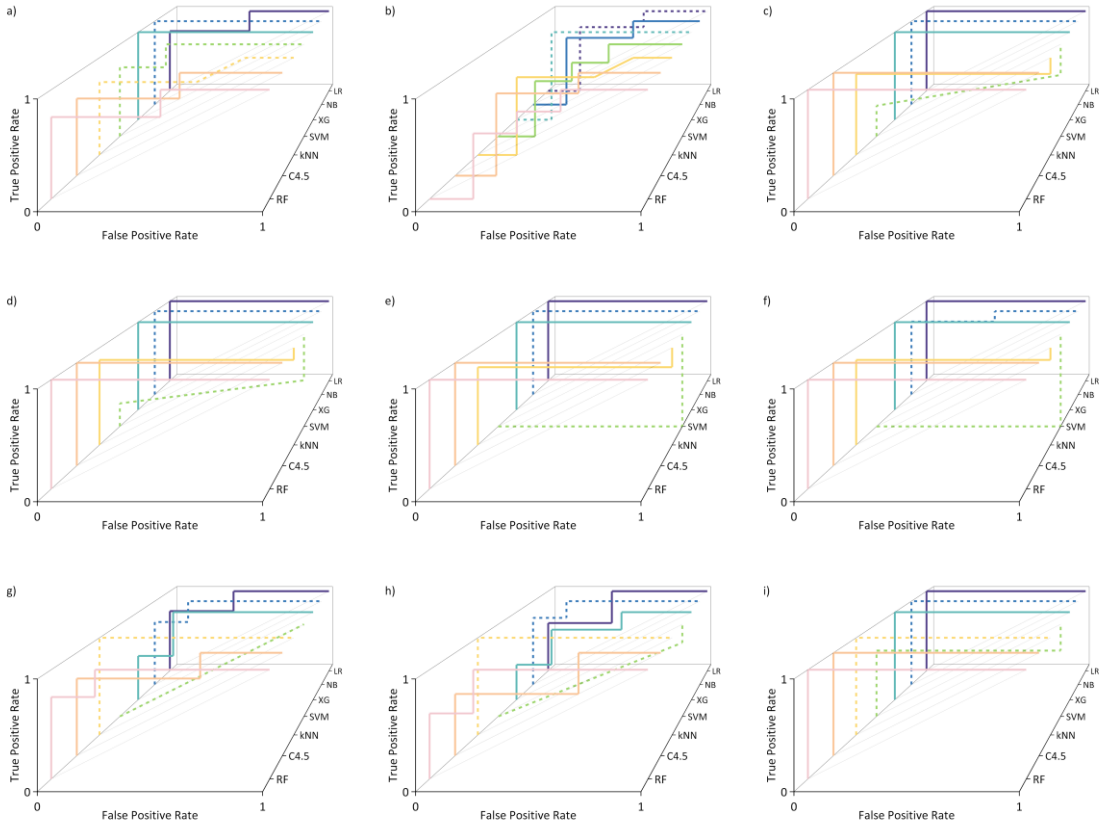
		Performance of optimal model on maxdist in-domain test set				
Model	Method	ROC AUC	Balanced Accuracy	MCC	K	Overall quality of model
Acetone	RF	0.680	0.700	0.408	0.400	Bad
	C4.5	0.720	0.700	0.408	0.400	Bad
	kNN	0.700	0.500	0.000	0.000	Bad
	SVM	0.680	0.500	ND	0.000	Bad
	XG	0.800	0.800	0.600	0.600	Good
	NB	0.720	0.400	-0.333	-0.200	Bad
	LR	0.720	0.700	0.408	0.400	Bad
DMF	RF	1.000	1.000	1.000	1.000	Good
	C4.5	1.000	1.000	1.000	1.000	Good
	kNN	0.875	0.875	0.612	0.545	Good
	SVM	0.375	0.500	ND	0.000	Bad
	XG	1.000	1.000	1.000	1.000	Good
	NB	1.000	0.625	0.250	0.118	Bad
	LR	1.000	1.000	1.000	1.000	Good
DMSO	RF	1.000	0.938	0.764	0.737	Good
	C4.5	1.000	1.000	1.000	1.000	Good
	kNN	0.875	0.938	0.764	0.737	Good
	SVM	0.000	0.500	ND	0.000	Bad
	XG	1.000	0.750	0.667	0.615	Good
	NB	0.938	0.625	0.250	0.118	Bad
	LR	1.000	0.938	0.764	0.737	Good
Ethanol	RF	0.920	0.700	0.408	0.400	Good
	C4.5	0.760	0.800	0.655	0.600	Good
	kNN	0.580	0.600	0.333	0.200	Bad
	SVM	0.400	0.500	ND	0.000	Bad
	XG	0.800	0.700	0.500	0.400	Good
	NB	0.96	0.500	ND	0.000	Bad
	LR	0.840	0.700	0.500	0.400	Good
THF	RF	1.000	1.000	1.000	1.000	Good
	C4.5	1.000	0.929	0.756	0.727	Good
	kNN	1.000	0.786	0.478	0.372	Bad
	SVM	0.714	0.500	ND	0.000	Bad
	XG	1.000	0.929	0.756	0.727	Good
	NB	1.000	0.643	0.286	0.151	Bad
	LR	1.000	1.000	1.000	1.000	Good

Optimisation and performance statistics of the QSPR models developed for the in-domain external test set compounds using the dsz applicability domain.

		Performance of optimal model on dsz in-domain test set				
Model	Method	ROC AUC	Balanced Accuracy	MCC	K	Overall quality of model
Acetone	RF	0.872	0.875	0.775	0.750	Good
	C4.5	0.875	0.875	0.755	0.750	Good
	kNN	0.844	0.625	0.378	0.250	Bad
	SVM	0.938	0.500	ND	0.000	Bad
	XG	1.000	0.875	0.775	0.750	Good
	NB	1.000	0.500	ND	0	Bad
	LR	0.875	0.875	0.775	0.750	Good
DMF	RF	1.000	1.000	1.000	1.000	Good
	C4.5	1.000	1.000	1.000	1.000	Good
	kNN	0.833	0.833	0.577	0.500	Good
	SVM	0.500	0.500	ND	0.000	Bad
	XG	1.000	1.000	1.000	1.000	Good
	NB	1.000	0.500	ND	0.000	Bad
	LR	1.000	1.000	1.000	1.000	Good
DMSO	RF	1.000	0.900	0.632	0.571	Good
	C4.5	1.000	1.000	1.000	1.000	Good
	kNN	0.800	0.900	0.632	0.571	Good
	SVM	0.000	0.500	ND	0.000	Bad
	XG	1.000	1.000	1.000	1.000	Good
	NB	1.000	0.500	ND	0.000	Bad
	LR	1.000	0.900	0.632	0.571	Good
Ethanol	RF	0.950	0.775	0.550	0.550	Good
	C4.5	0.850	0.875	0.791	0.769	Good
	kNN	0.575	0.625	0.395	0.270	Bad
	SVM	0.400	0.500	ND	0.000	Bad
	XG	0.900	0.750	0.598	0.526	Good
	NB	0.960	0.500	ND	0.000	Bad
	LR	0.900	0.750	0.598	0.526	Good
THF	RF	1.000	1.000	1.000	1.000	Good
	C4.5	1.000	0.929	0.756	0.727	Good
	kNN	1.000	0.786	0.478	0.372	Bad
	SVM	0.714	0.500	ND	0.000	Bad
	XG	1.000	0.929	0.756	0.727	Good
	NB	1.000	0.643	0.286	0.151	Bad
	LR	1.000	1.000	1.000	1.000	Good

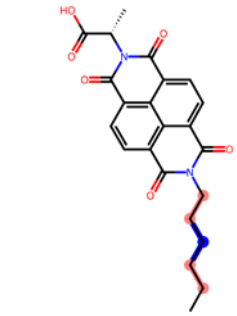
A 3.11 Receiver Operator Characteristic Curves

Receiver operator characteristic (ROC) curves for all reported optimised models. Curves with a solid line represent models with good predictive statistics, while curves with a dashed line represents models with bad predictive statistics, as described in Table S3 and S4. Curves have been separated on the basis of solvent model and applicability domain as follows: a) acetone, dsz; b) acetone, maxdist; c) DMF, dsz; d) DMF, maxdist; e) DMSO, dsz; f) DMSO, maxdist; g) ethanol, dsz; h) ethanol, maxdist; i) THF, dsz and maxdist (identical ROC curves).

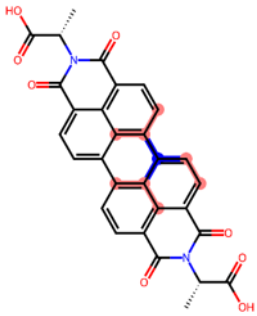


A 3.12 Feature Importance

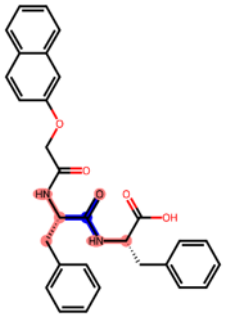
Descriptor feature importance of the 9 most important substructures for the acetone model. The core atom represented by the fingerprint bit of interest is highlighted in blue, with the extended connectivity captured by the structural descriptor highlighted in red. The quantified importance (Equation S8) of each of these descriptors is also provided.



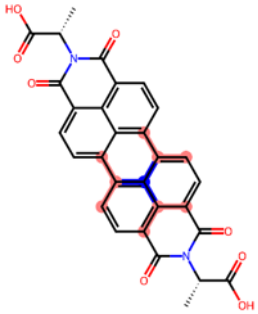
Bit 119 (1.48)



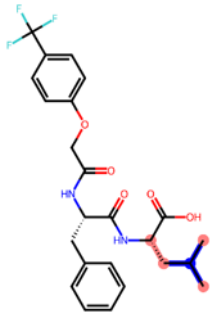
Bit 1001 (1.25)



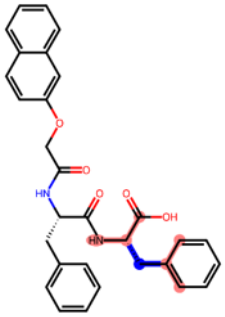
Bit 86 (0.90)



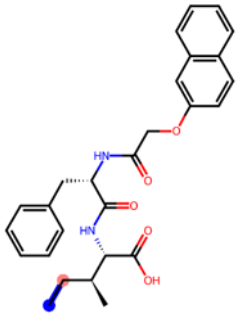
Bit 145 (0.87)



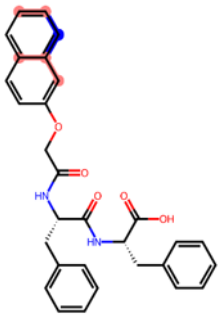
Bit 759 (0.62)



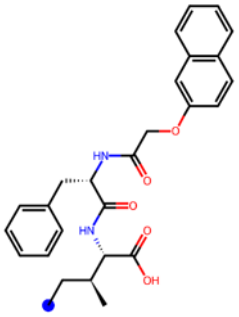
Bit 982 (0.50)



Bit 294 (0.49)

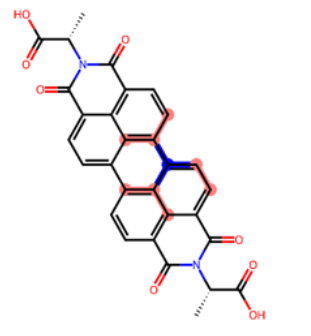


Bit 63 (0.45)

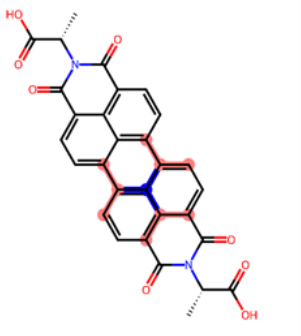


Bit 33 (0.45)

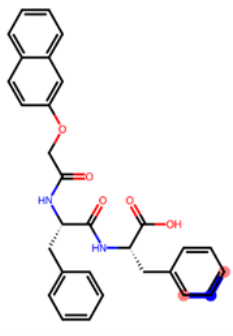
Descriptor feature importance of the 9 most important substructures for the DMF model. The core atom represented by the fingerprint bit of interest is highlighted in blue, with the extended connectivity captured by the structural descriptor highlighted in red. The quantified importance (Equation S8) of each of these descriptors is also provided.



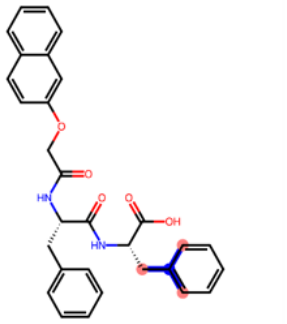
Bit 1001 (1.81)



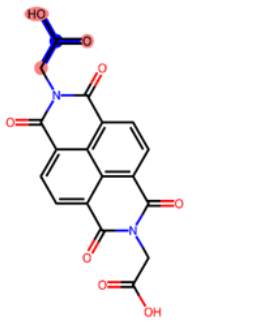
Bit 145 (1.49)



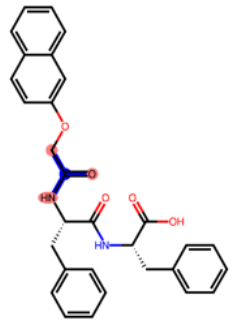
Bit 64 (1.23)



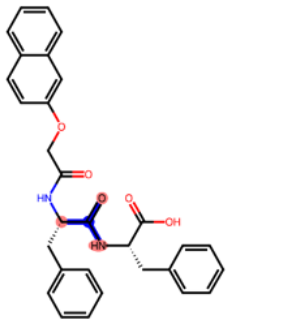
Bit 730 (1.04)



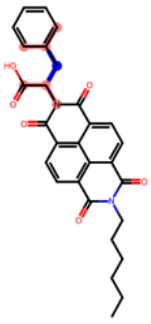
Bit 713 (0.95)



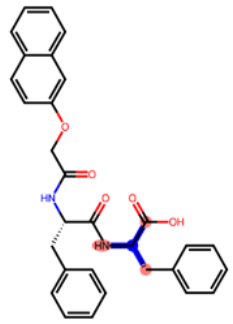
Bit 41 (0.65)



Bit 573 (0.55)

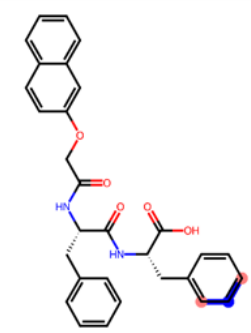


Bit 267 (0.46)

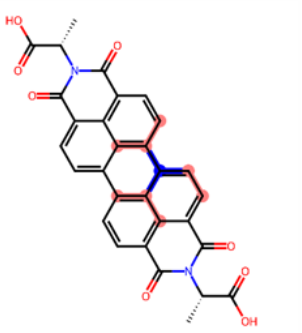


Bit 736 (0.38)

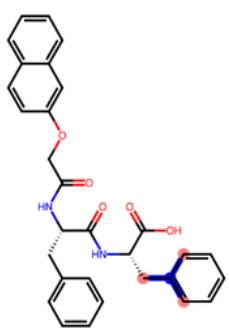
Descriptor feature importance of the 9 most important substructures for the DMSO model. The core atom represented by the fingerprint bit of interest is highlighted in blue, with the extended connectivity captured by the structural descriptor highlighted in red. The quantified importance (Equation S8) of each of these descriptors is also provided.



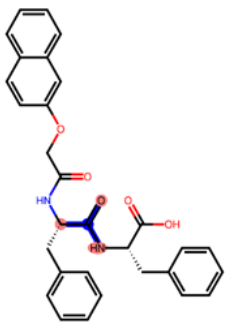
Bit 64 (1.36)



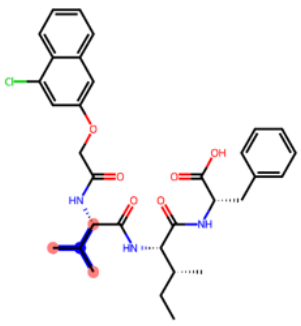
Bit 1001 (1.21)



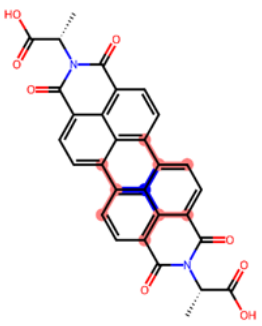
Bit 730 (1.19)



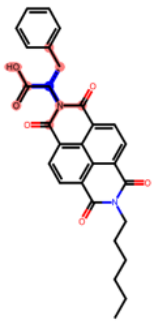
Bit 573 (0.95)



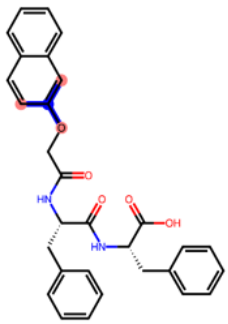
Bit 749 (0.82)



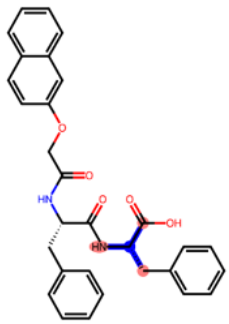
Bit 145 (0.73)



Bit 892 (0.61)



Bit 322 (0.54)

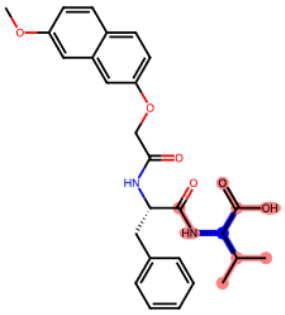


Bit 736 (0.49)

Descriptor feature importance of the 9 most important substructures for the ethanol model. The core atom represented by the fingerprint bit of interest is highlighted in blue, with the extended connectivity captured by the structural descriptor highlighted in red. The quantified importance (Equation S8) of each of these descriptors is also provided.

Rotatable Bonds (1.20)

Labute Atomic Surface Area (1.18)

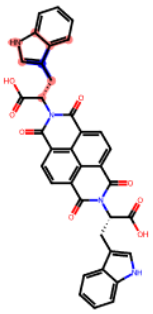


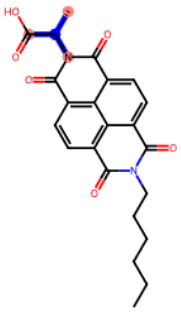
Number of Atoms (1.15)

SLogP (1.08)

Bit 752 (0.856)

Number of Rings (0.848)



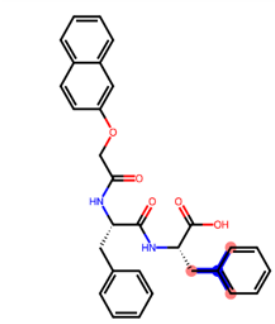


Total Polar Surface Area (0.68)

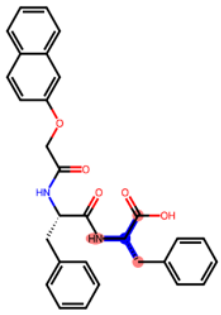
Bit 946 (0.48)

Bit 708 (0.48)

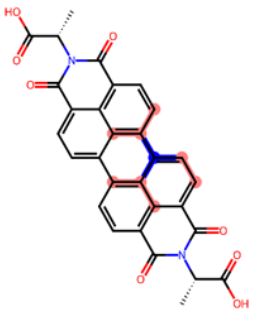
Descriptor feature importance of the 9 most important substructures for the THF model. The core atom represented by the fingerprint bit of interest is highlighted in blue, with the extended connectivity captured by the structural descriptor highlighted in red. The quantified importance (Equation S8) of each of these descriptors is also provided.



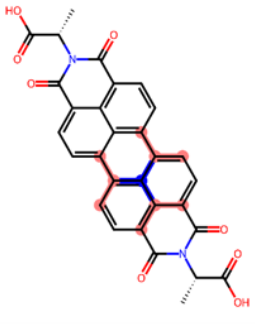
Bit 730 (1.31)



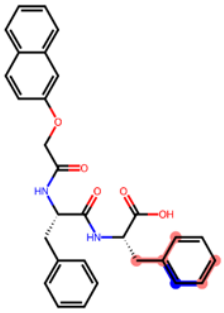
Bit 736 (1.14)



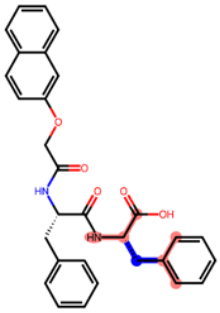
Bit 1001 (1.08)



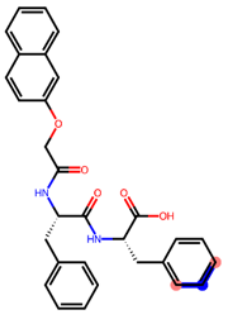
Bit 145 (0.97)



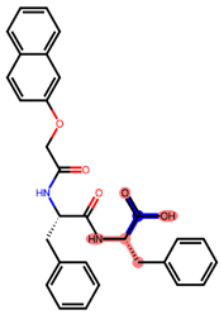
Bit 42 (0.80)



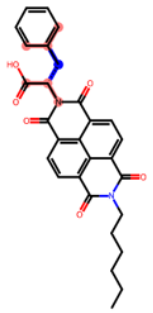
Bit 982 (0.759)



Bit 64 (0.66)



Bit 600 (0.64)

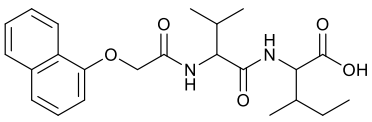


Bit 267 (0.58)

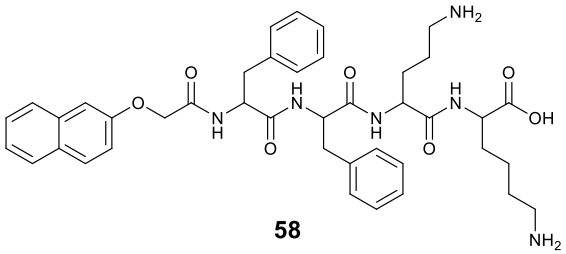
Chapter 4 Appendices

A 4.1 Model Molecules

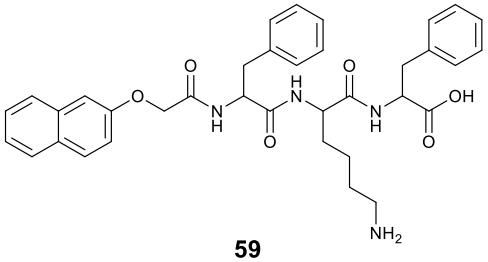
In addition to the molecules described in Chapter 3, four additional molecules made previously by members of the Draper and Adams groups were characterised.



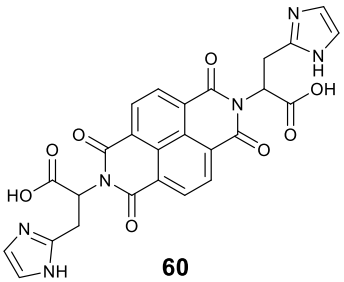
57



58



59



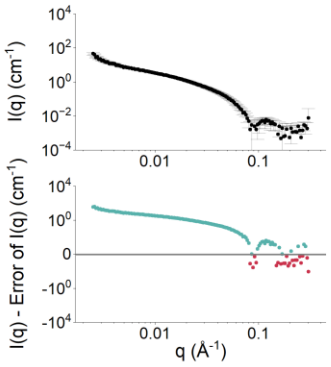
60

A 4.2 Scattering Classification

A 4.2.1 5 mgmL^{-1} Aqueous Solution Classifications

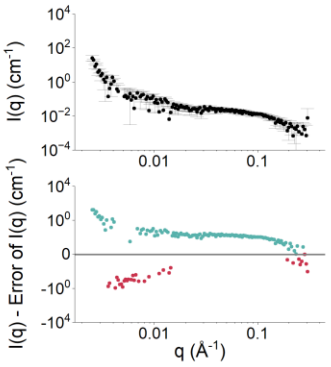
1. 2Nap-FF H_2O 5 mgmL^{-1}

Classification: Good Scattering



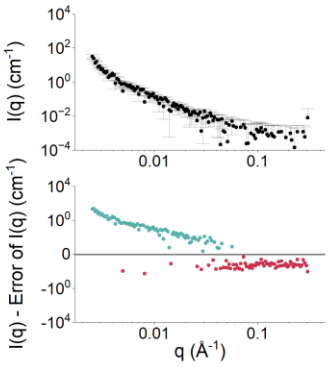
2. 2Nap-V H_2O 5 mgmL^{-1}

Classification: Bad Scattering



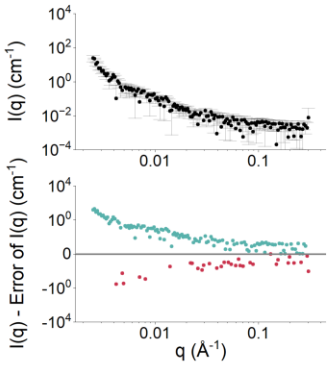
4. 2Nap-IF H_2O 5 mgmL^{-1}

Classification: Bad Scattering



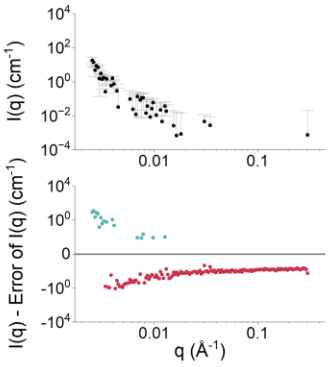
5. 6MeO2Nap-FI H_2O 5 mgmL^{-1}

Classification: Bad Scattering



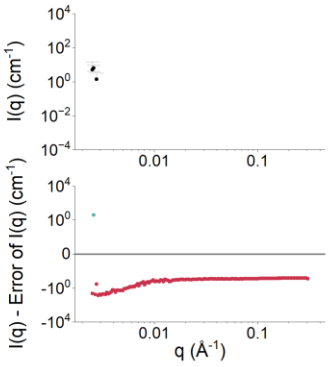
6. 7MeO2Nap-FV H_2O 5 mgmL^{-1}

Classification: Bad Scattering



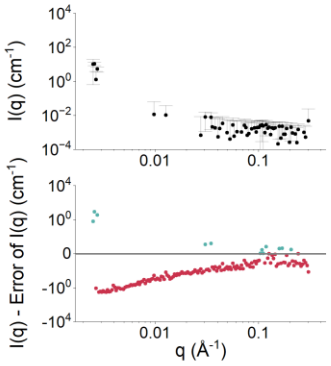
9. NDI-F H_2O 5 mgmL^{-1}

Classification: Bad Scattering



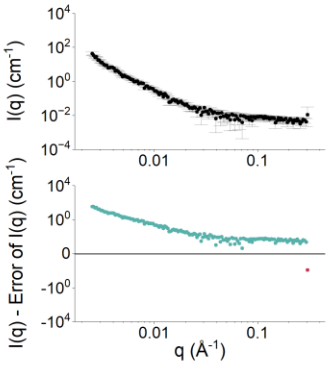
11. NDI-G H_2O 5 mgmL^{-1}

Classification: Bad Scattering



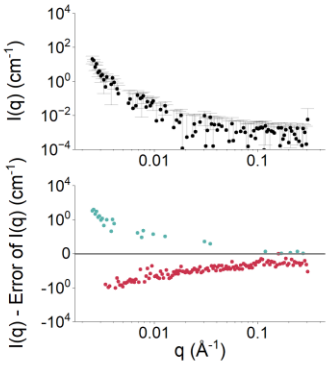
12. NDI-GF H_2O 5 mgmL^{-1}

Classification: Good Scattering



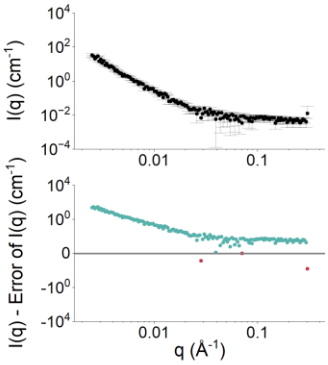
13. NDI-A H_2O 5 mgmL^{-1}

Classification: Bad Scattering



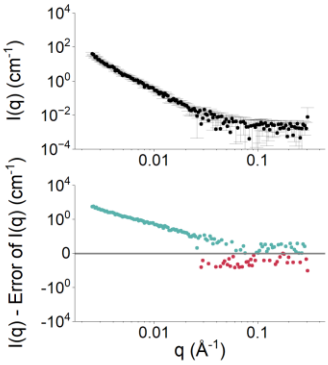
14. NDI-I H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



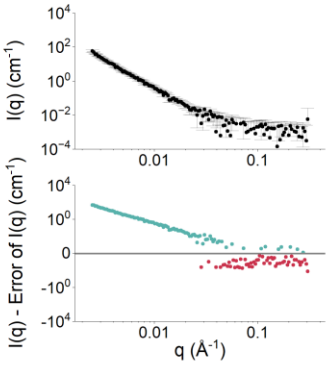
15. NDI-L H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



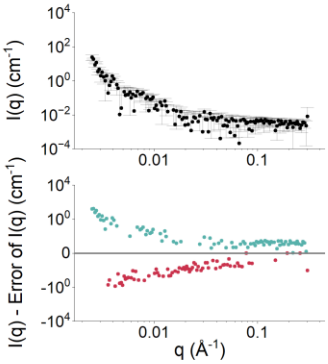
16. NDI-M H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



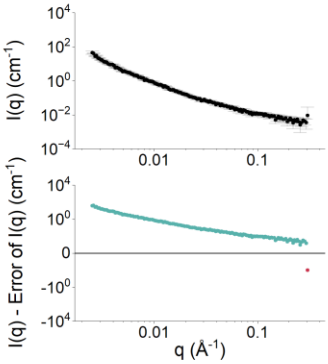
17. NDI-V H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



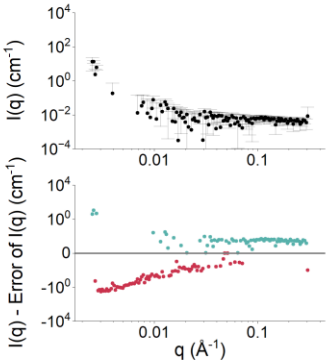
18. NDI-W H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



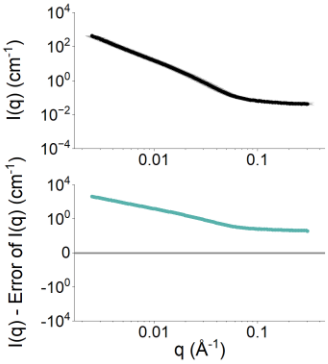
19. NDI-Y H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



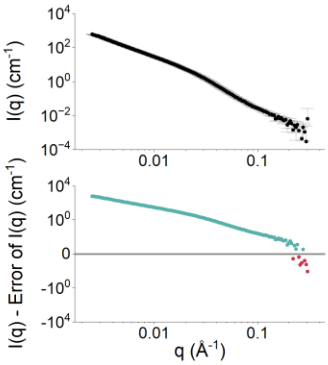
20. PBI-A H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



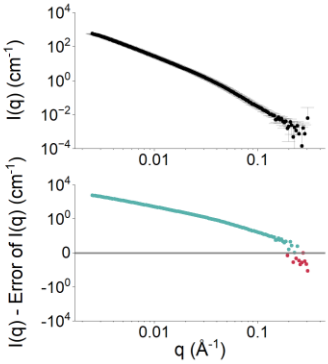
21. PBI-D H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



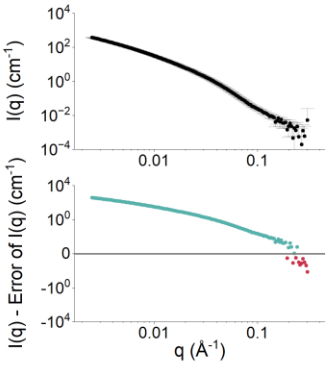
22. PBI-DOPA H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



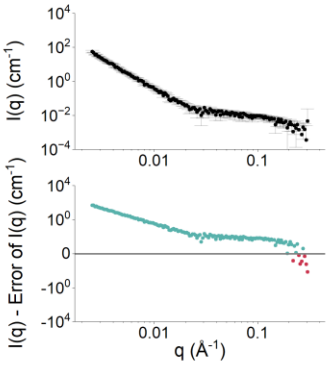
23. PBI-E H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



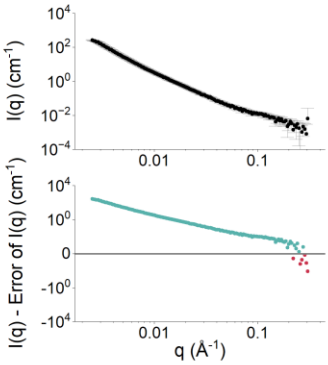
24. PBI-F H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



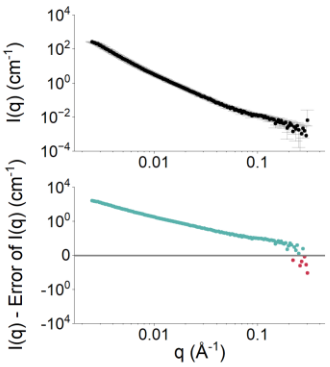
25. PBI-G H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



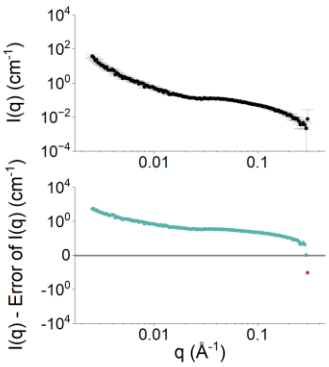
25. PBI-G H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



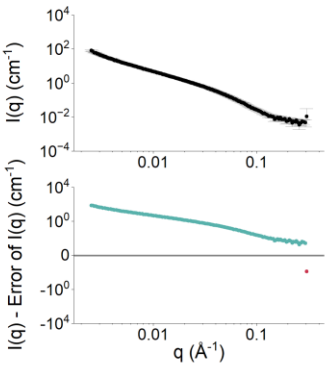
26. PBI-H H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



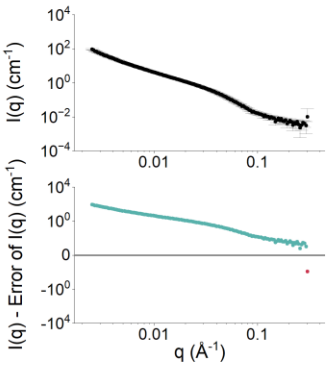
27. PBI-I H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



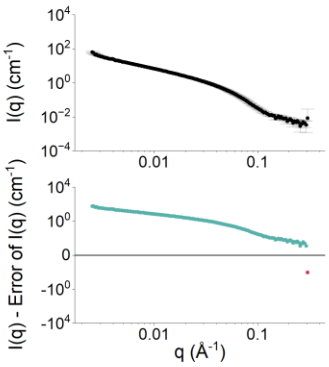
28. PBI-L H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



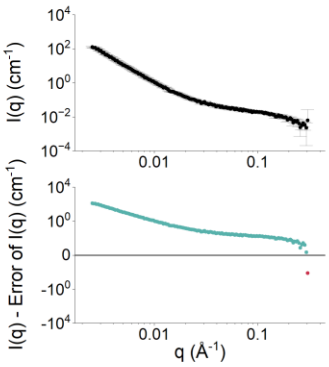
29. PBI-M H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



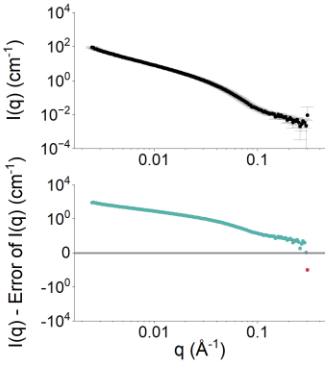
30. PBI-S H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



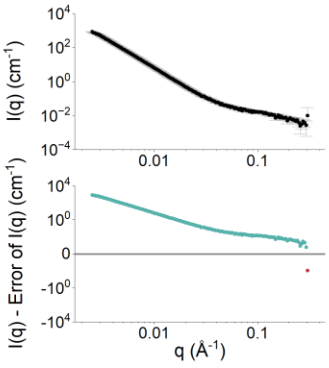
31. PBI-V H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



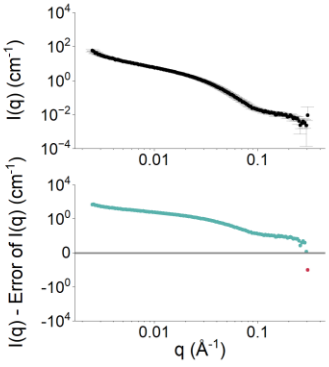
32. PBI-W H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



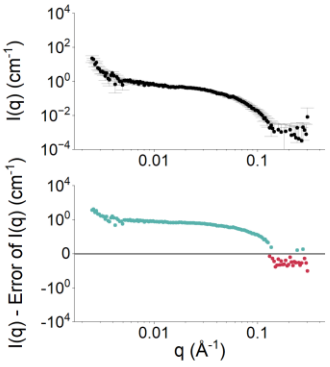
33. PBI-Y H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



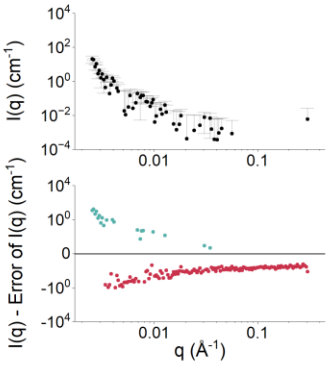
34. ThNap-FF H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



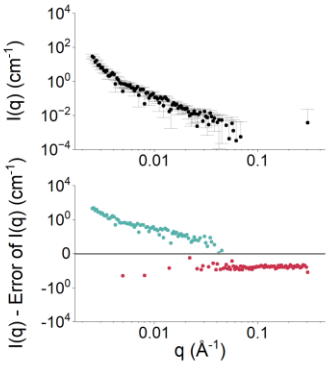
35. 2Nap-VL H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



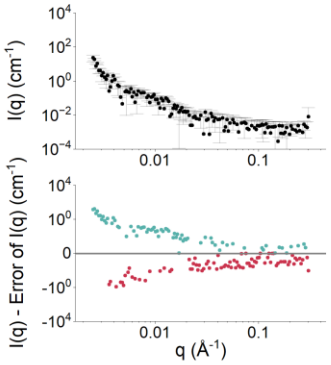
36. 2Nap-FL H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



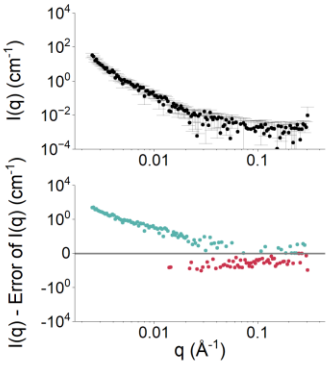
37. 2Nap-FI H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



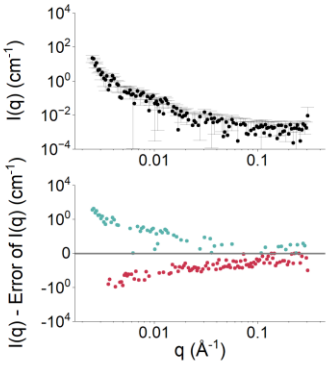
38. 2Nap-VF H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



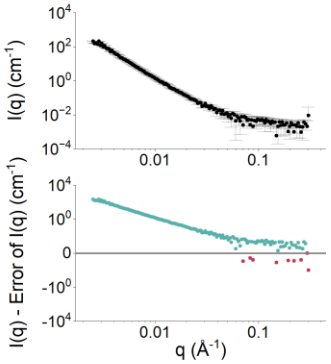
39. 2Nap-IV H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



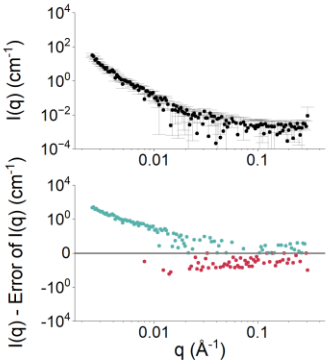
41. CF3Ph-FV H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



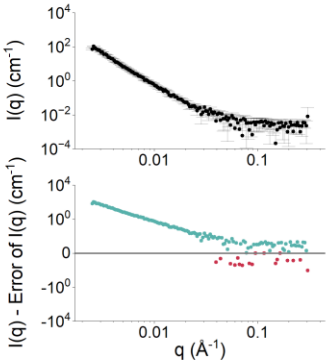
42. 2Nap-VI H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



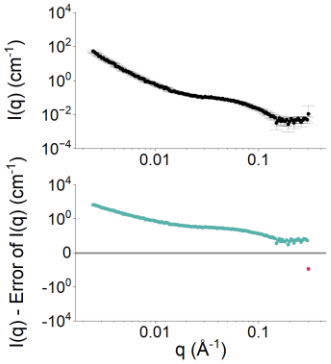
43. C6-NDI-A H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



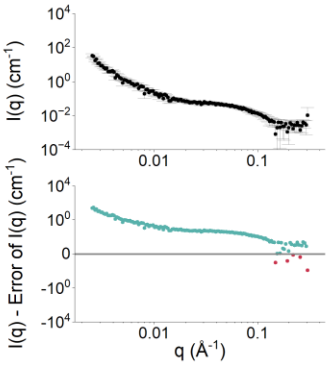
44. C6-NDI-F H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



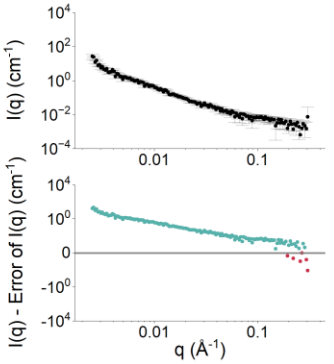
45. C6-NDI-L H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



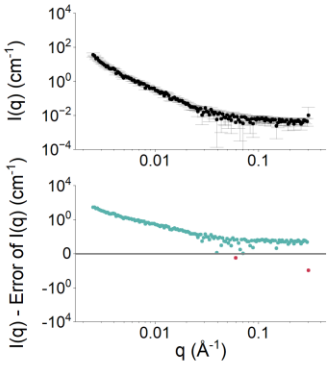
46. NDI-GA H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



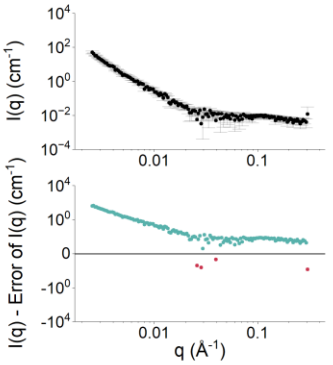
48. NDI-LA H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



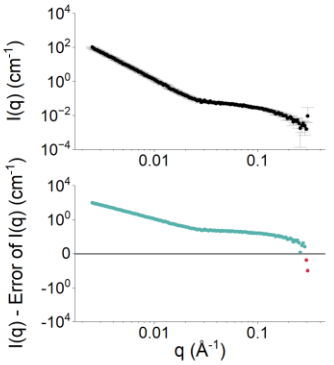
49. NDI-YI H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



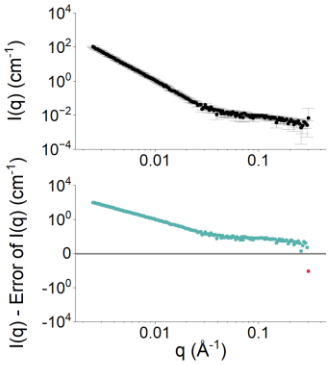
50. PBI-A 1Py H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



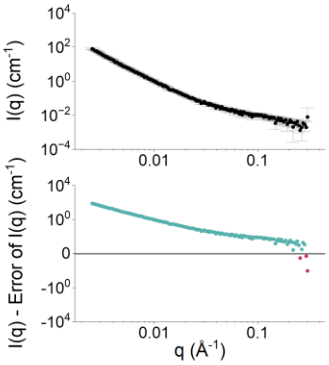
51. PBI-I 1Py H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



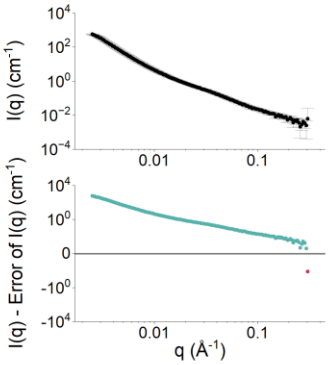
52. PBI-L 1Py H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



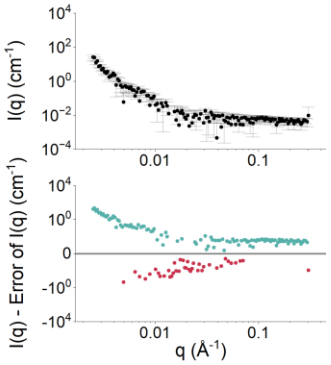
53. PBI-V 1Py H_2O 5 $mgmL^{-1}$

Classification: Good Scattering



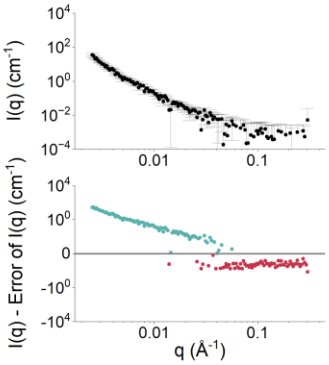
56. NDI-VI H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



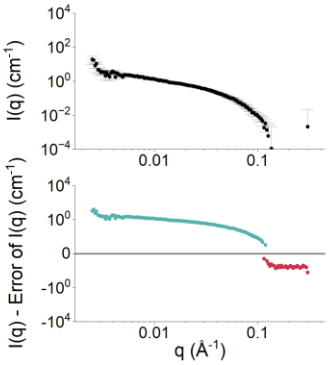
57. 1NapFL H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



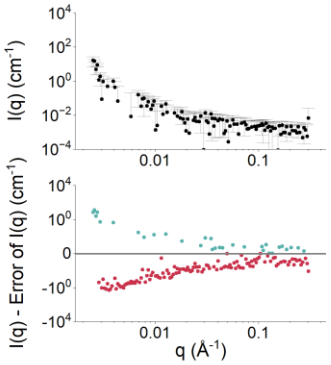
58. 2NapFFKK H_2O 5 $mgmL^{-1}$

Classification: Bad Scattering



60. NDIH H_2O 5 $mgmL^{-1}$

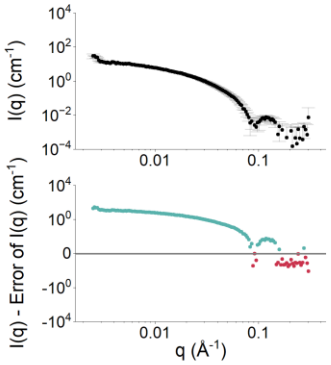
Classification: Bad Scattering



A 4.2.2 10 mgmL^{-1} Aqueous Solution Classifications

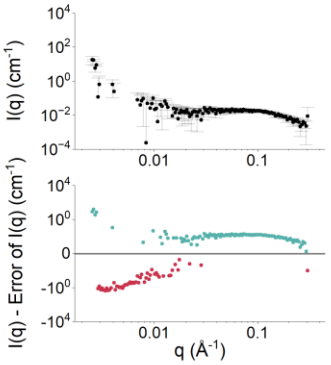
1. 2Nap-FF H_2O 10 mgmL^{-1}

Classification: Good Scattering



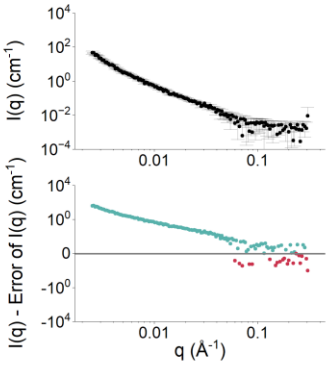
2. 2Nap-V H_2O 10 mgmL^{-1}

Classification: Bad Scattering



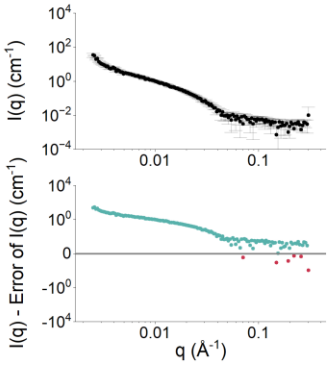
3. 2Nap-VVV H_2O 10 mgmL^{-1}

Classification: Good Scattering



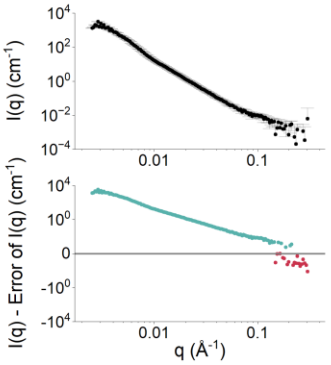
4. 2Nap-IF H_2O 10 mgmL^{-1}

Classification: Good Scattering



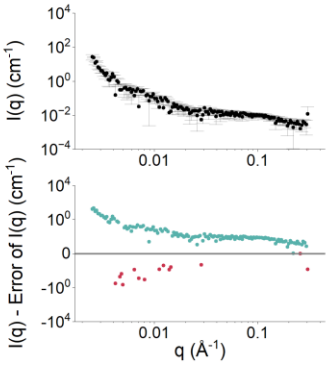
5. 6Br2Nap-VF H_2O 10 mgmL^{-1}

Classification: Good Scattering



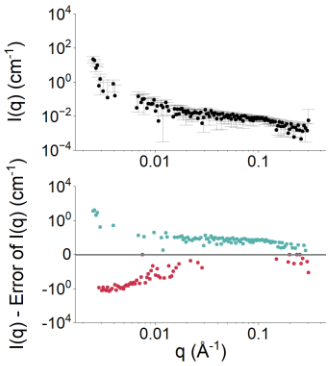
6. 6MeO2Nap-FI H_2O 10 mgmL^{-1}

Classification: Bad Scattering



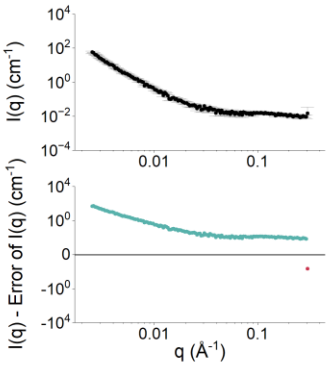
7. 7MeO2NapFV H_2O 10 mgmL^{-1}

Classification: Bad Scattering



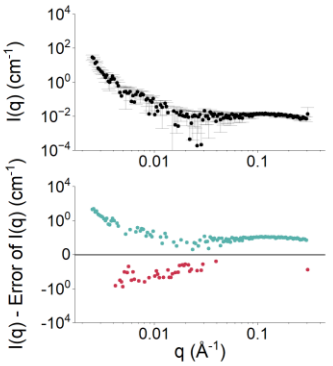
9. NDI-F H_2O 10 mgmL^{-1}

Classification: Good Scattering



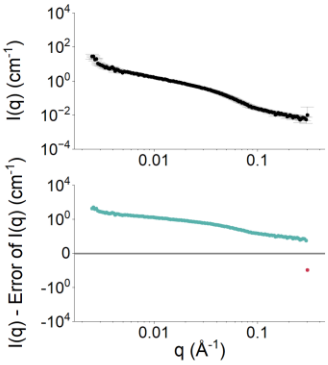
11. NDI-G H_2O 10 mgmL^{-1}

Classification: Bad Scattering



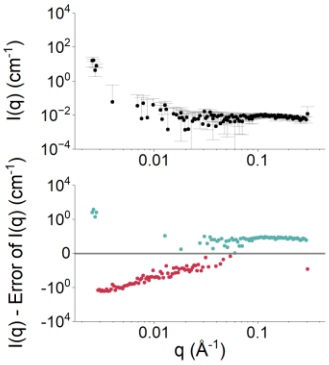
12. NDI-GF H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



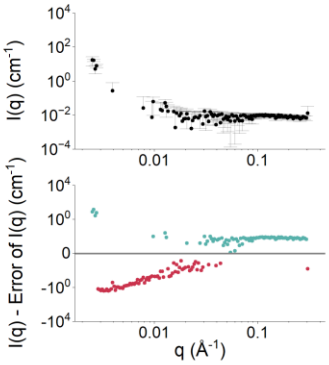
13. NDI-A H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



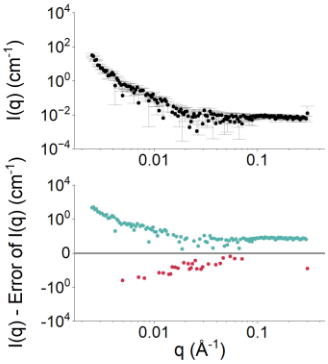
14. NDI-I H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



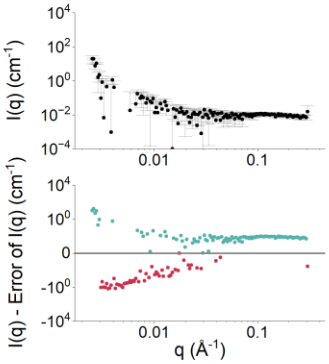
15. NDI-L H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



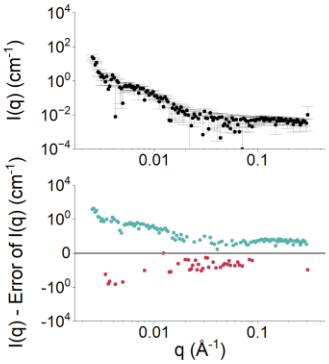
16. NDI-M H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



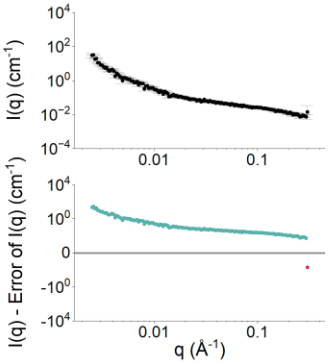
17. NDI-V H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



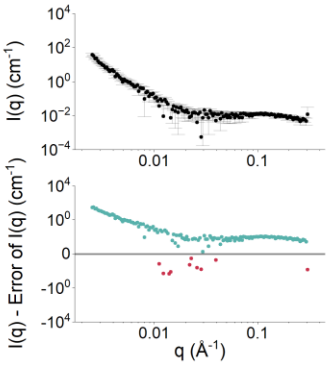
18. NDI-W H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



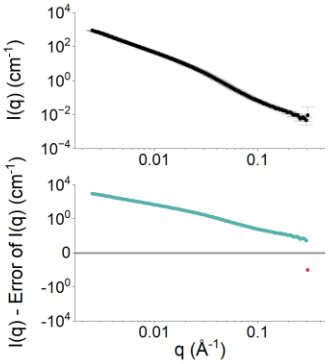
19. NDI-Y H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



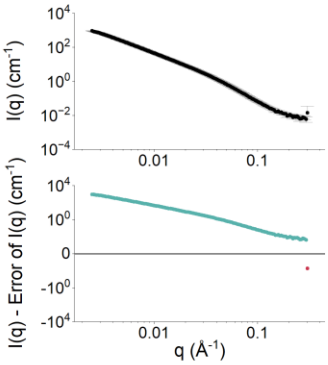
21. PBI-D H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



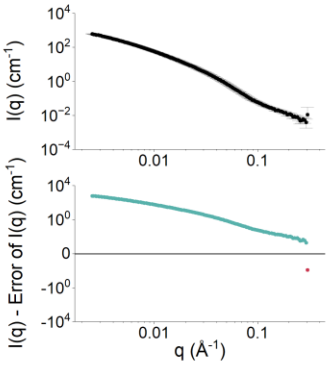
22. PBI-DOPA H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



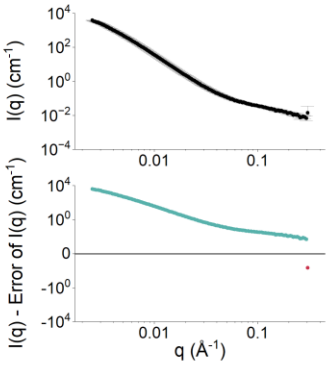
23. PBI-E H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



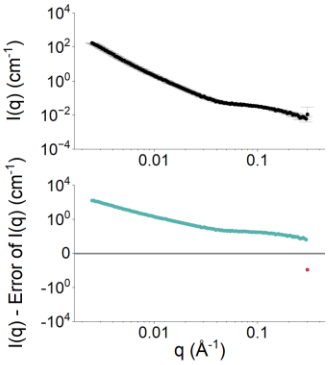
24. PBI-F H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



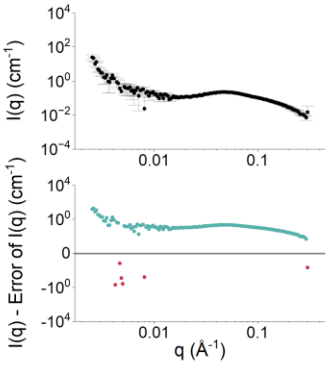
25. PBI-G H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



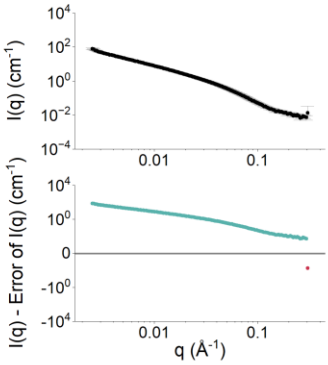
26. PBI-H H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



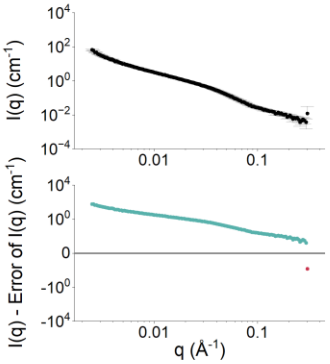
27. PBI-I H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



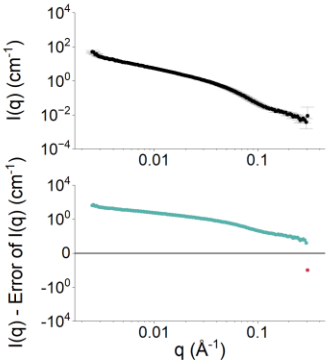
28. PBI-L H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



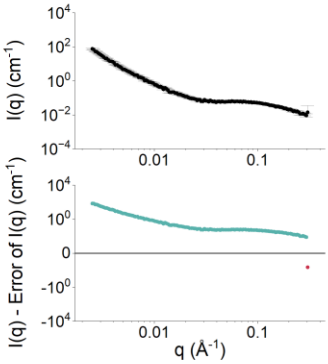
29. PBI-M H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



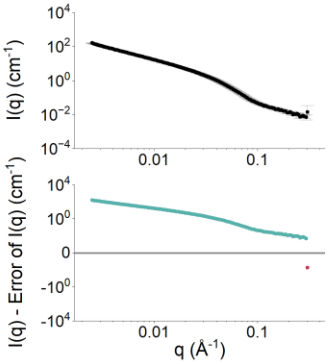
30. PBI-S H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



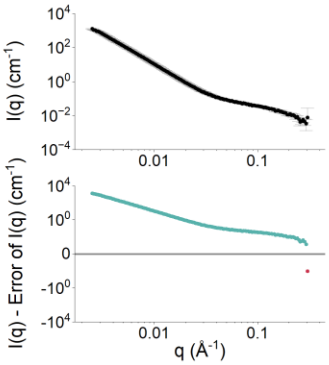
31. PBI-V H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



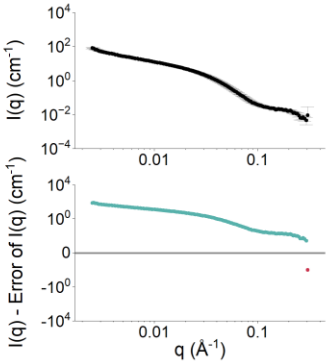
32. PBI-W H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



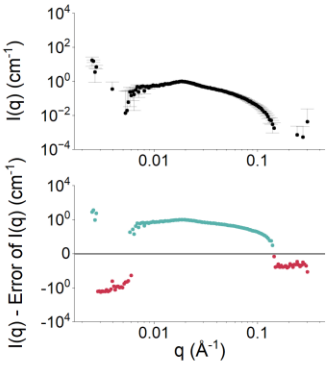
33. PBI-Y H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



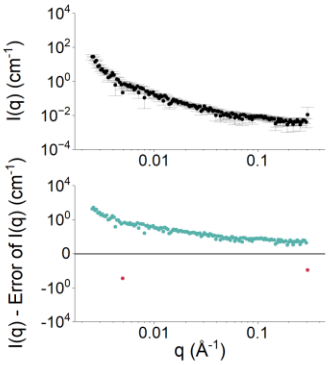
34. ThNap-FF H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



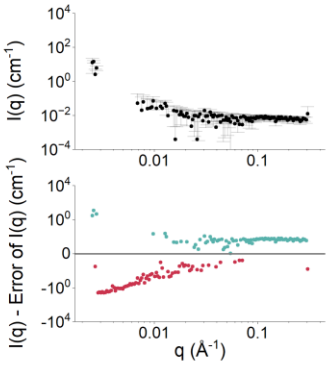
37. 2Nap-FI H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



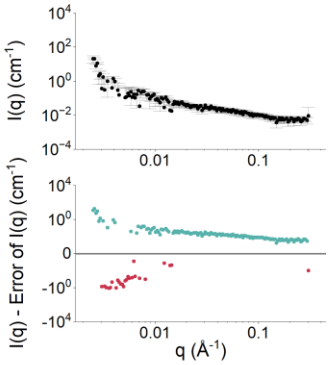
43. C6-NDI-A H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



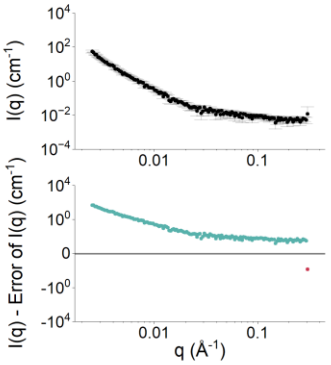
44. C6-NDI-F H_2O 10 $mgmL^{-1}$

Classification: Bad Scattering



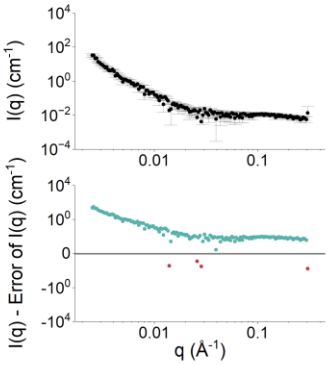
45. C6-NDI-L H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



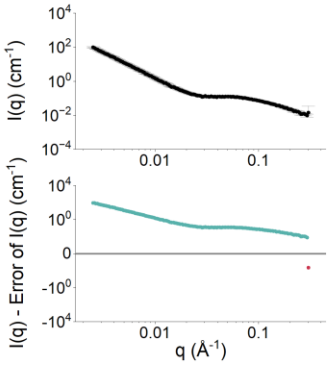
49. NDI-YI H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



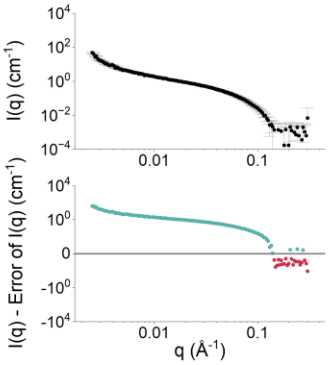
50. PBI-A 1Py H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



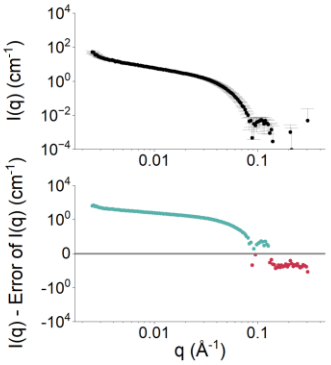
58. 2Nap-FFKK H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



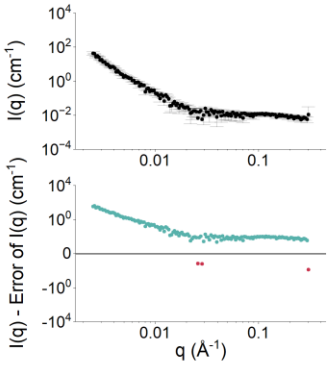
59. 2Nap-FKF H_2O 10 $mgmL^{-1}$

Classification: Good Scattering



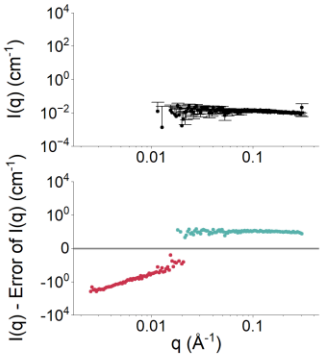
60. NDI-H H_2O 10 $mgmL^{-1}$

Classification: Good Scattering

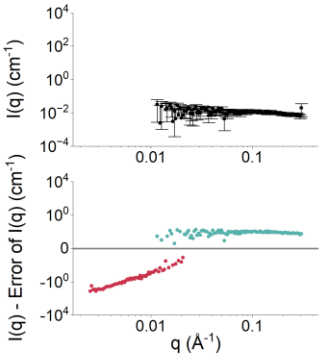


A 4.2.3 Acetone Solution Classifications

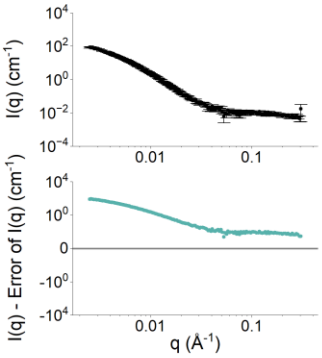
2. 2Nap-V Acetone 5 mgmL^{-1}
Classification: Bad Scattering



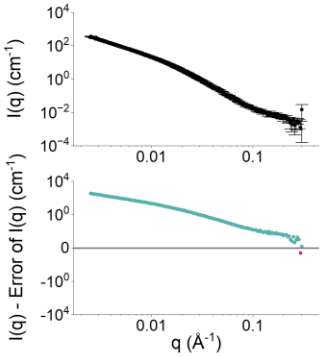
4. 2Nap-IF Acetone 5 mgmL^{-1}
Classification: Bad Scattering



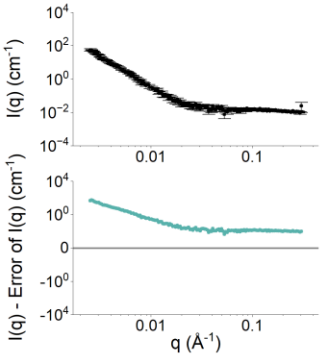
5. 6Br2Nap-VF Acetone 5 mgmL^{-1}
Classification: Good Scattering



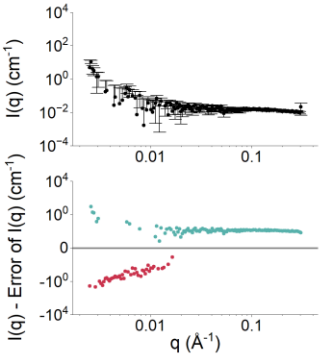
6. 6MeO2Nap-FI Acetone 5 mgmL^{-1}
Classification: Good Scattering



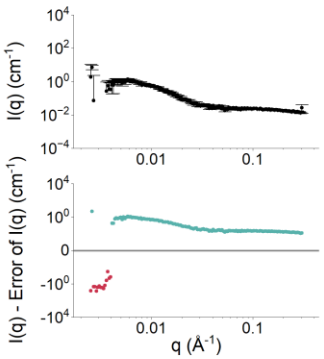
14. NDI-I Acetone 5 mgmL^{-1}
Classification: Good Scattering



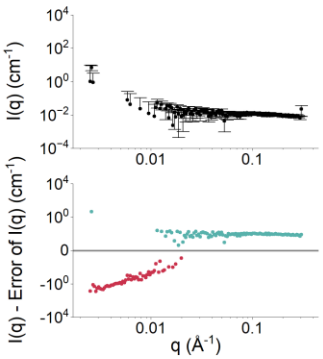
15. NDI-L Acetone 5 mgmL^{-1}
Classification: Bad Scattering



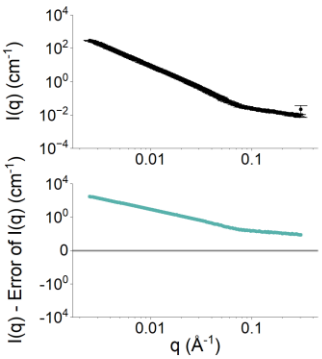
16. NDI-M Acetone 5 mgmL^{-1}
Classification: Bad Scattering



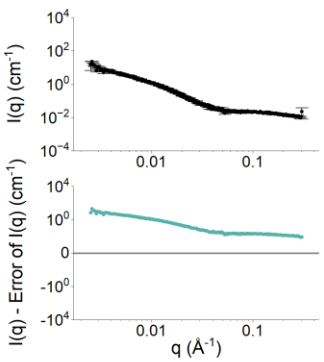
17. NDI-V Acetone 5 mgmL^{-1}
Classification: Bad Scattering



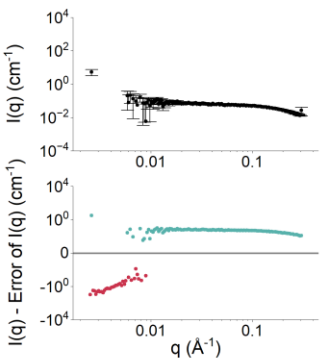
18. NDI-W Acetone 5 mgmL^{-1}
Classification: Good Scattering



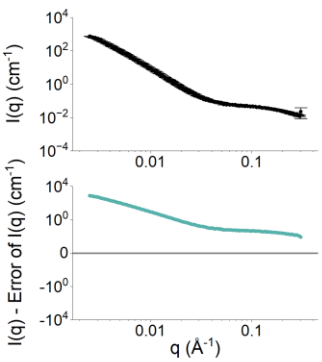
19. NDI-Y Acetone 5 mgmL^{-1}
Classification: Good Scattering



22. PBI-DOPA Acetone 5 mgmL^{-1}
Classification: Bad Scattering

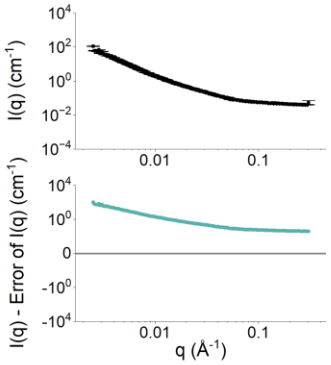


32. PBI-W Acetone 5 mgmL^{-1}
Classification: Good Scattering



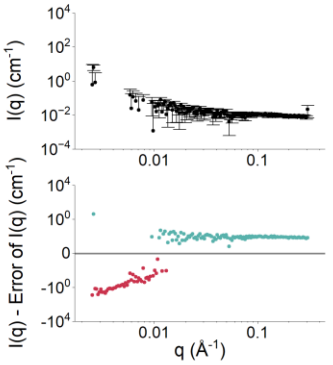
36. 2Nap-FL Acetone 5 mgmL^{-1}

Classification: Good Scattering



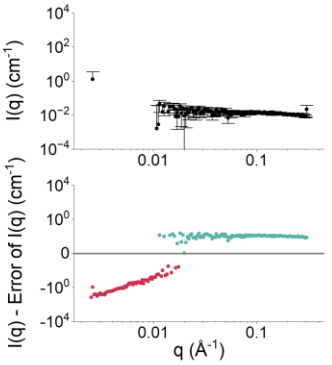
37. 2Nap-FI Acetone 5 mgmL^{-1}

Classification: Bad Scattering



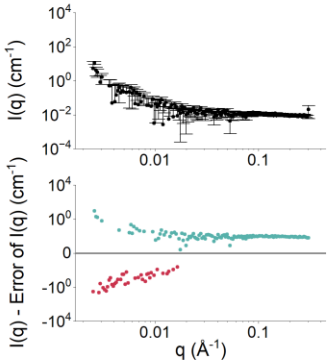
41. CF3Ph-FV Acetone 5 mgmL^{-1}

Classification: Bad Scattering



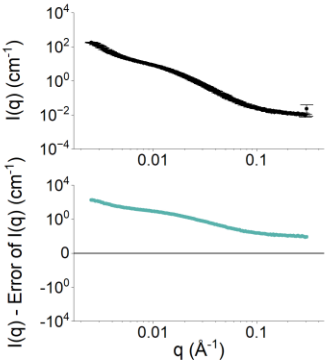
42. 2Nap-VI Acetone 5 mgmL^{-1}

Classification: Bad Scattering



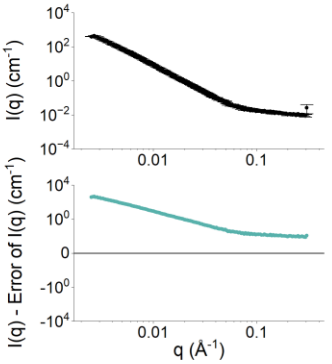
43. C6-NDI-A Acetone 5 mgmL^{-1}

Classification: Good Scattering



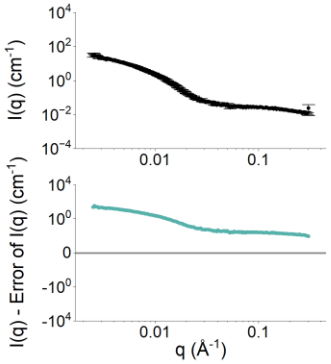
45. C6-NDI-L Acetone 5 mgmL^{-1}

Classification: Good Scattering



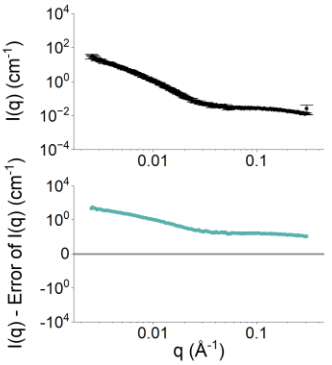
51. PBI-I 1Py Acetone 5 mgmL^{-1}

Classification: Good Scattering



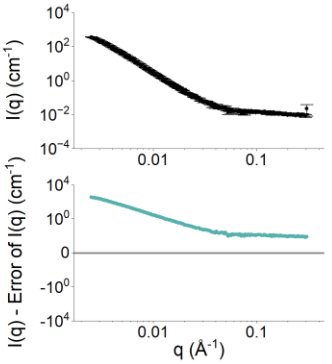
52. PBI-L 1Py Acetone 5 mgmL^{-1}

Classification: Good Scattering



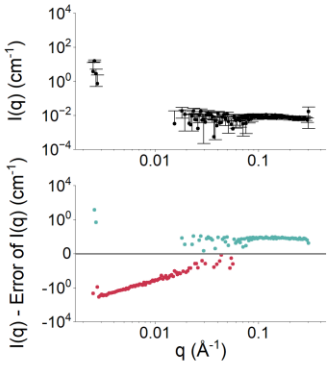
53. PBI-V 1Py Acetone 5 mgmL^{-1}

Classification: Good Scattering



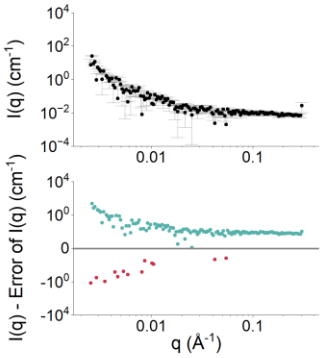
57. 1Nap-FL Acetone 5 mgmL^{-1}

Classification: Bad Scattering

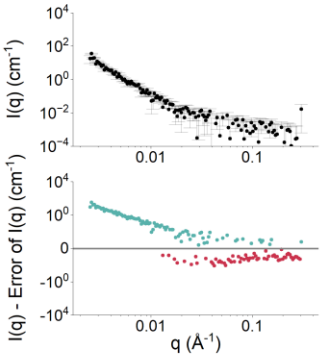


A 4.2.4 DMF Solution Classifications

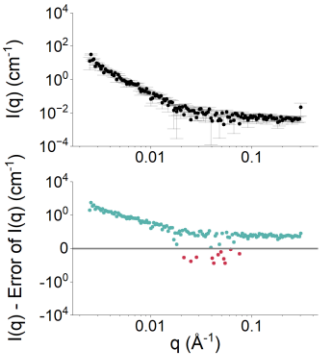
1. 2Nap-FF DMF 5 mgmL^{-1}
Classification: Bad Scattering



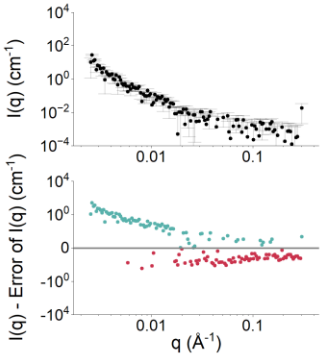
2. 2Nap-V DMF 5 mgmL^{-1}
Classification: Bad Scattering



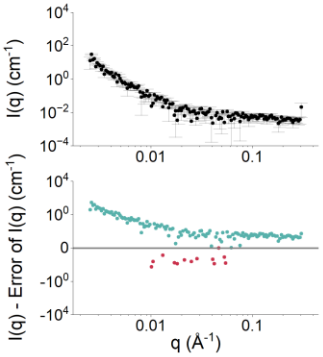
3. 2Nap-VVV DMF 5 mgmL^{-1}
Classification: Bad Scattering



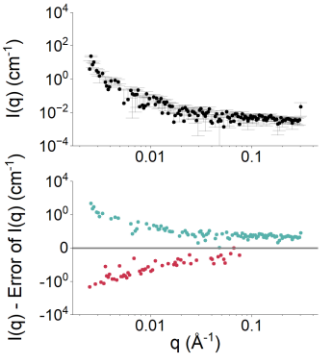
4. 2Nap-IF DMF 5 mgmL^{-1}
Classification: Bad Scattering



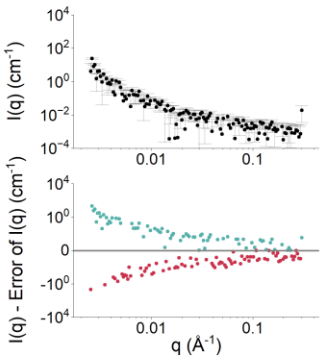
5. 6Br2Nap-VF DMF 5 mgmL^{-1}
Classification: Bad Scattering



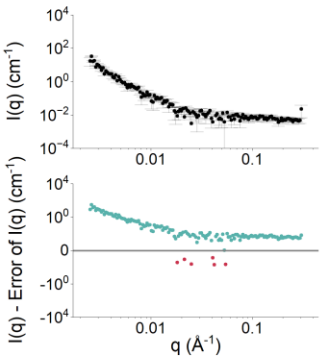
6. 6MeO2Nap-FI DMF 5 mgmL^{-1}
Classification: Bad Scattering



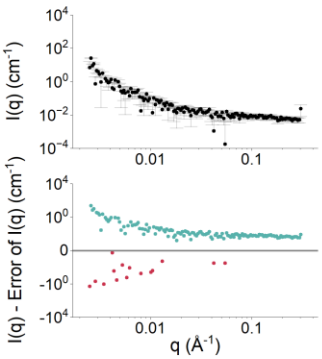
7. 7MeO2Nap-FV DMF 5 mgmL^{-1}
Classification: Bad Scattering



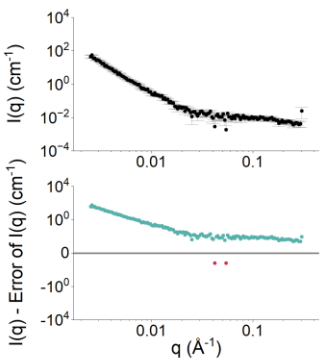
9. NDI-F DMF 5 mgmL^{-1}
Classification: Good Scattering



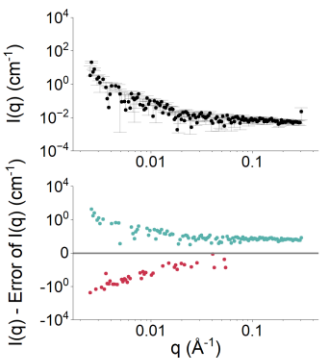
11. NDI-G DMF 5 mgmL^{-1}
Classification: Bad Scattering



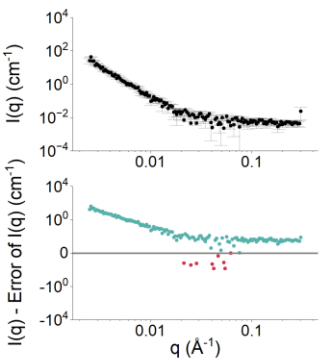
12. NDI-GF DMF 5 mgmL^{-1}
Classification: Good Scattering



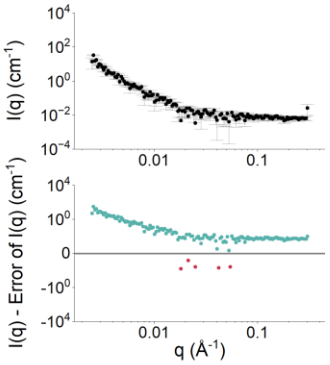
13. NDI-A DMF 5 mgmL^{-1}
Classification: Bad Scattering



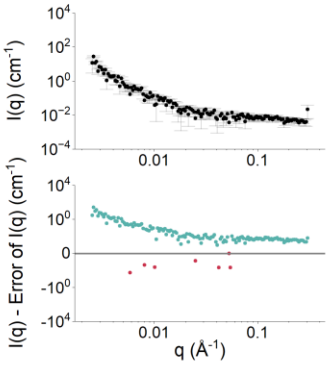
14. NDI-I DMF 5 mgmL^{-1}
Classification: Bad Scattering



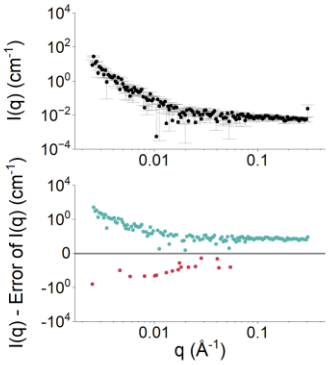
15. NDI-L DMF 5 mgmL^{-1}
Classification: Good Scattering



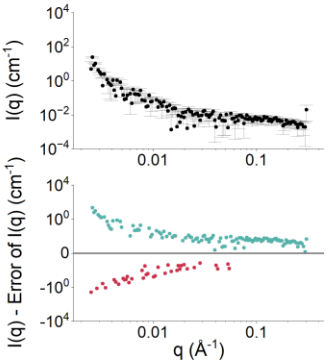
16. NDI-M DMF 5 mgmL^{-1}
Classification: Good Scattering



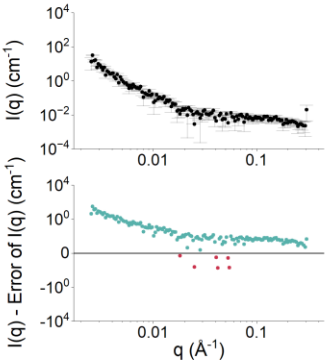
17. NDI-V DMF 5 mgmL^{-1}
Classification: Bad Scattering



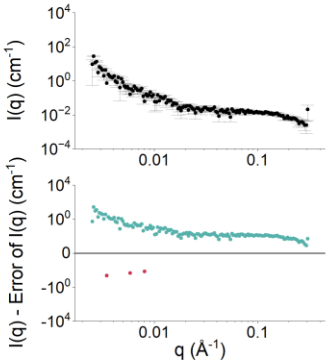
18. NDI-W DMF 5 mgmL^{-1}
Classification: Bad Scattering



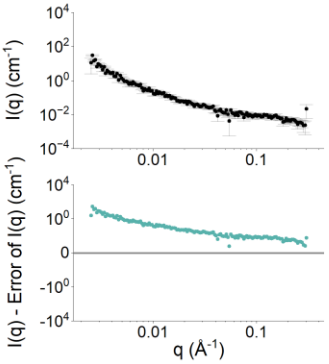
19. NDI-Y DMF 5 mgmL^{-1}
Classification: Good Scattering



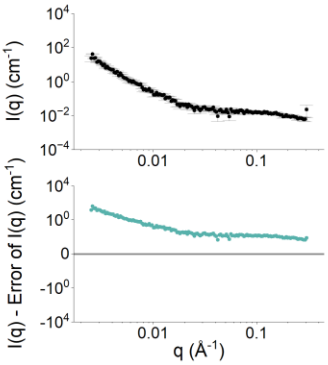
22. PBI-DOPA DMF 5 mgmL^{-1}
Classification: Good Scattering



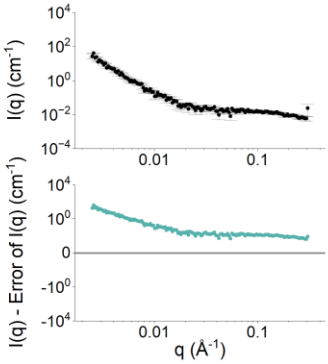
24. PBI-F DMF 5 mgmL^{-1}
Classification: Good Scattering



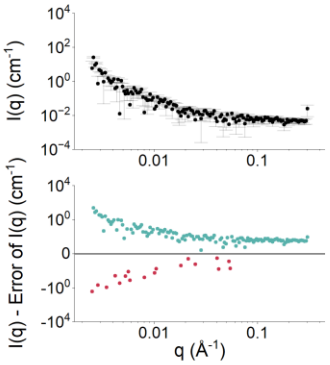
32. PBI-W DMF 5 mgmL^{-1}
Classification: Good Scattering



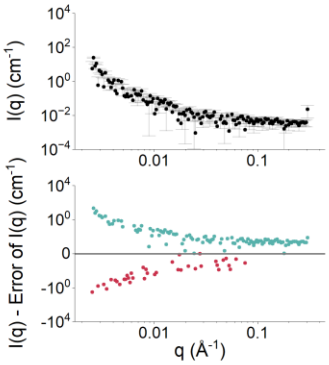
33. PBI-Y DMF 5 mgmL^{-1}
Classification: Good Scattering



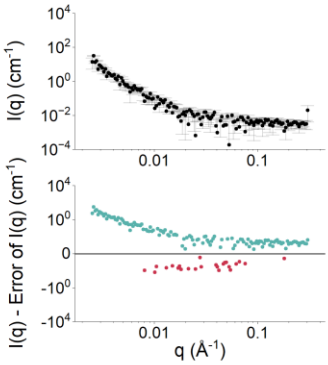
34. ThNapFF DMF 5 mgmL^{-1}
Classification: Bad Scattering



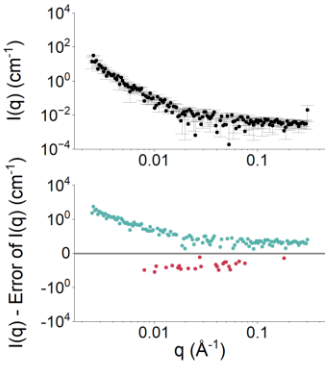
35. 2Nap-VL DMF 5 mgmL^{-1}
Classification: Bad Scattering



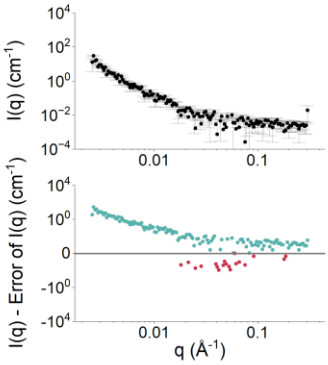
36. 2Nap-FL DMF 5 mgmL^{-1}
Classification: Bad Scattering



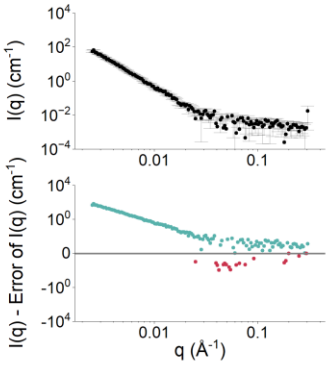
36. 2Nap-FL DMF 5 mgmL^{-1}
Classification: Bad Scattering



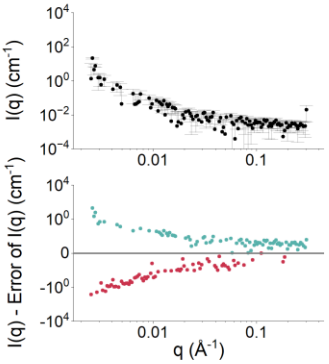
37. 2NapFl DMF 5 mgmL^{-1}
Classification: Bad Scattering



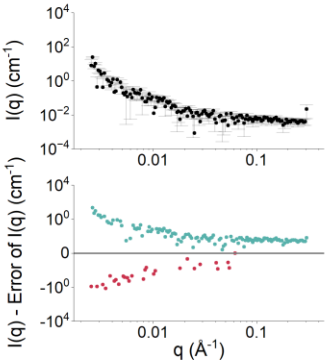
38. 2Nap-VF DMF 5 mgmL^{-1}
Classification: Bad Scattering



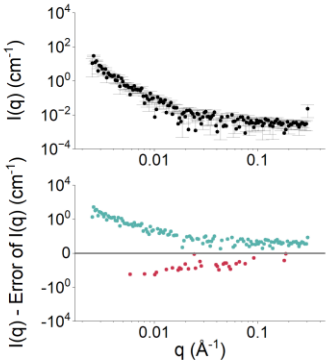
39. 2Nap-IV DMF 5 mgmL^{-1}
Classification: Bad Scattering



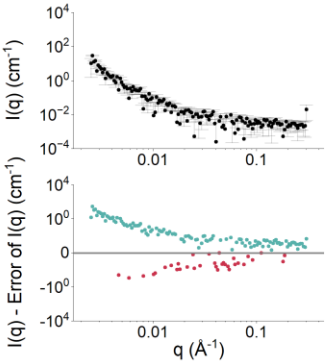
41. CF3PhFV DMF 5 mgmL^{-1}
Classification: Bad Scattering



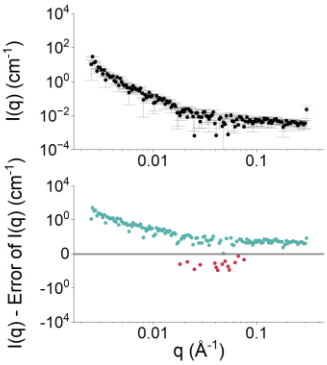
42. 2Nap-VI DMF 5 mgmL^{-1}
Classification: Bad Scattering



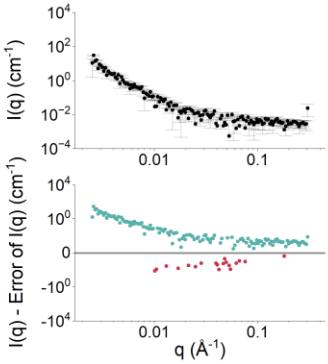
43. C6-NDI-A DMF 5 mgmL^{-1}
Classification: Bad Scattering



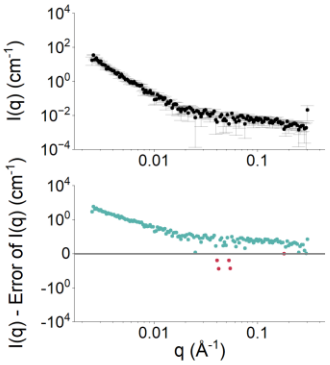
44. C6-NDI-F DMF 5 mgmL^{-1}
Classification: Bad Scattering



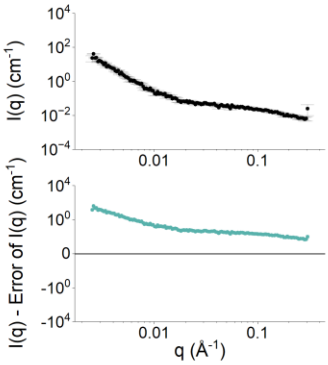
45. C6-NDI-L DMF 5 mgmL^{-1}
Classification: Bad Scattering



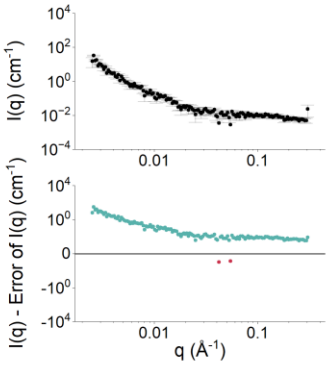
49. NDI-YI DMF 5 mgmL^{-1}
Classification: Good Scattering



50. PBI-A 1Py DMF 5 mgmL^{-1}
Classification: Good Scattering

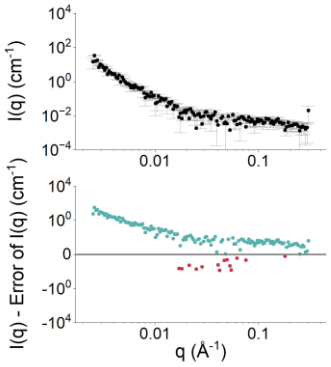


51. PBI-I 1Py DMF 5 mgmL^{-1}
Classification: Good Scattering



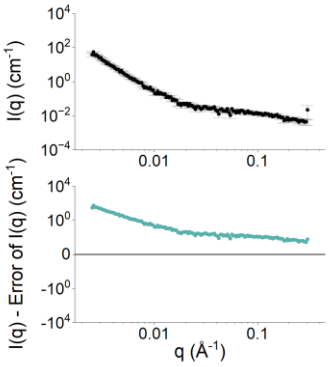
52. PBI-L 1Py DMF 5 mgmL^{-1}

Classification: Bad Scattering



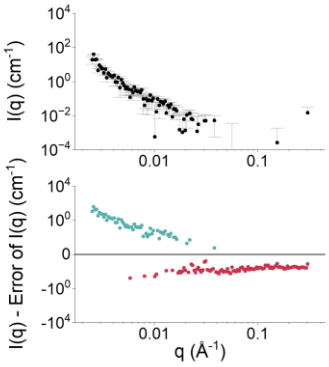
53. PBI-V 1Py DMF 5 mgmL^{-1}

Classification: Good Scattering



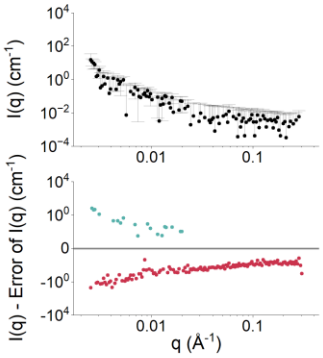
57. 1Nap-FL DMF 5 mgmL^{-1}

Classification: Bad Scattering

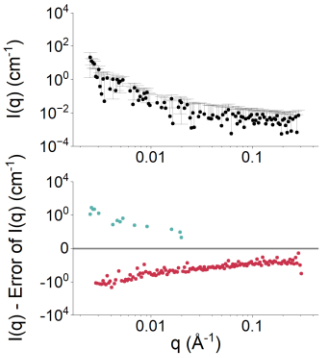


A 4.2.5 DMSO Solution Classifications

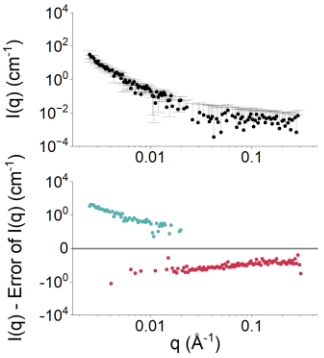
1. 2Nap-FF DMSO 5 mgmL^{-1}
Classification: Bad Scattering



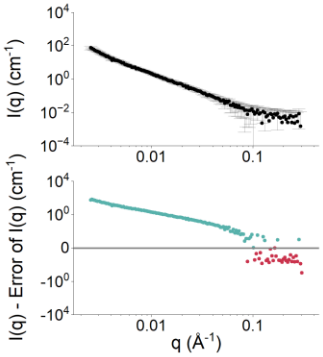
2. 2Nap-V DMSO 5 mgmL^{-1}
Classification: Bad Scattering



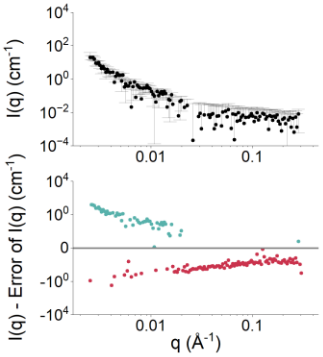
3. 2Nap-VVV DMSO 5 mgmL^{-1}
Classification: Bad Scattering



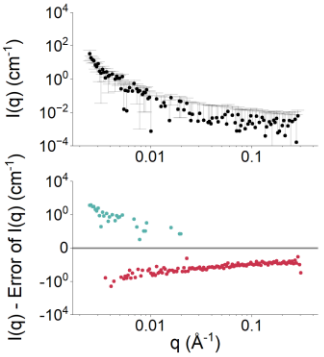
4. 2Nap-IF DMSO 5 mgmL^{-1}
Classification: Good Scattering



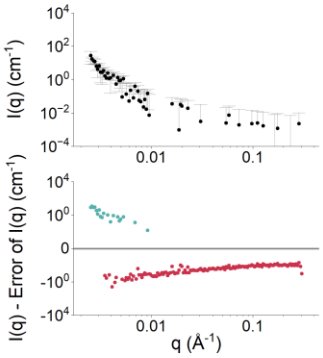
5. 6Br2Nap-VF DMSO 5 mgmL^{-1}
Classification: Bad Scattering



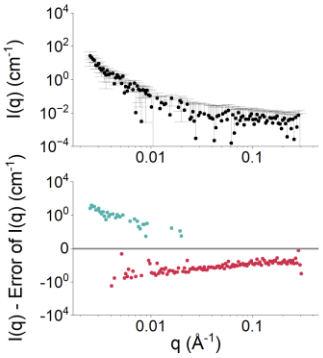
6. 6MeO2Nap-FI DMSO 5 mgmL^{-1}
Classification: Bad Scattering



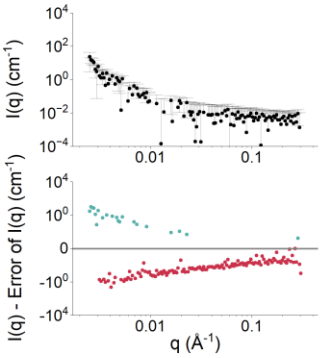
7. 7MeO2Nap-FV DMSO 5 mgmL^{-1}
Classification: Bad Scattering



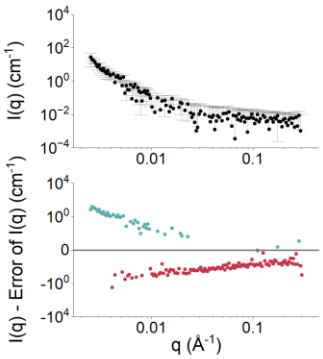
9. NDI-F DMSO 5 mgmL^{-1}
Classification: Bad Scattering



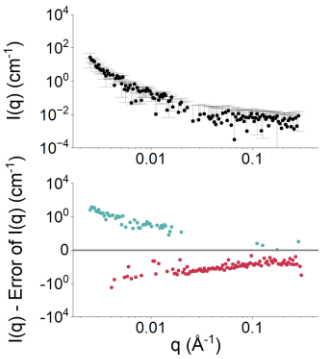
11. NDI-G DMSO 5 mgmL^{-1}
Classification: Bad Scattering



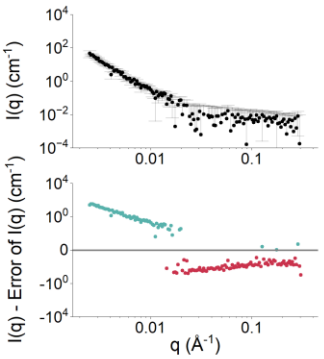
12. NDI-GF DMSO 5 mgmL^{-1}
Classification: Bad Scattering



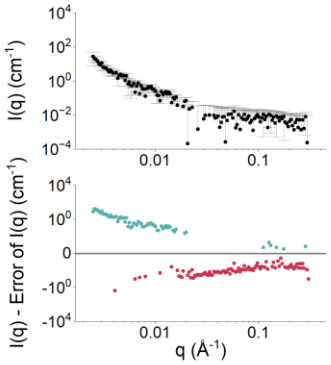
13. NDI-A DMSO 5 mgmL^{-1}
Classification: Bad Scattering



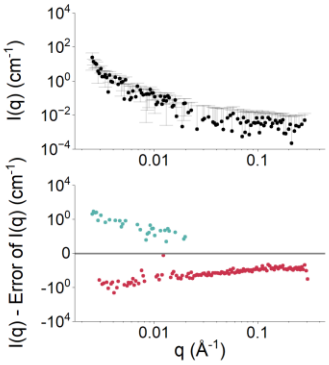
14. NDI-I DMSO 5 mgmL^{-1}
Classification: Bad Scattering



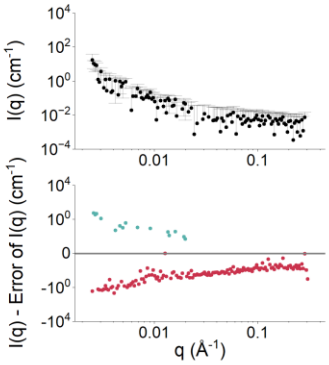
15. NDI-L DMSO 5 mgmL^{-1}
Classification: Bad Scattering



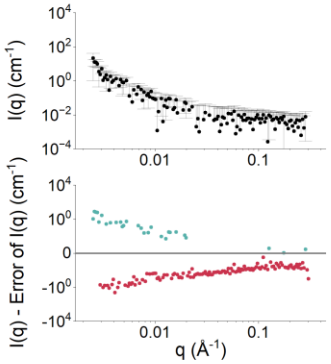
16. NDI-M DMSO 5 mgmL^{-1}
Classification: Bad Scattering



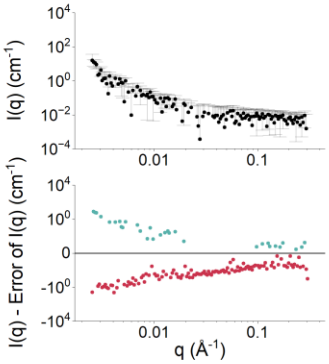
17. NDI-V DMSO 5 mgmL^{-1}
Classification: Bad Scattering



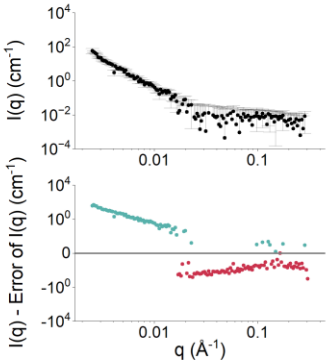
18. NDI-W DMSO 5 mgmL^{-1}
Classification: Bad Scattering



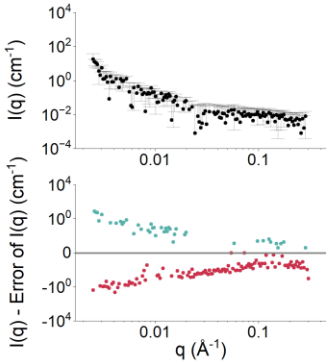
19. NDI-Y DMSO 5 mgmL^{-1}
Classification: Bad Scattering



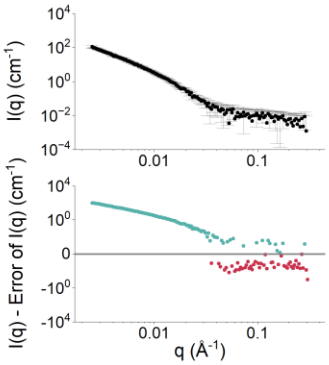
20. PBI-A DMSO 5 mgmL^{-1}
Classification: Bad Scattering



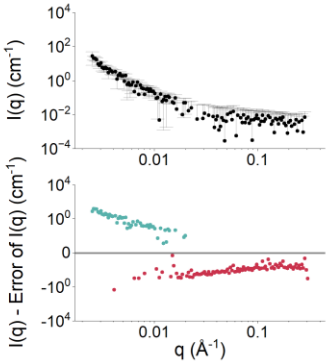
22. PBI-DOPA DMSO 5 mgmL^{-1}
Classification: Bad Scattering



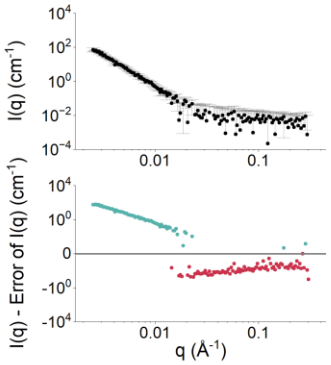
23. PBI-E DMSO 5 mgmL^{-1}
Classification: Bad Scattering



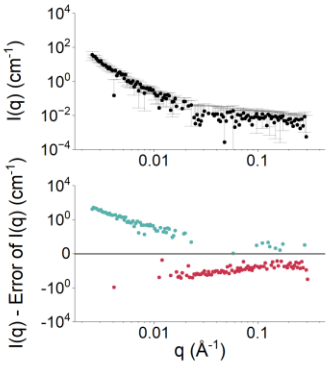
24. PBI-F DMSO 5 mgmL^{-1}
Classification: Bad Scattering



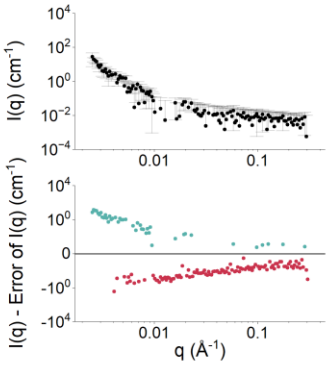
28. PBI-L DMSO 5 mgmL^{-1}
Classification: Bad Scattering



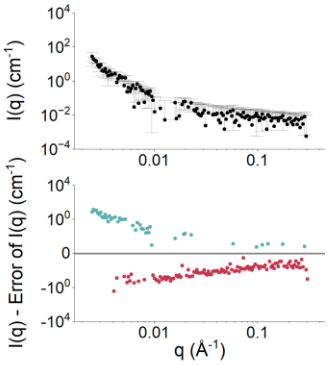
31. PBI-V DMSO 5 mgmL^{-1}
Classification: Bad Scattering



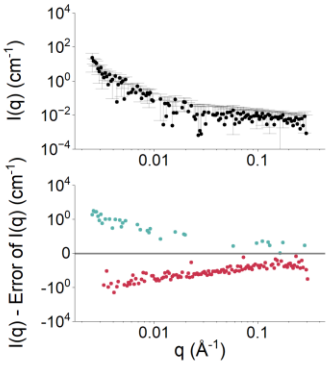
32. PBI-W DMSO 5 mgmL^{-1}
Classification: Bad Scattering



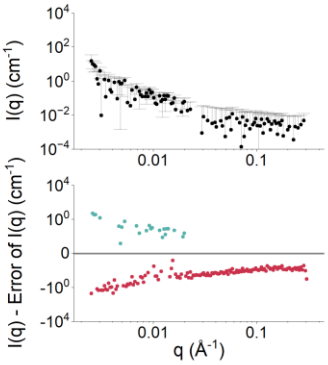
32. PBI-W DMSO 5 mgmL^{-1}
Classification: Bad Scattering



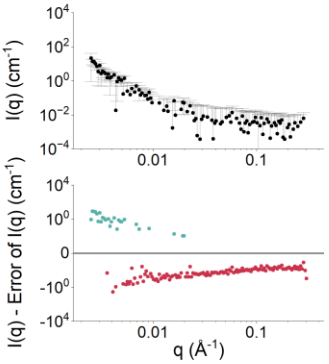
33. PBI-Y DMSO 5 mgmL^{-1}
Classification: Bad Scattering



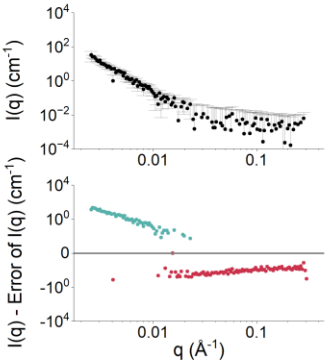
34. ThNap-FF DMSO 5 mgmL^{-1}
Classification: Bad Scattering



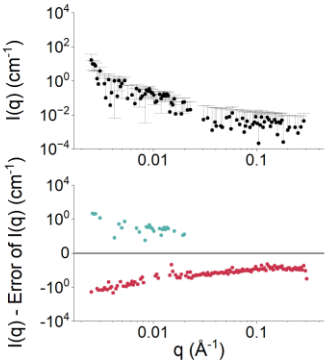
35. 2Nap-VL DMSO 5 mgmL^{-1}
Classification: Bad Scattering



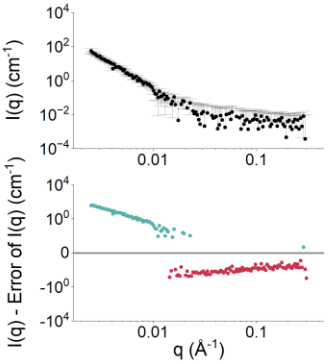
36. 2Nap-FL DMSO 5 mgmL^{-1}
Classification: Bad Scattering



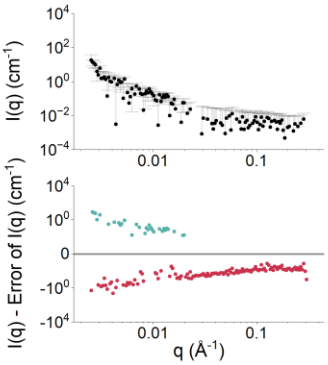
37. 2Nap-FI DMSO 5 mgmL^{-1}
Classification: Bad Scattering



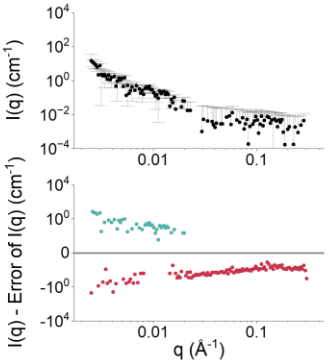
38. 2NapVF DMSO 5 mgmL^{-1}
Classification: Bad Scattering



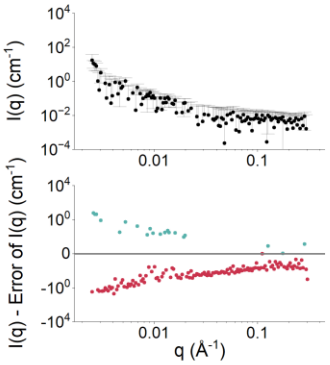
39. 2Nap-IV DMSO 5 mgmL^{-1}
Classification: Bad Scattering



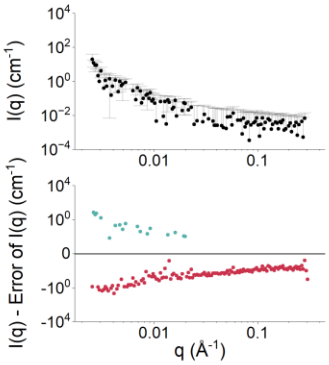
41. CF3-Ph-FV DMSO 5 mgmL^{-1}
Classification: Bad Scattering



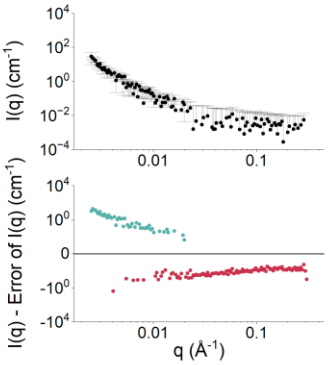
43. C6-NDI-A DMSO 5 mgmL^{-1}
Classification: Bad Scattering



44. C6-NDI-F DMSO 5 mgmL^{-1}
Classification: Bad Scattering

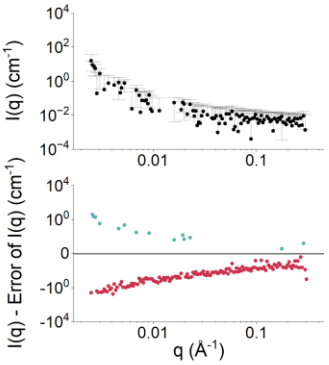


45. C6-NDI-L DMSO 5 mgmL^{-1}
Classification: Bad Scattering



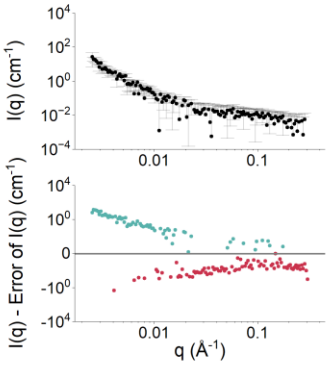
49. NDI-YI DMSO 5 mgmL^{-1}

Classification: Bad Scattering



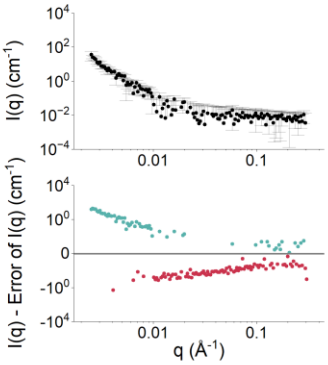
50. PBI-A 1Py DMSO 5 mgmL^{-1}

Classification: Bad Scattering



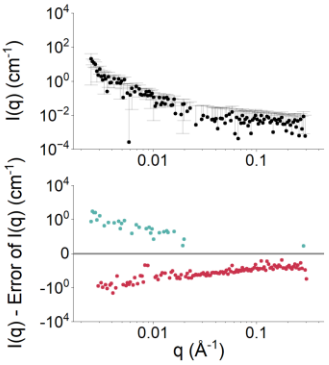
51. PBI-I 1Py DMSO 5 mgmL^{-1}

Classification: Bad Scattering



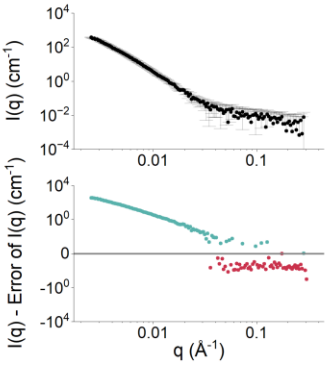
52. PBI-L1Py DMSO 5 mgmL^{-1}

Classification: Bad Scattering



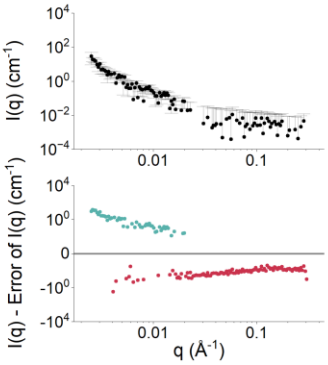
53. PBI-V 1Py DMSO 5 mgmL^{-1}

Classification: Bad Scattering



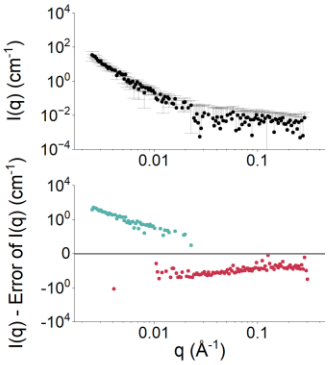
57. 1Nap-FL DMSO 5 mgmL^{-1}

Classification: Bad Scattering



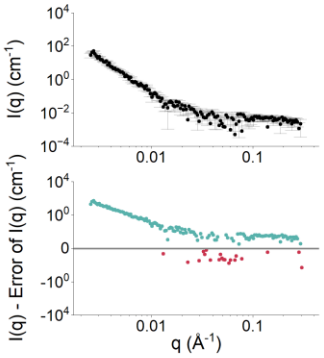
60. NDI-H DMSO 5 mgmL^{-1}

Classification: Bad Scattering

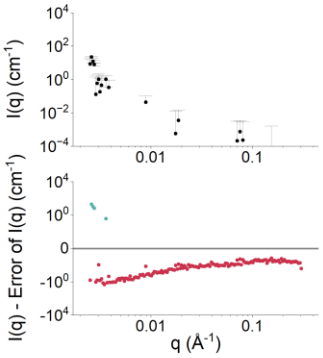


A 4.2.6 Ethanol Solution Classifications

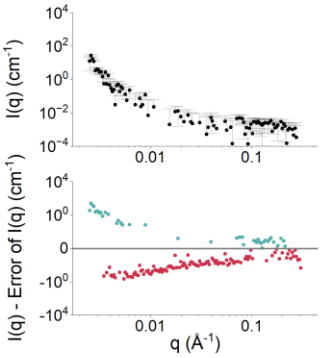
3. 2Nap-VVV Ethanol 5 mgmL^{-1}
Classification: Bad Scattering



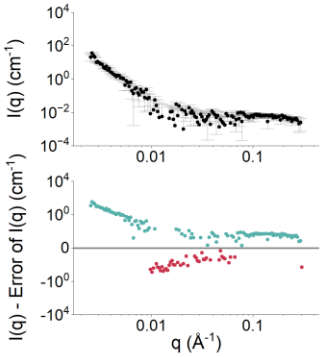
4. 2Nap-IF Ethanol 5 mgmL^{-1}
Classification: Bad Scattering



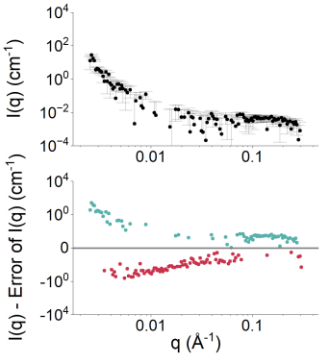
6. 6MeO2Nap-FI Ethanol 5 mgmL^{-1}
Classification: Bad Scattering



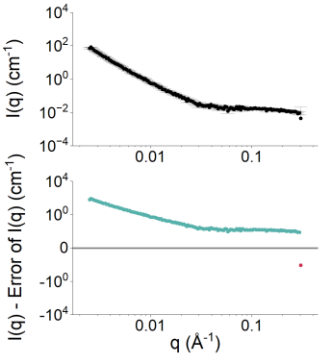
14. NDI-I Ethanol 5 mgmL^{-1}
Classification: Bad Scattering



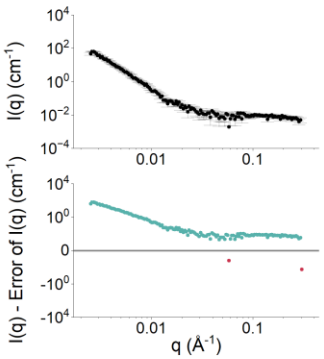
15. NDI-L Ethanol 5 mgmL^{-1}
Classification: Bad Scattering



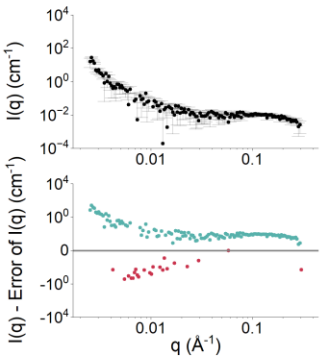
16. NDI-M Ethanol 5 mgmL^{-1}
Classification: Good Scattering



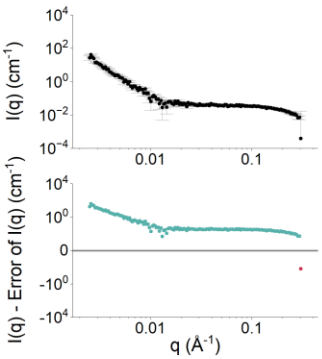
17. NDI-V Ethanol 5 mgmL^{-1}
Classification: Good Scattering



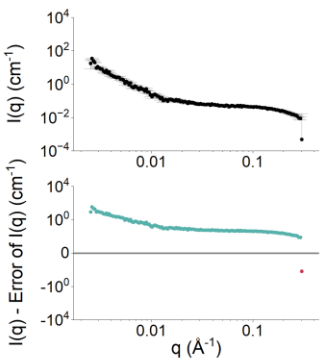
19. NDI-Y Ethanol 5 mgmL^{-1}
Classification: Bad Scattering



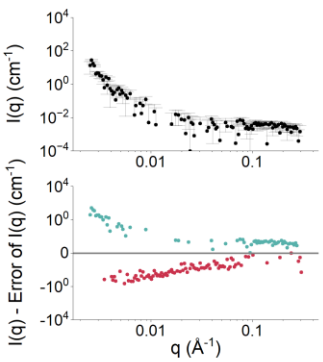
22. PBI-DOPA Ethanol 5 mgmL^{-1}
Classification: Good Scattering



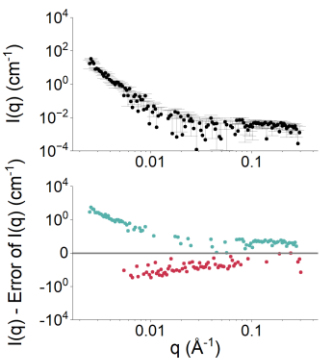
33. PBI-Y Ethanol 5 mgmL^{-1}
Classification: Good Scattering



35. 2Nap-VL Ethanol 5 mgmL^{-1}
Classification: Bad Scattering

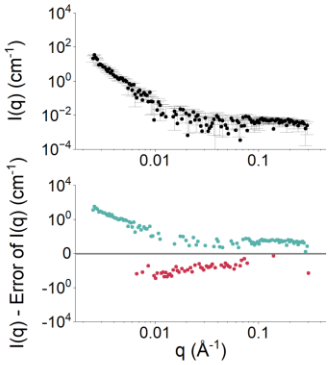


36. 2Nap-FL Ethanol 5 mgmL^{-1}
Classification: Bad Scattering



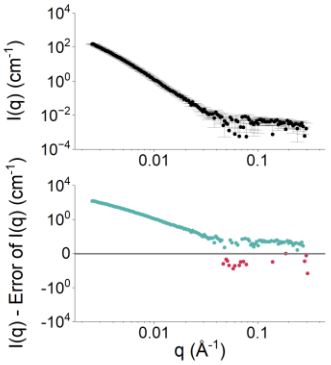
37. 2Nap-FI Ethanol 5 mgmL^{-1}

Classification: Bad Scattering



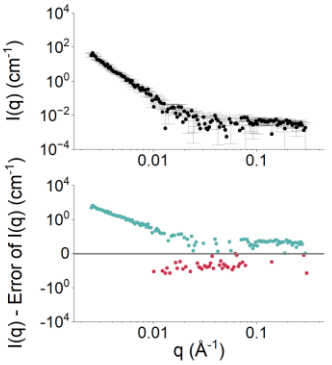
38. 2Nap-VF Ethanol 5 mgmL^{-1}

Classification: Bad Scattering



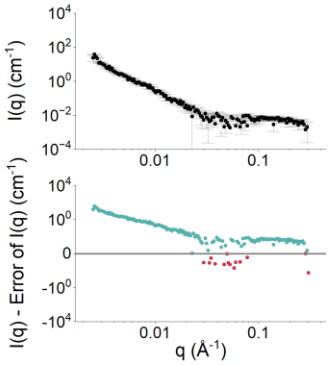
41. CF3Ph-FV Ethanol 5 mgmL^{-1}

Classification: Bad Scattering



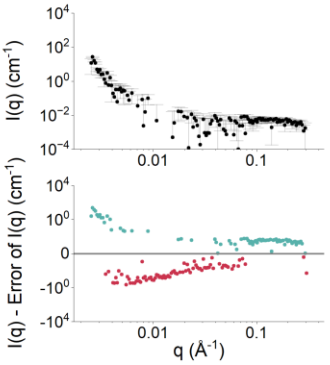
44. C6-NDI-F Ethanol 5 mgmL^{-1}

Classification: Bad Scattering



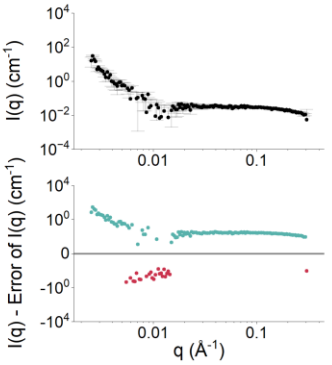
45. C6-NDI-L Ethanol 5 mgmL^{-1}

Classification: Bad Scattering



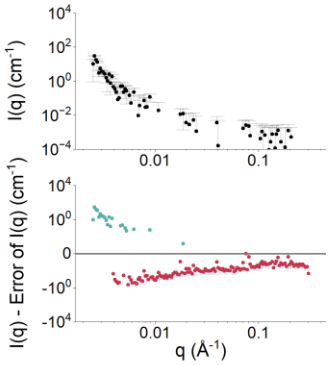
51. PBI-I 1Py Ethanol 5 mgmL^{-1}

Classification: Bad Scattering



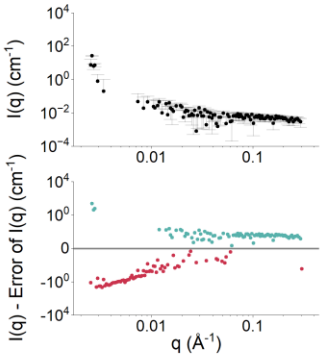
57. 1Nap-FL Ethanol 5 mgmL^{-1}

Classification: Bad Scattering

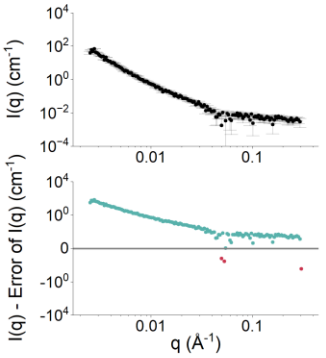


A 4.2.7 THF Solution Classifications

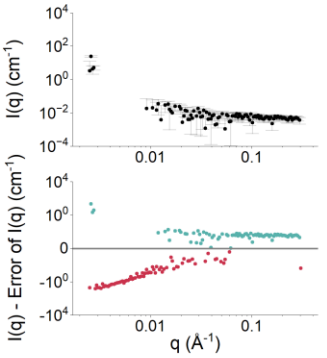
1. 2Nap-FF THF 5 mgmL^{-1}
Classification: Bad Scattering



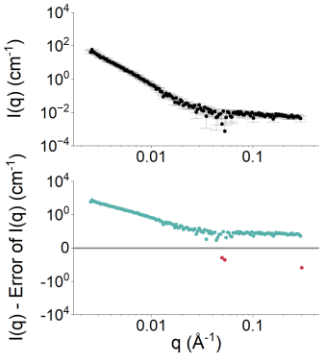
3. 2Nap-VVV THF 5 mgmL^{-1}
Classification: Good Scattering



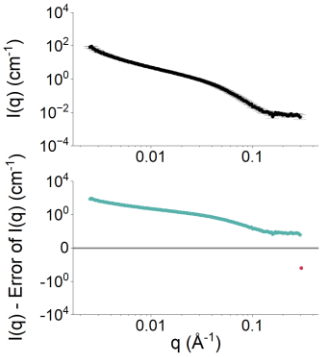
4. 2Nap-IF THF 5 mgmL^{-1}
Classification: Bad Scattering



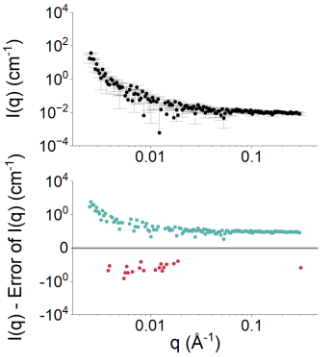
5. 6Br2Nap-VF THF 5 mgmL^{-1}
Classification: Good Scattering



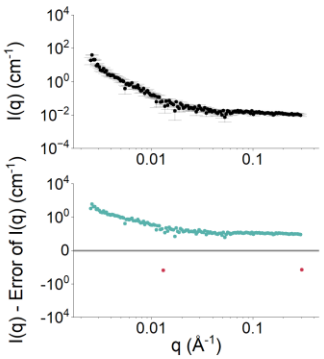
6. 6MeO2Nap-FI THF 5 mgmL^{-1}
Classification: Good Scattering



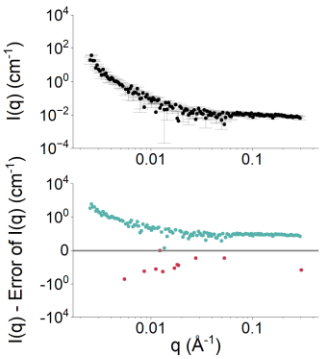
7. 7MeO2Nap-FV THF 5 mgmL^{-1}
Classification: Bad Scattering



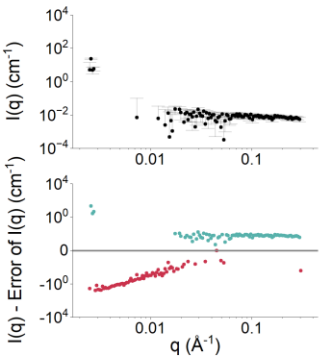
9. NDI-F THF 5 mgmL^{-1}
Classification: Good Scattering



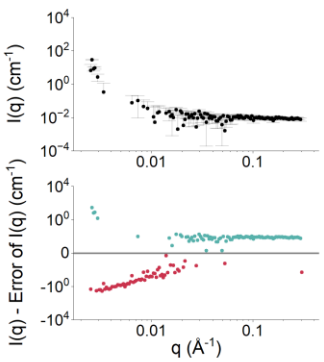
13. NDI-A THF 5 mgmL^{-1}
Classification: Bad Scattering



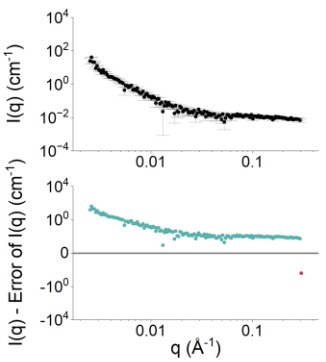
14. NDI-I THF 5 mgmL^{-1}
Classification: Bad Scattering



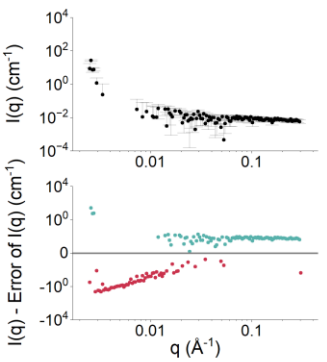
15. NDI-L THF 5 mgmL^{-1}
Classification: Bad Scattering



16. NDI-M THF 5 mgmL^{-1}
Classification: Good Scattering

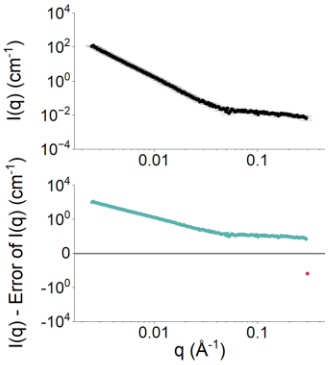


17. NDI-V THF 5 mgmL^{-1}
Classification: Bad Scattering



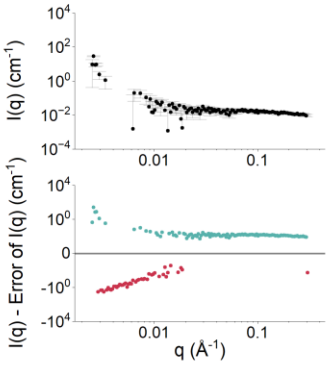
18. NDI-W THF 5 mgmL^{-1}

Classification: Good Scattering



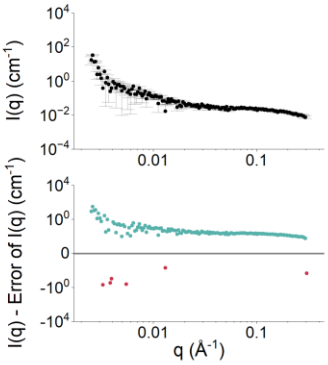
19. NDI-Y THF 5 mgmL^{-1}

Classification: Bad Scattering



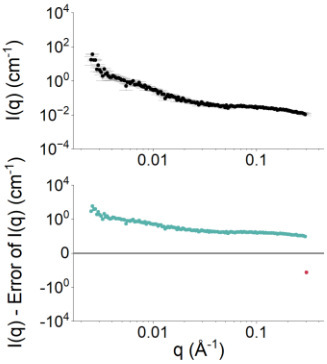
22. PBI-DOPA THF 5 mgmL^{-1}

Classification: Good Scattering



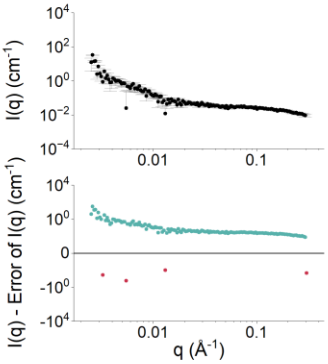
32. PBI-W THF 5 mgmL^{-1}

Classification: Good Scattering



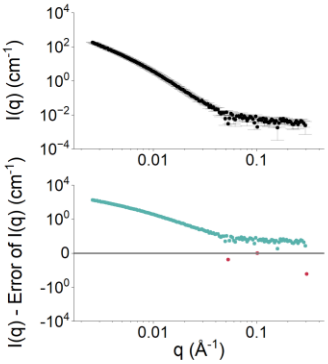
33. PBI-Y THF 5 mgmL^{-1}

Classification: Good Scattering



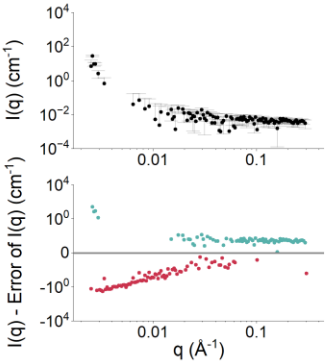
34. ThNapFF THF 5 mgmL^{-1}

Classification: Good Scattering



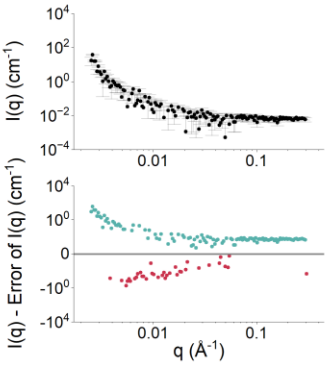
35. 2Nap-VL THF 5 mgmL^{-1}

Classification: Bad Scattering



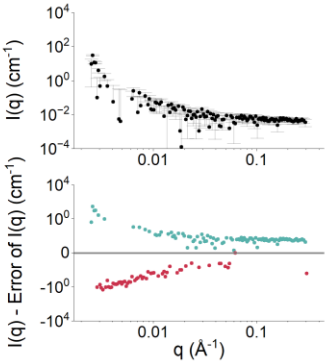
36. 2Nap-FL THF 5 mgmL^{-1}

Classification: Bad Scattering



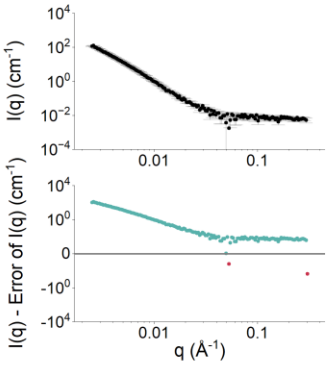
37. 2Nap-FI THF 5 mgmL^{-1}

Classification: Bad Scattering



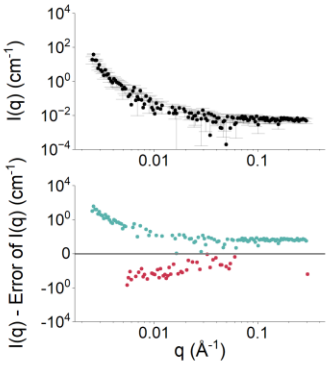
38. 2Nap-VF THF 5 mgmL^{-1}

Classification: Good Scattering



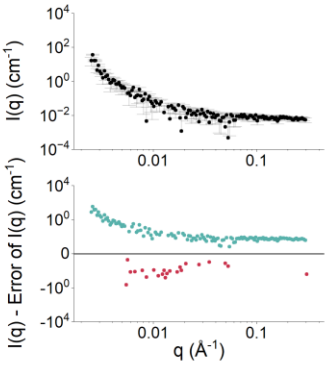
39. 2Nap-IV THF 5 mgmL^{-1}

Classification: Bad Scattering



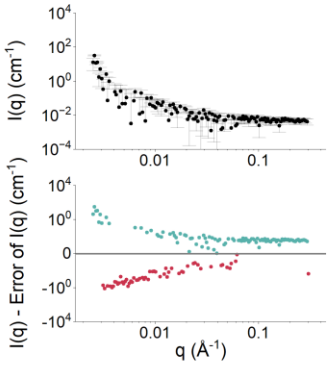
41. CF3Ph-FV THF 5 mgmL^{-1}

Classification: Bad Scattering



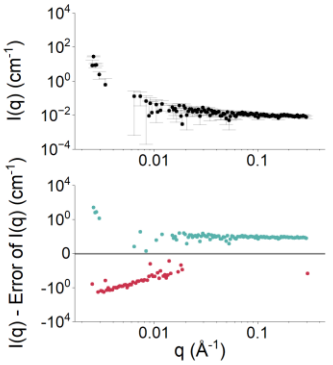
42. 2Nap-VI THF 5 mgmL^{-1}

Classification: Bad Scattering



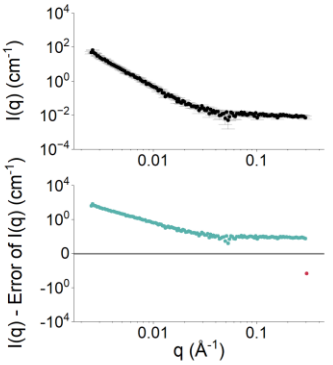
43. C6-NDI-A THF 5 mgmL^{-1}

Classification: Bad Scattering



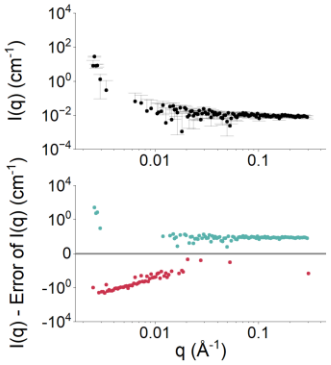
44. C6-NDI-F THF 5 mgmL^{-1}

Classification: Good Scattering



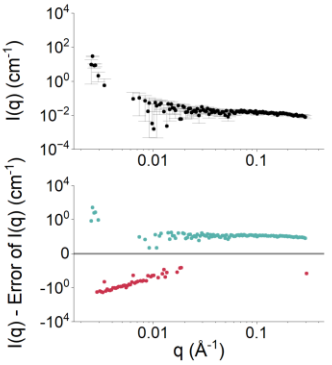
45. C6-NDI-L THF 5 mgmL^{-1}

Classification: Bad Scattering



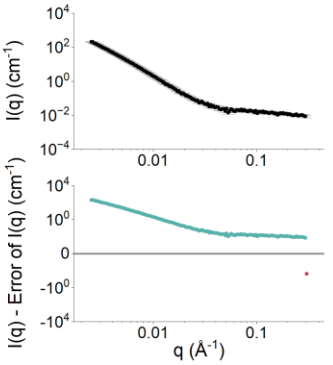
49. NDI-YI THF 5 mgmL^{-1}

Classification: Bad Scattering



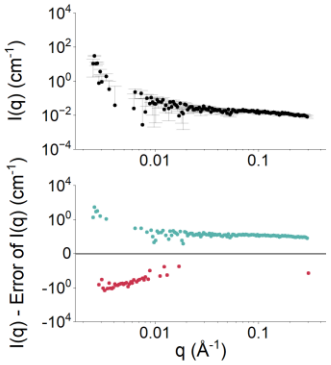
50. PBI-A 1Py THF 5 mgmL^{-1}

Classification: Good Scattering



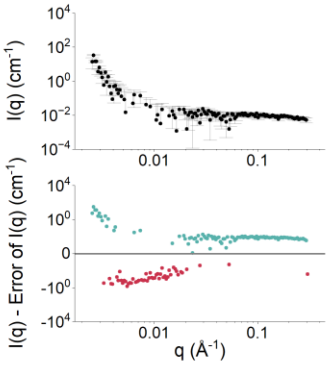
51. PBI-I 1Py THF 5 mgmL^{-1}

Classification: Bad Scattering



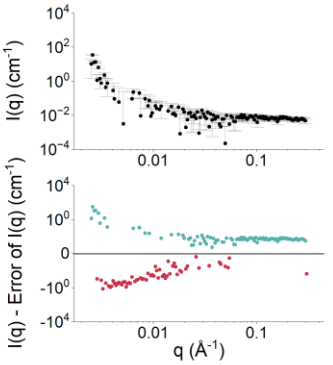
52. PBI-L1Py THF 5 mgmL^{-1}

Classification: Bad Scattering



57. 1Nap-FL THF 5 mgmL^{-1}

Classification: Bad Scattering

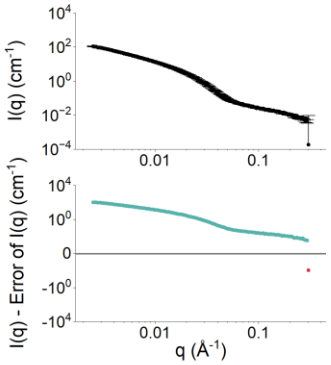


A 4.2.8 Deuterated Solution Classifications

A 4.2.8.1 5 mgmL^{-1} Aqueous Solution Classifications

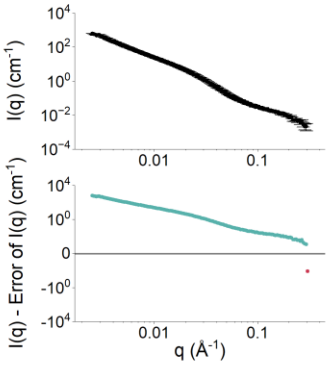
20. PBI-A D_2O 5 mgmL^{-1}

Classification: Good Scattering



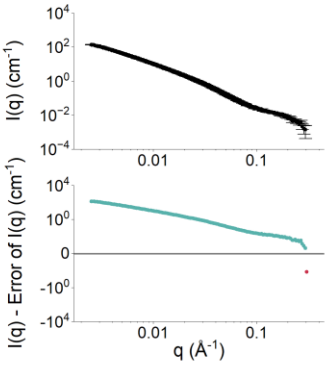
21. PBI-D D_2O 5 mgmL^{-1}

Classification: Good Scattering



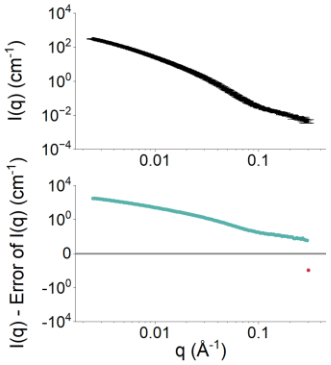
22. PBI-DOPA D_2O 5 mgmL^{-1}

Classification: Good Scattering



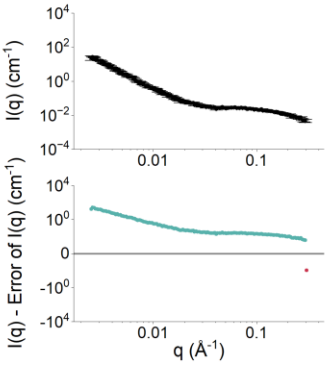
23. PBI-E D_2O 5 mgmL^{-1}

Classification: Good Scattering



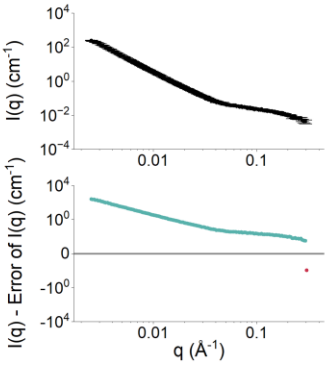
24. PBI-F D_2O 5 mgmL^{-1}

Classification: Good Scattering



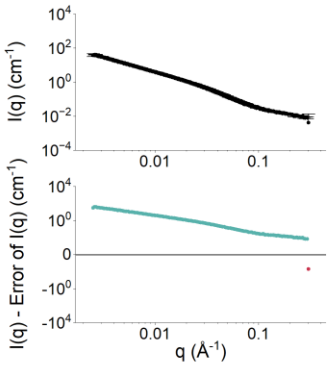
25. PBI-G D_2O 5 mgmL^{-1}

Classification: Good Scattering



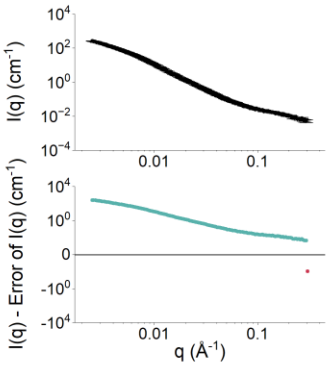
27. PBI-I D_2O 5 mgmL^{-1}

Classification: Good Scattering



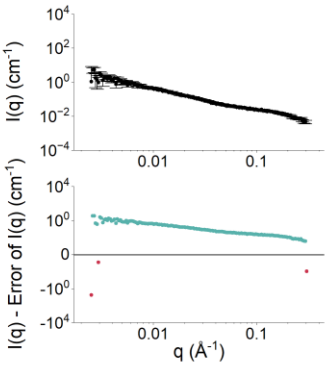
28. PBI-L D_2O 5 mgmL^{-1}

Classification: Good Scattering



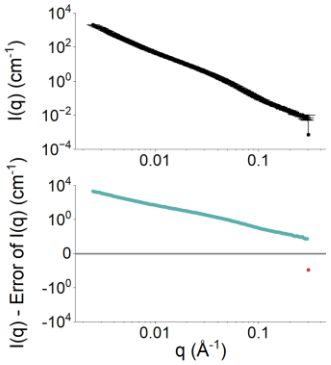
29. PBI-M D_2O 5 mgmL^{-1}

Classification: Good Scattering



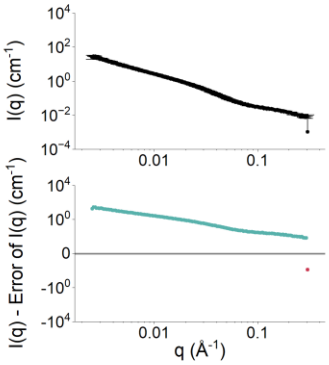
30. PBI-S D_2O 5 $mgmL^{-1}$

Classification: Good Scattering



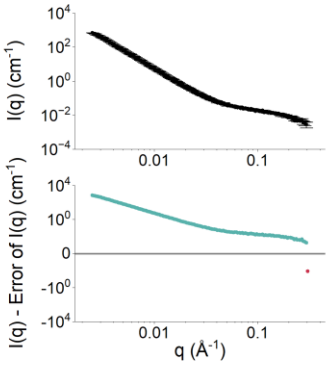
31. PBI-V D_2O 5 $mgmL^{-1}$

Classification: Good Scattering



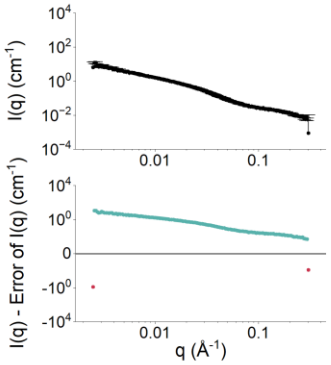
32. PBI-W D_2O 5 $mgmL^{-1}$

Classification: Good Scattering



33. PBI-Y D_2O 5 $mgmL^{-1}$

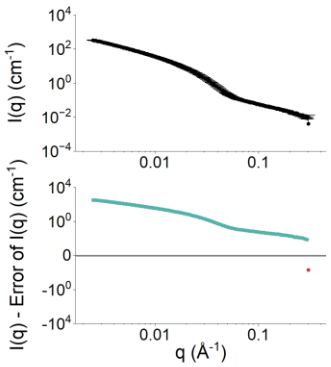
Classification: Good Scattering



A 4.2.8.2 10 mgmL^{-1} Aqueous Solution Classifications

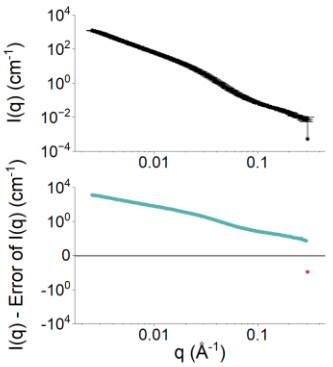
20. PBI-A D_2O 10 mgmL^{-1}

Classification: Good Scattering



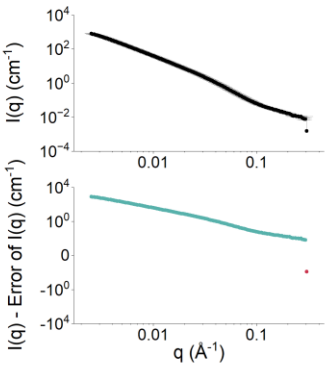
21. PBI-D D_2O 10 mgmL^{-1}

Classification: Good Scattering



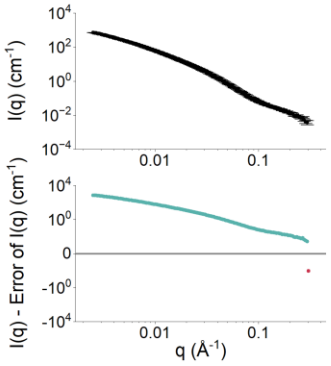
22. PBI-DOPA D_2O 10 mgmL^{-1}

Classification: Good Scattering



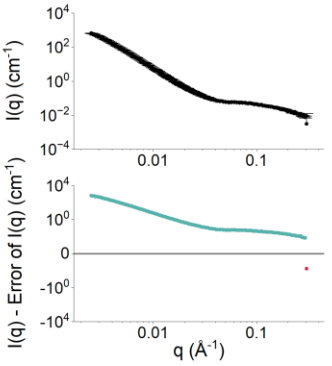
23. PBI-E D_2O 10 mgmL^{-1}

Classification: Good Scattering



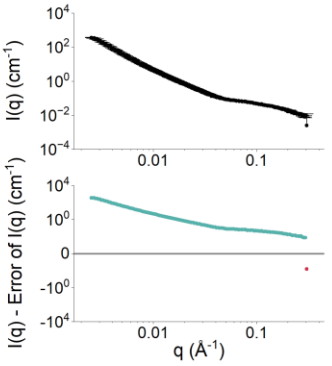
24. PBI-F D_2O 10 mgmL^{-1}

Classification: Good Scattering



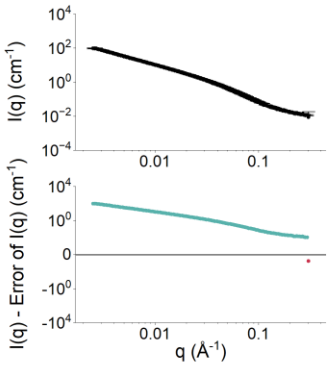
25. PBI-G D_2O 10 mgmL^{-1}

Classification: Good Scattering



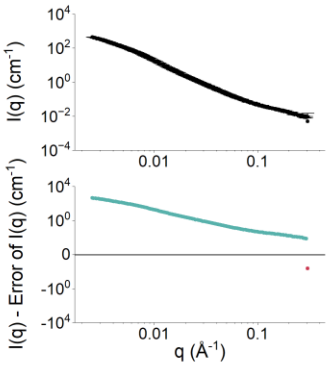
27. PBI-I D_2O 10 mgmL^{-1}

Classification: Good Scattering



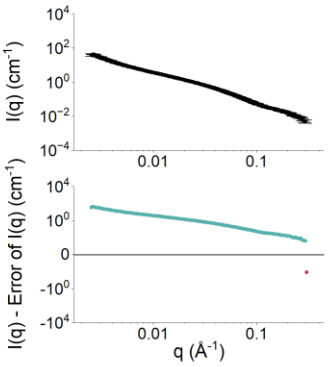
28. PBI-L D_2O 10 mgmL^{-1}

Classification: Good Scattering



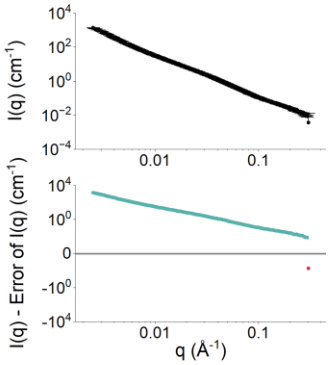
29. PBI-M D_2O 10 mgmL^{-1}

Classification: Good Scattering



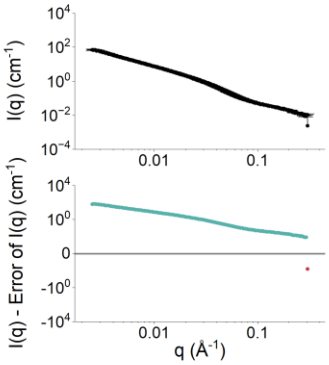
30. PBI-S D_2O 10 $mgmL^{-1}$

Classification: Good Scattering



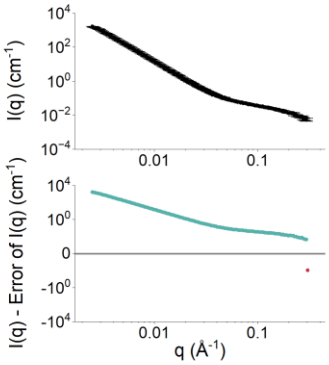
31. PBI-V D_2O 10 $mgmL^{-1}$

Classification: Good Scattering



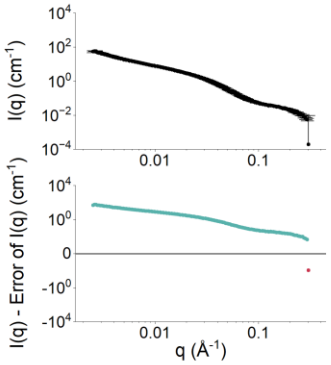
32. PBI-W D_2O 10 $mgmL^{-1}$

Classification: Good Scattering



33. PBI-Y D_2O 10 $mgmL^{-1}$

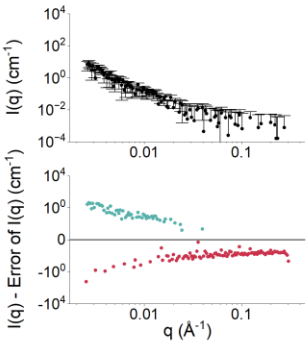
Classification: Good Scattering



A 4.2.8.3 d_6-DMSO Solution Classifications

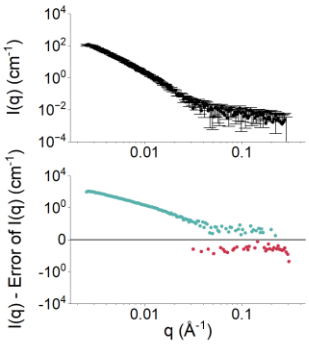
4. 2Nap-IF d_6-DMSO 5 mgmL^{-1}

Classification: Bad Scattering



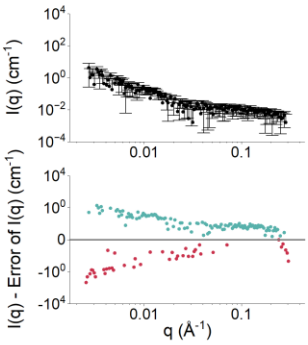
20. PBI-A d_6-DMSO 5 mgmL^{-1}

Classification: Bad Scattering



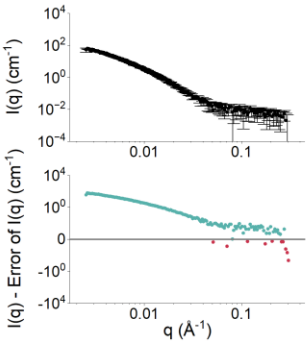
22. PBI-DOPA d_6-DMSO 5 mgmL^{-1}

Classification: Bad Scattering



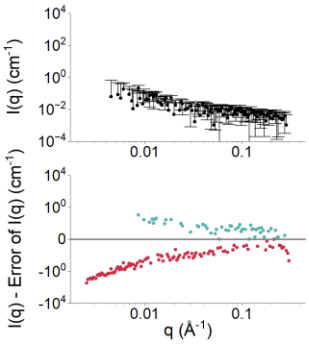
23. PBI-E d_6-DMSO 5 mgmL^{-1}

Classification: Bad Scattering



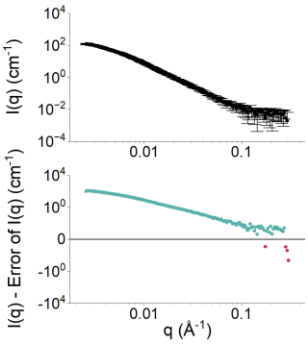
24. PBI-F d_6-DMSO 5 mgmL^{-1}

Classification: Bad Scattering



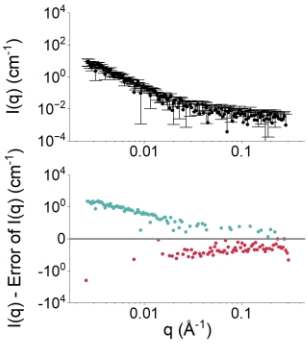
28. PBI-L d_6-DMSO 5 mgmL^{-1}

Classification: Good Scattering



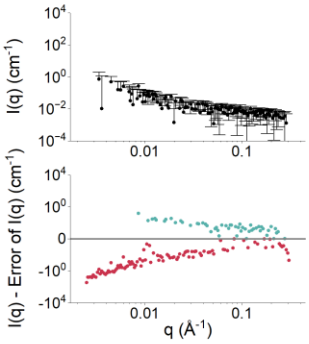
31. PBI-V d_6-DMSO 5 mgmL^{-1}

Classification: Bad Scattering



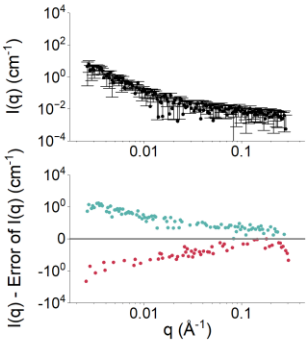
32. PBI-W d_6-DMSO 5 mgmL^{-1}

Classification: Bad Scattering



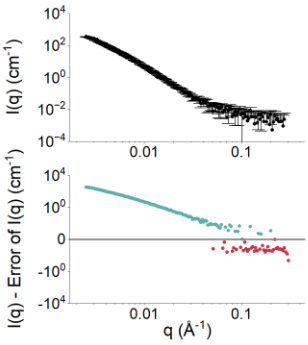
33. PBI-Y d_6-DMSO 5 mgmL^{-1}

Classification: Bad Scattering



53. PBI-V 1Py d_6-DMSO 5 mgmL^{-1}

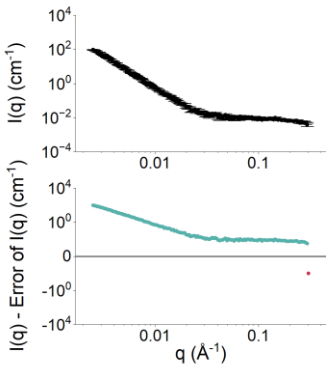
Classification: Bad Scattering



A 4.2.8.4 d_6-Ethanol Solution Classifications

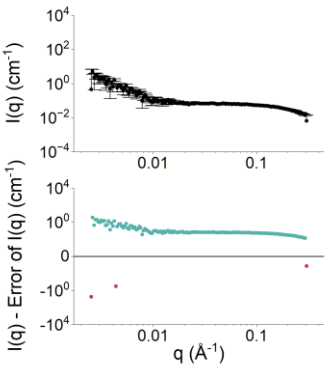
17. NDI-V d_6-Ethanol 5 mgmL^{-1}

Classification: Good Scattering



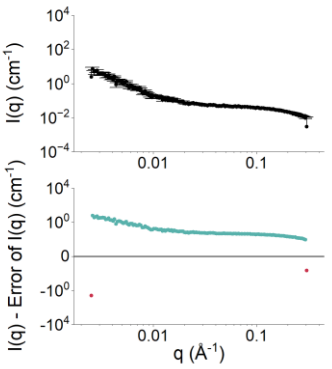
22. PBI-DOPA d_6-Ethanol 5 mgmL^{-1}

Classification: Good Scattering



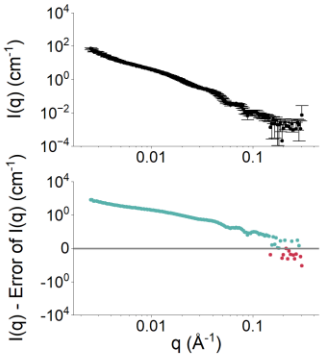
33. PBI-Y d_6-Ethanol 5 mgmL^{-1}

Classification: Good Scattering

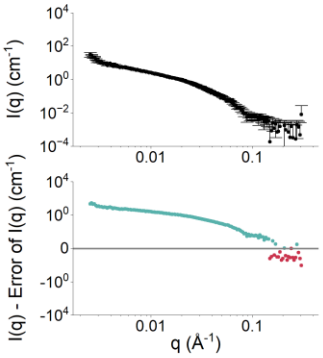


A 4.2.9 Multicomponent Solution Classifications

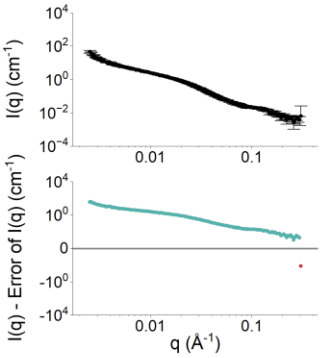
2Nap-FI (37) + 2Nap-FF (1)
Classification: Good Scattering



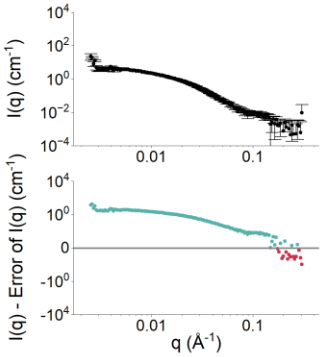
2Nap-IF (4) + 2Nap-FF (1)
Classification: Good Scattering



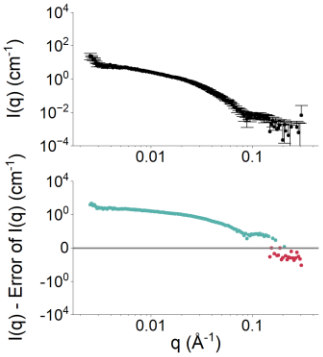
2Nap-V (2) + 2Nap-FF (1)
Classification: Good Scattering



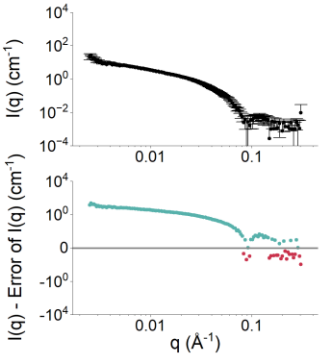
6MeO2Nap-FI (6) + 2Nap-FF (1)
Classification: Good Scattering



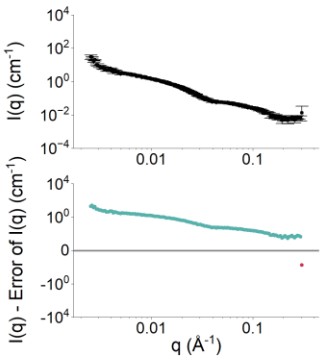
7MeO2Nap-FV (7) + 2Nap-FF (1)
Classification: Good Scattering



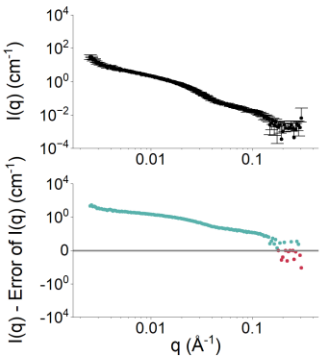
C6-NDI-A (43) + 2Nap-FF (1)
Classification: Good Scattering



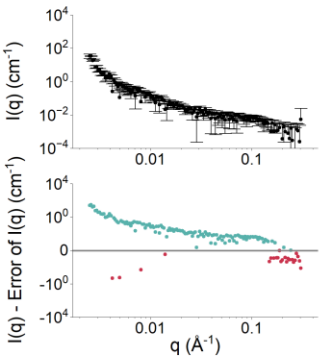
C6-NDI-F (44) + 2Nap-FF (1)
Classification: Good Scattering



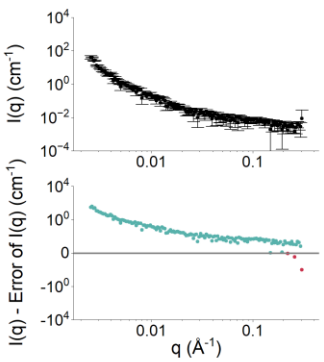
C6-NDI-L (45) + 2Nap-FF (1)
Classification: Good Scattering



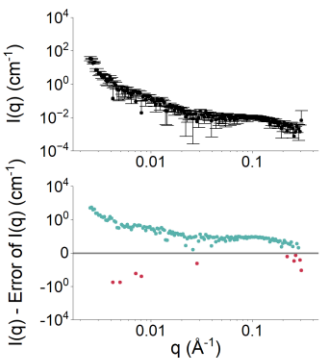
C6-NDI-L (45) + 2Nap-FI (37)
Classification: Bad Scattering



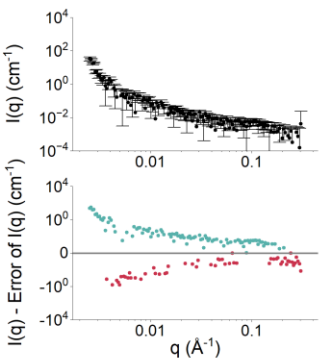
C6-NDI-L (45) + 2Nap-IF (4)
Classification: Good Scattering



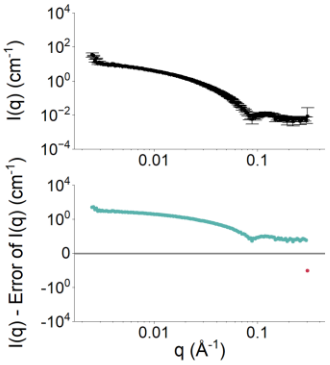
C6-NDI-L (45) + 2Nap-V (2)
Classification: Bad Scattering



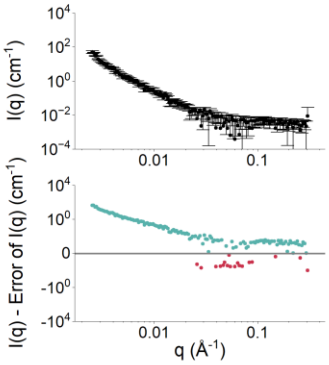
C6-NDI-L (45) + 6MeO2Nap-FI (6)
Classification: Bad Scattering



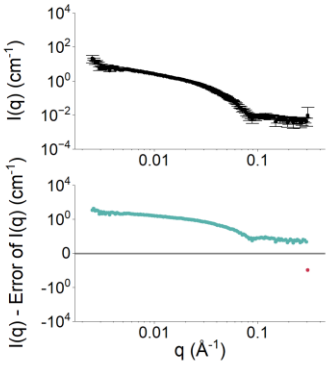
NDI-A (13) + 2Nap-FF (1)
Classification: Good Scattering



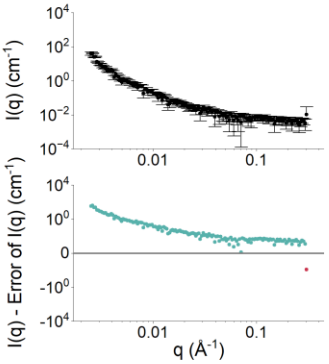
NDI-A (13) + C6-NDI-L (45)
Classification: Bad Scattering



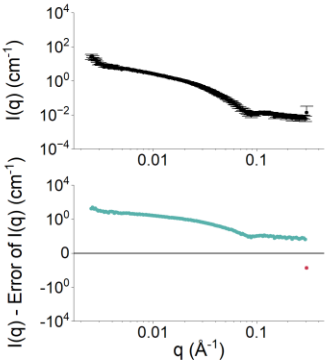
NDI-F (9) + 2Nap-FF (1)
Classification: Good Scattering



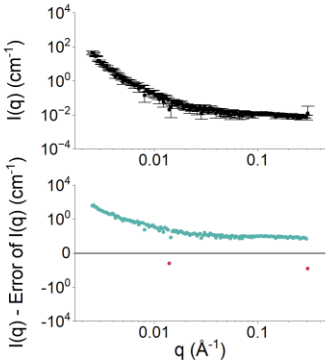
NDI-F (9) + C6-NDI-L (45)
Classification: Good Scattering



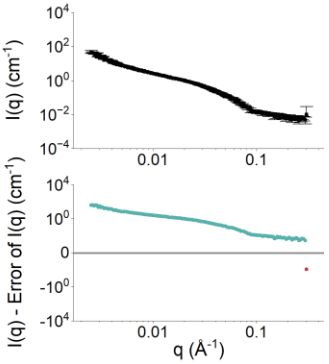
NDI-G (11) + 2Nap-FF (1)
Classification: Good Scattering



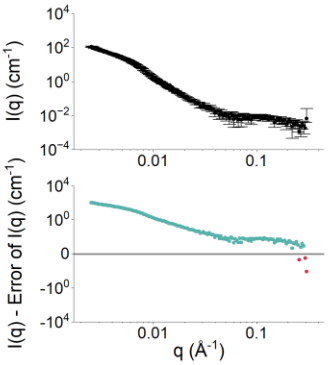
NDI-G (11) + C6-NDI-L (45)
Classification: Good Scattering



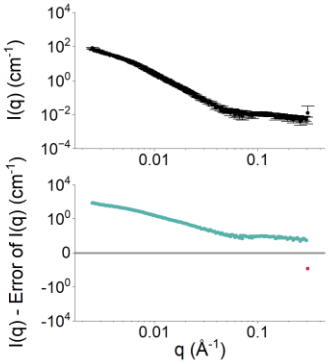
NDI-GF (12) + 2Nap-FF (1)
Classification: Good Scattering



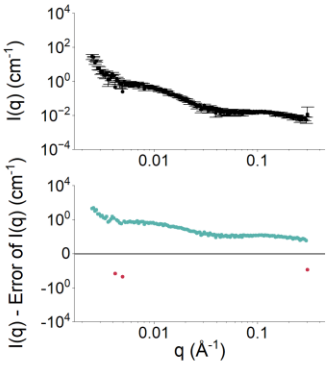
NDI-GF (12) + 2Nap-FI (37)
Classification: Good Scattering



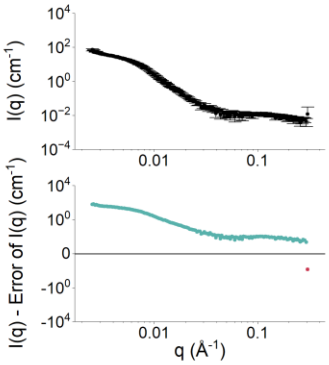
NDI-GF (12) + 2Nap-IF (4)
Classification: Good Scattering



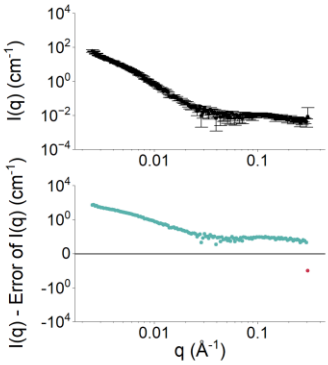
NDI-GF (12) + 2Nap-V (2)
Classification: Good Scattering



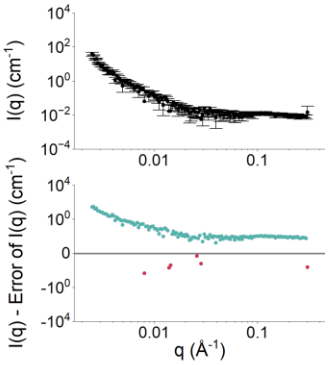
NDI-GF (12) + 6MeO2Nap-FI (6)
Classification: Good Scattering



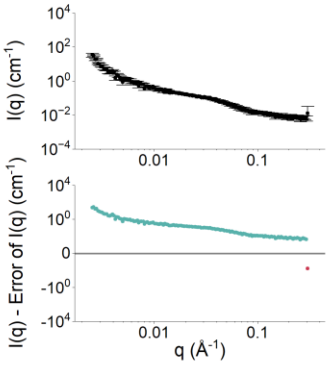
NDI-GF (12) + 7MeO2Nap-FV (7)
Classification: Good Scattering



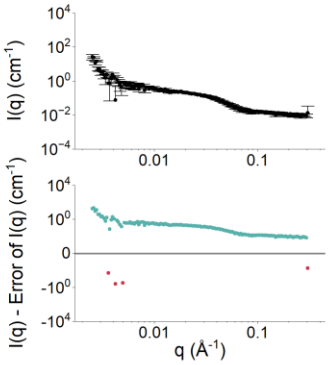
NDI-GF (12) + C6-NDI-A (43)
Classification: Good Scattering



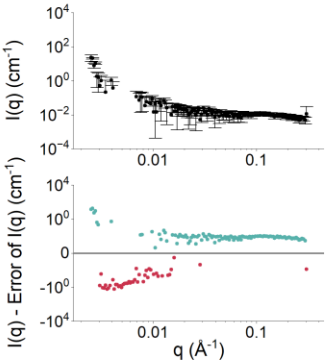
NDI-GF (12) + C6-NDI-F (44)
Classification: Good Scattering



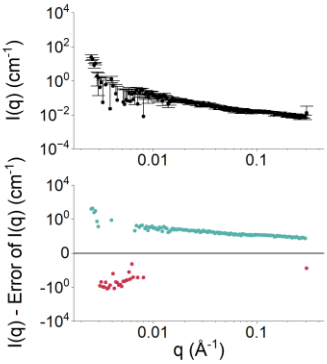
NDI-GF (12) + C6-NDI-L (45)
Classification: Good Scattering



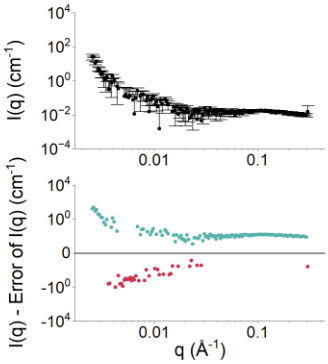
NDI-GF (12) + NDI-A (13)
Classification: Bad Scattering



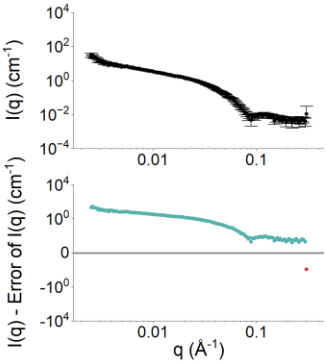
NDI-GF (12) + NDI-F (9)
Classification: Bad Scattering



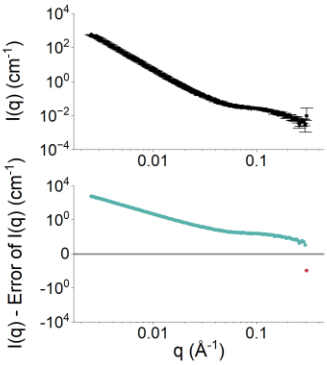
NDI-GF (12) + NDI-G (11)
Classification: Bad Scattering



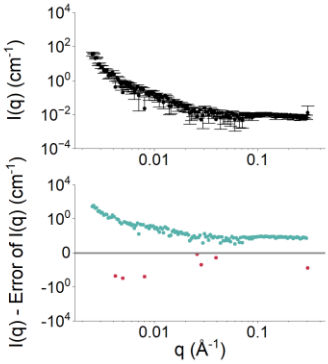
NDI-H (60) + 2Nap-FF (1)
Classification: Good Scattering



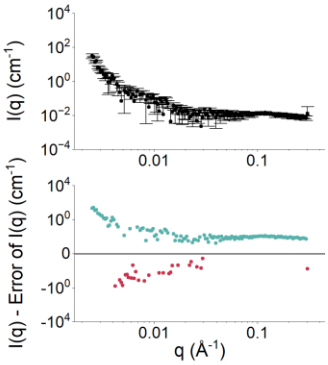
NDI-H (60) + C6-NDI-F (44)
Classification: Good Scattering



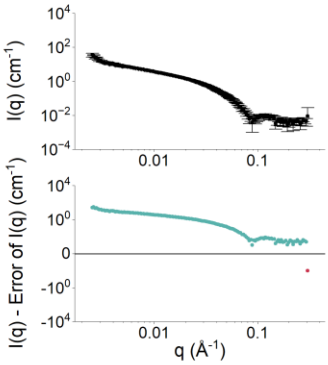
NDI-H (60) + C6-NDI-L (45)
Classification: Good Scattering



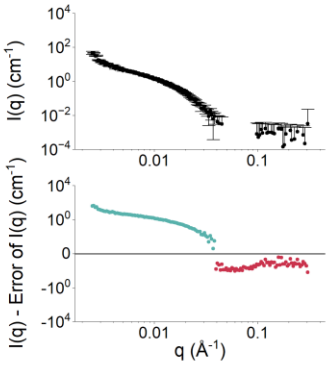
NDI-H (60) + NDI-GF (12)
Classification: Bad Scattering



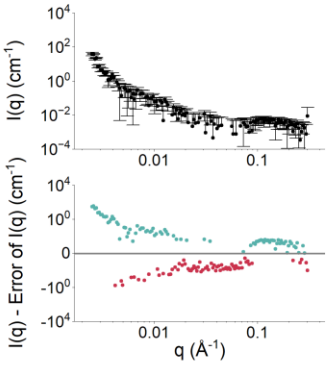
NDI-I (14) + 2Nap-FF (1)
Classification: Good Scattering



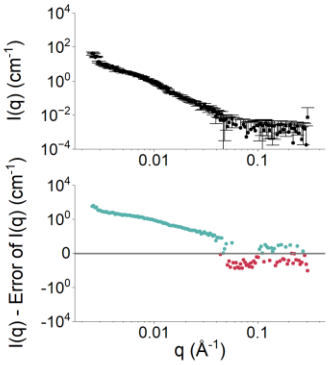
NDI-I (14) + 2Nap-IF (4)
Classification: Bad Scattering



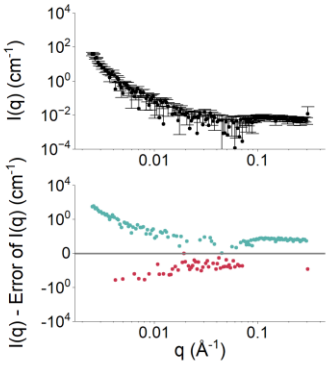
NDI-I (14) + 2Nap-V (2)
Classification: Bad Scattering



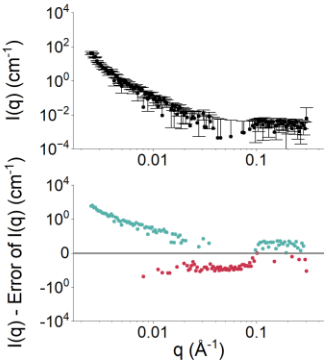
NDI-I (14) + 6MeO2Nap-FI (6)
Classification: Bad Scattering



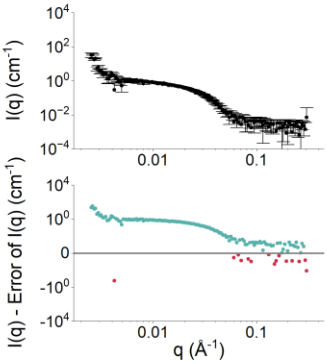
NDI-I (14) + 7MeO2Nap-FV (7)
Classification: Bad Scattering



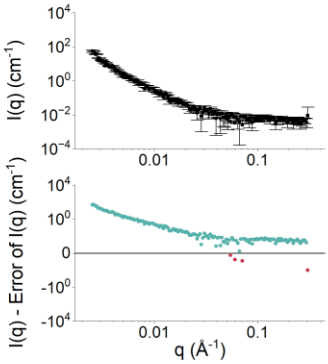
NDI-I (14) + C6-NDI-A (43)
Classification: Bad Scattering



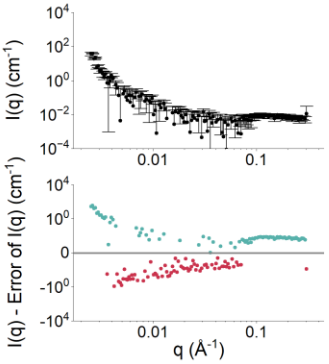
NDI-I (14) + C6-NDI-F (44)
Classification: Bad Scattering



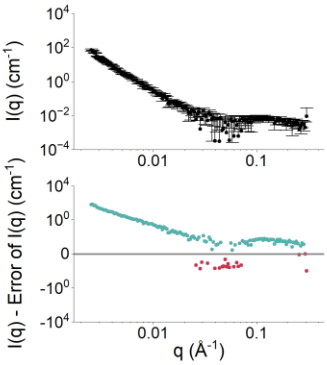
NDI-I (14) + C6-NDI-L (45)
Classification: Good Scattering



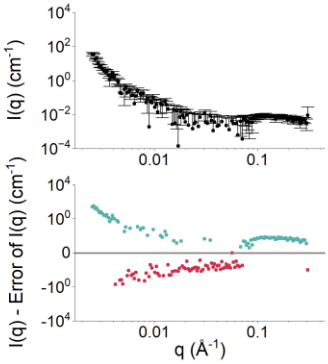
NDI-I (14) + NDI-A (13)
Classification: Bad Scattering



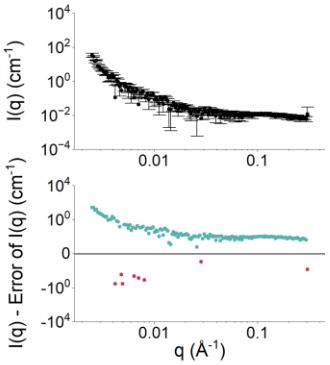
NDI-I (14) + NDI-F (9)
Classification: Bad Scattering



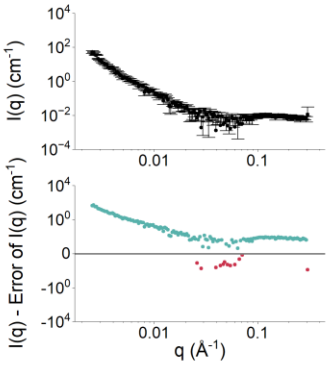
NDI-I (14) + NDI-G (11)
Classification: Bad Scattering



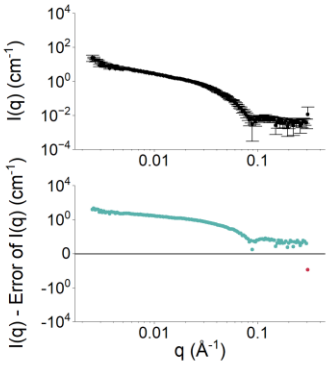
NDI-I (14) + NDI-GF (12)
Classification: Good Scattering

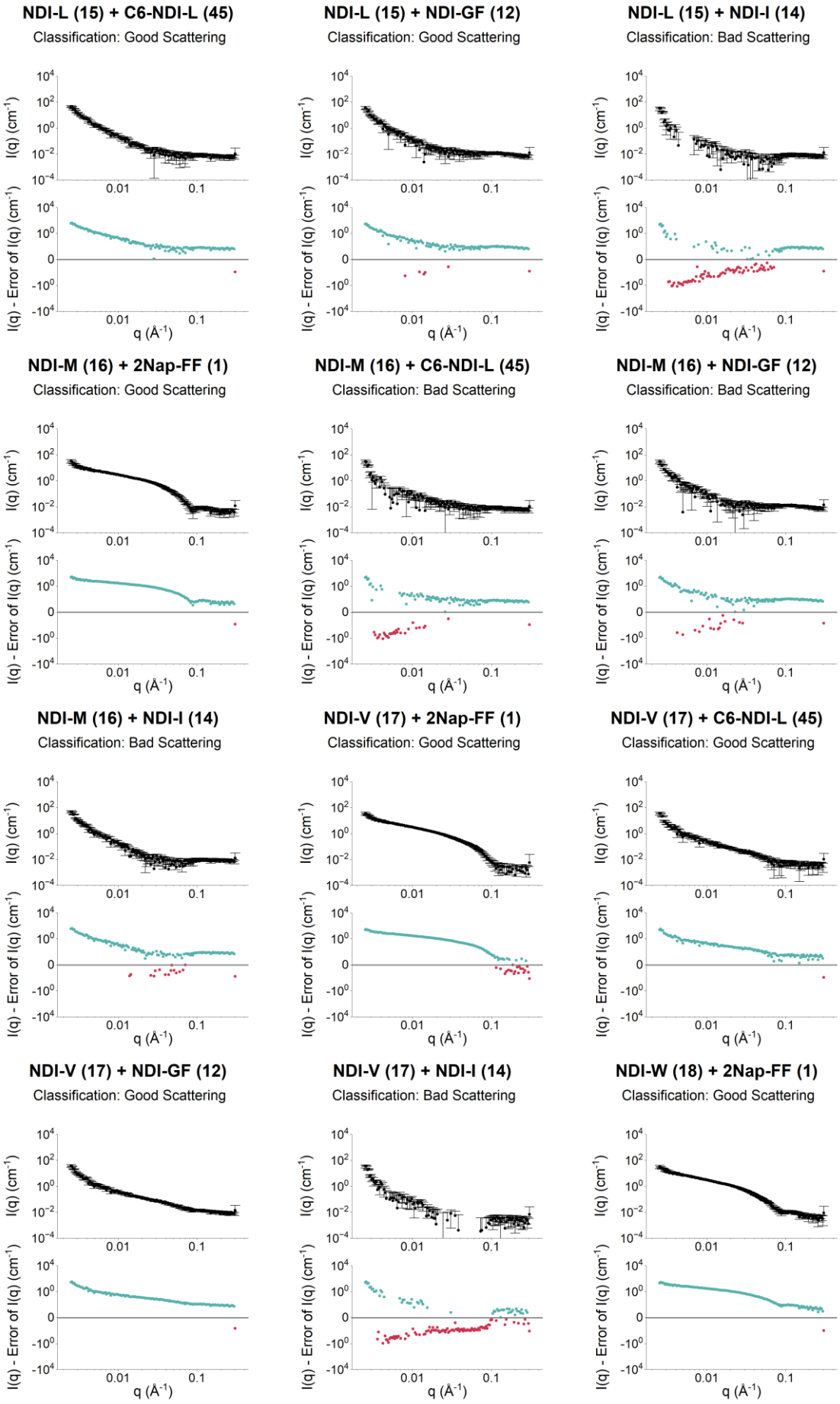


NDI-I (14) + NDI-H (60)
Classification: Bad Scattering

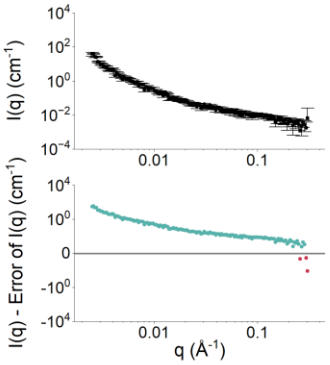


NDL-L (15) + 2Nap-FF (1)
Classification: Good Scattering

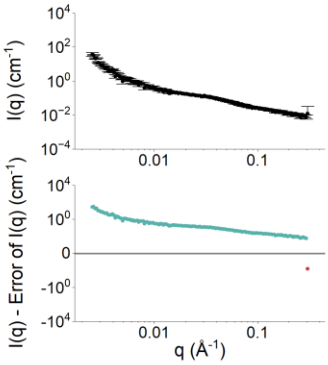



NDI-L (15) + C6-NDI-L (45)
Classification: Good Scattering
NDI-L (15) + NDI-GF (12)
Classification: Good Scattering
NDI-L (15) + NDI-I (14)
Classification: Bad Scattering
NDI-M (16) + 2Nap-FF (1)
Classification: Good Scattering
NDI-M (16) + C6-NDI-L (45)
Classification: Bad Scattering
NDI-M (16) + NDI-GF (12)
Classification: Bad Scattering
NDI-M (16) + NDI-I (14)
Classification: Bad Scattering
NDI-V (17) + 2Nap-FF (1)
Classification: Good Scattering
NDI-V (17) + C6-NDI-L (45)
Classification: Good Scattering
NDI-V (17) + NDI-GF (12)
Classification: Good Scattering
NDI-V (17) + NDI-I (14)
Classification: Bad Scattering
NDI-W (18) + 2Nap-FF (1)
Classification: Good Scattering
I(q) (cm-1)
I(q) - Error of I(q) (cm-1)
q (Å-1)
0.01
0.1

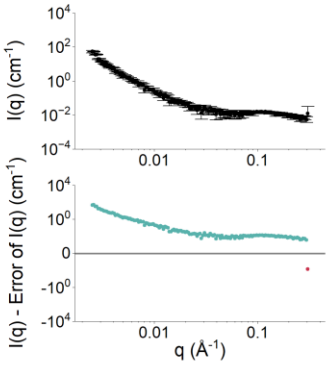
NDI-W (18) + C6-NDI-L (45)
Classification: Good Scattering



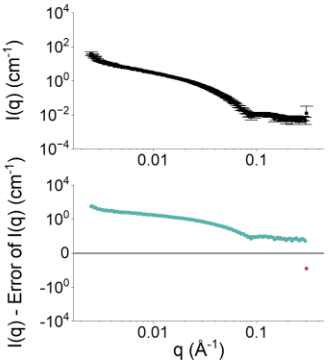
NDI-W (18) + NDI-GF (12)
Classification: Good Scattering



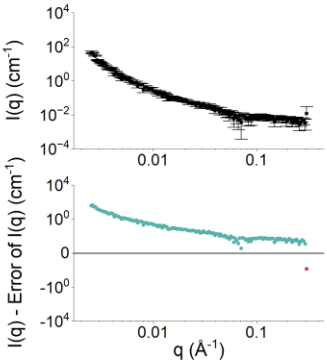
NDI-W (18) + NDI-I (14)
Classification: Good Scattering



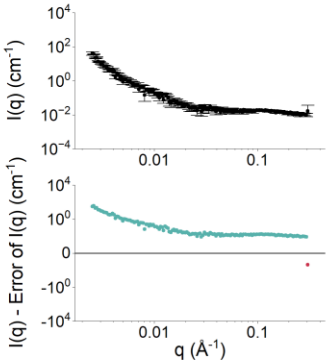
NDI-Y (19) + 2Nap-FF (1)
Classification: Good Scattering



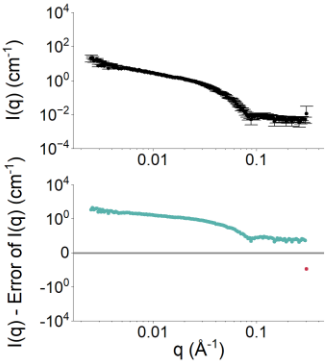
NDI-Y (19) + C6-NDI-L (45)
Classification: Good Scattering



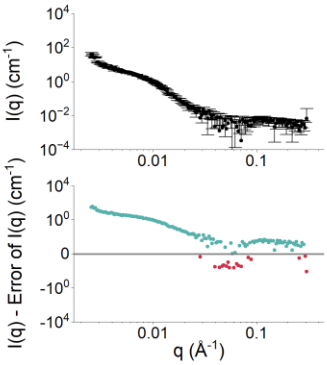
NDI-Y (19) + NDI-GF (12)
Classification: Good Scattering



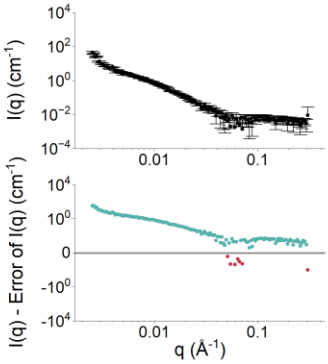
NDI-YI (49) + 2Nap-FF (1)
Classification: Good Scattering



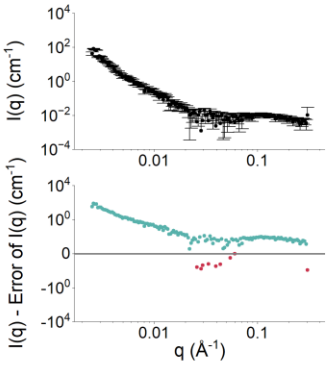
NDI-YI (49) + 2Nap-FI (37)
Classification: Bad Scattering



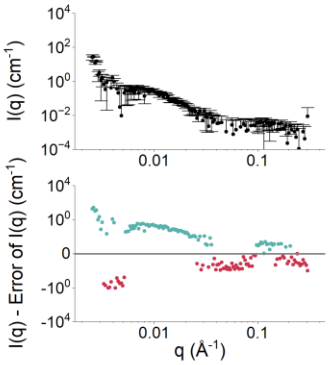
NDI-YI (49) + 2Nap-IF (4)
Classification: Good Scattering



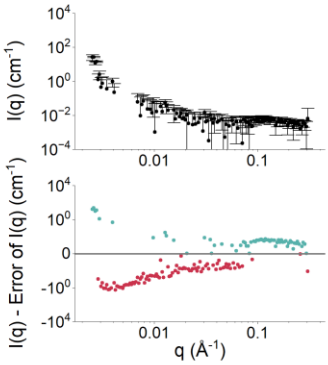
NDI-YI (49) + 2Nap-V (2)
Classification: Bad Scattering



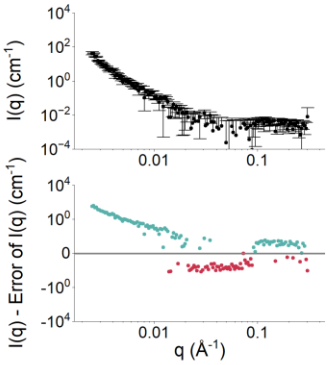
NDI-YI (49) + 6MeO2Nap-FI (6)
Classification: Bad Scattering



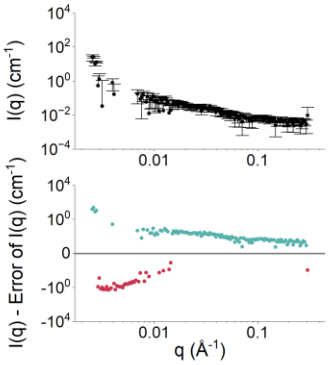
NDI-YI (49) + 7MeO2Nap-FV (7)
Classification: Bad Scattering



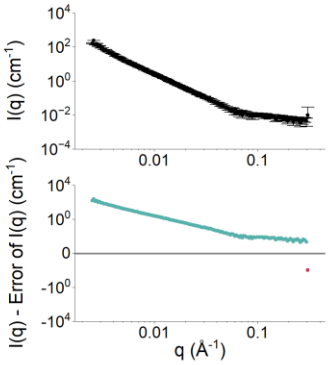
NDI-YI (49) + C6-NDI-A (43)
Classification: Bad Scattering



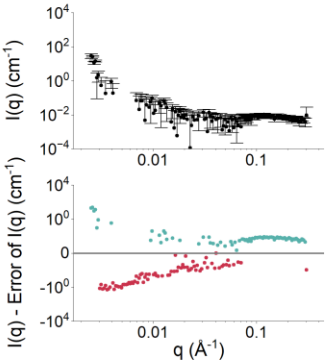
NDI-YI (49) + C6-NDI-F (44)
Classification: Bad Scattering



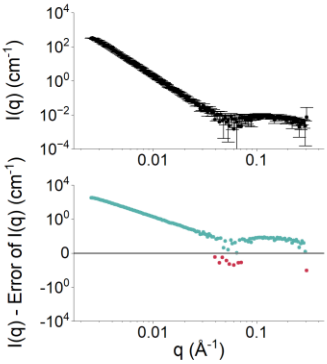
NDI-YI (49) + C6-NDI-L (45)
Classification: Good Scattering



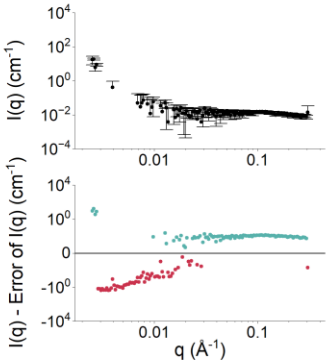
NDI-YI (49) + NDI-A (13)
Classification: Bad Scattering



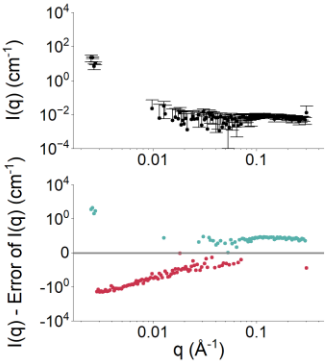
NDI-YI (49) + NDI-G (11)
Classification: Bad Scattering



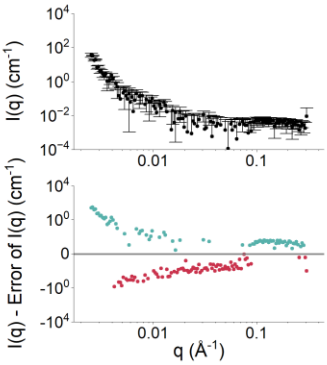
NDI-YI (49) + NDI-GF (12)
Classification: Bad Scattering



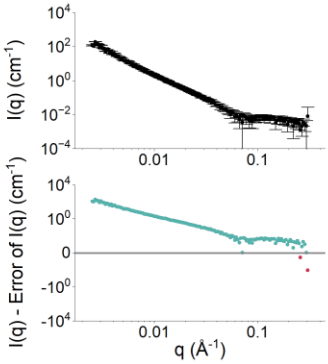
NDI-YI (49) + NDI-H (60)
Classification: Bad Scattering



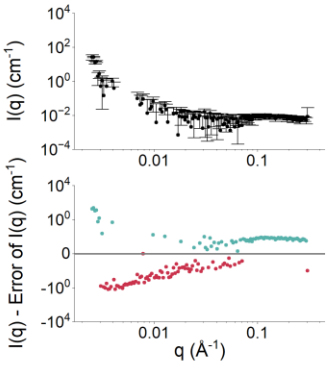
NDI-YI (49) + NDI-I (14)
Classification: Bad Scattering



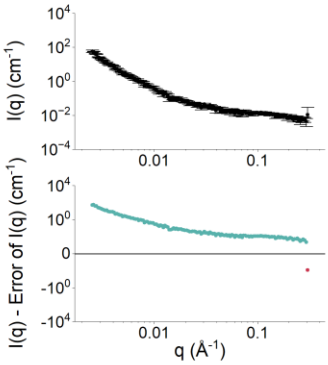
NDI-YI (49) + NDI-L (15)
Classification: Good Scattering



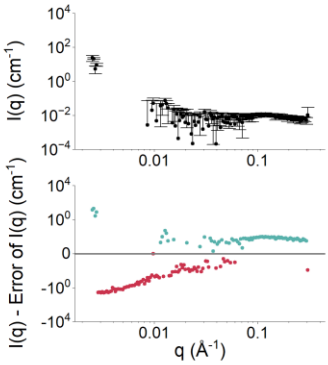
NDI-YI (49) + NDI-M (16)
Classification: Bad Scattering



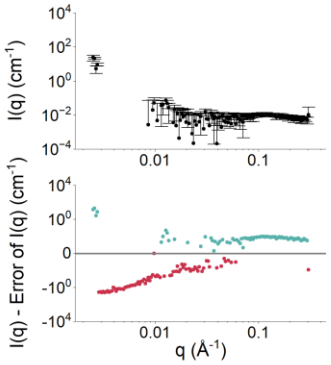
NDI-YI (49) + NDI-W (18)
Classification: Good Scattering



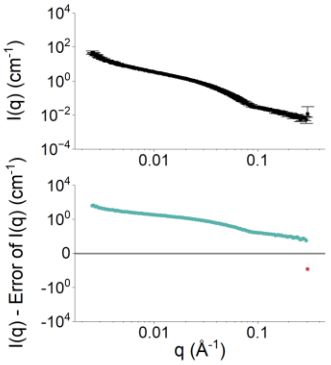
NDI-YI (49) + NDI-Y (19)
Classification: Bad Scattering



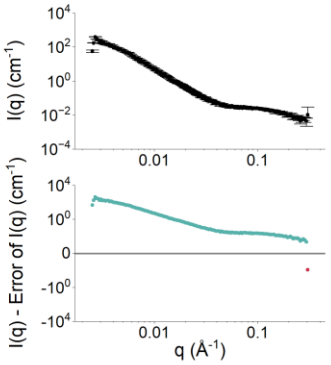
NDI-YI (49) + NDI-Y (19)
Classification: Bad Scattering



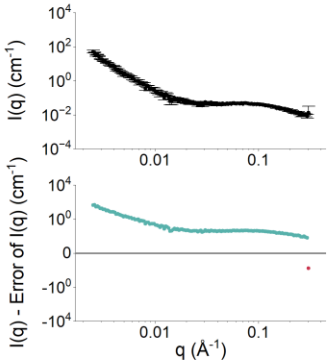
PBI-A 1Py (50) + 2Nap-FF (1)
Classification: Good Scattering



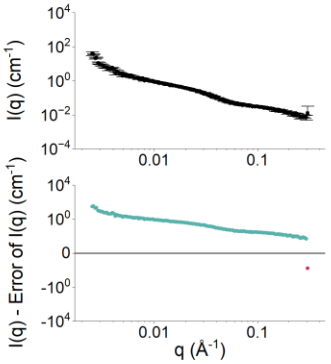
PBI-A 1Py (50) + 2Nap-FI (37)
Classification: Good Scattering



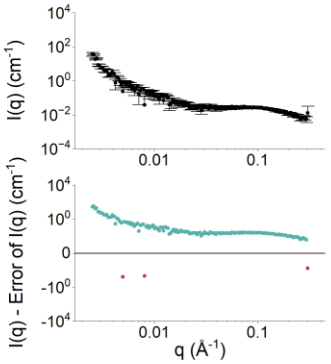
PBI-A 1Py (50) + 2Nap-V (2)
Classification: Good Scattering



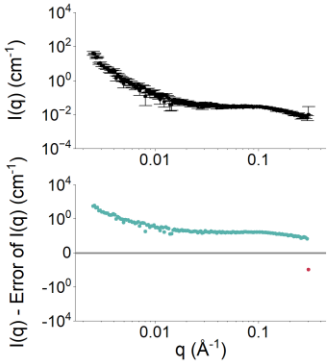
PBI-A 1Py (50) + 6MeO2Nap-FI (6)
Classification: Good Scattering



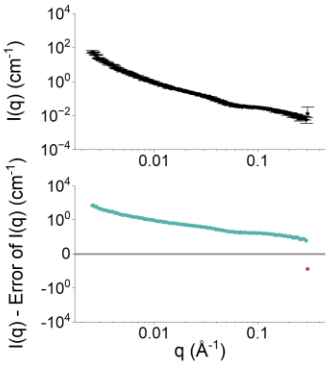
PBI-A 1Py (50) + 7MeO2Nap-FV (7)
Classification: Good Scattering



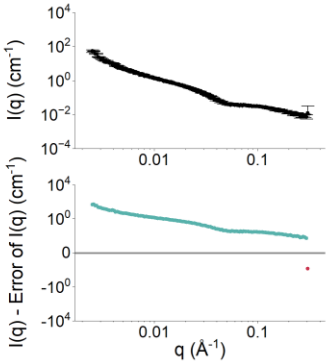
PBI-A 1Py (50) + C6-NDI-A (43)
Classification: Good Scattering



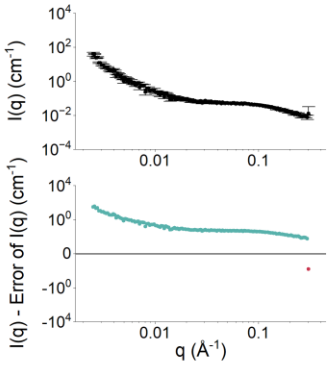
PBI-A 1Py (50) + C6-NDI-F (44)
Classification: Good Scattering



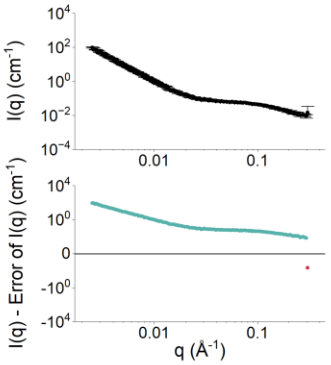
PBI-A 1Py (50) + C6-NDI-L (45)
Classification: Good Scattering



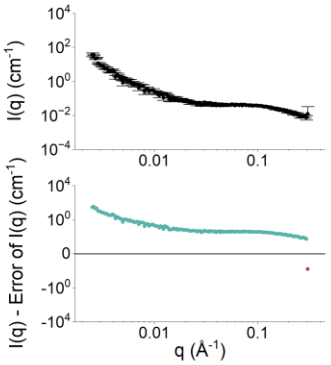
PBI-A 1Py (50) + NDI-A (13)
Classification: Good Scattering



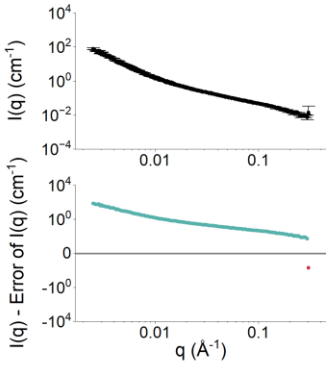
PBI-A 1Py (50) + NDI-F (9)
Classification: Good Scattering



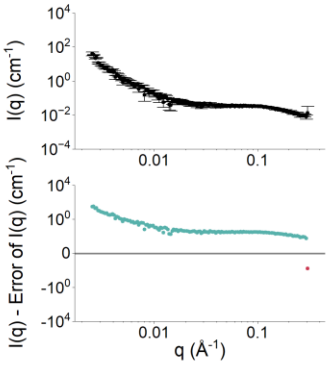
PBI-A 1Py (50) + NDI-G (11)
Classification: Good Scattering



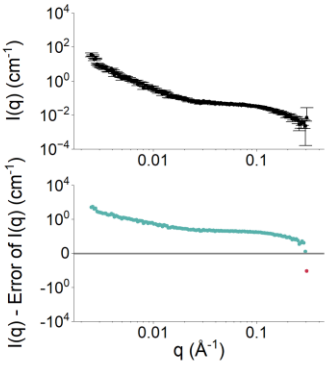
PBI-A 1Py (50) + NDI-GF (12)
Classification: Good Scattering



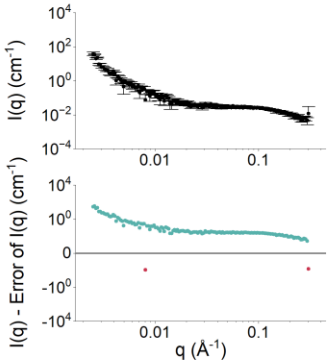
PBI-A 1Py (50) + NDI-H (60)
Classification: Good Scattering



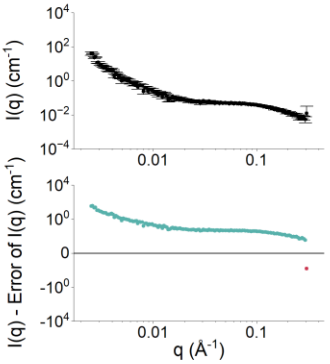
PBI-A 1Py (50) + NDI-I (14)
Classification: Good Scattering



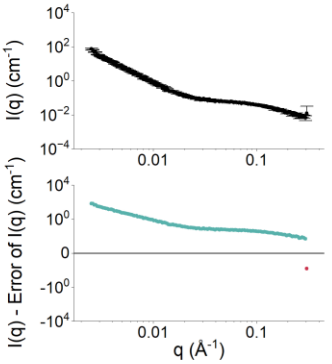
PBI-A 1Py (50) + NDI-L (15)
Classification: Good Scattering



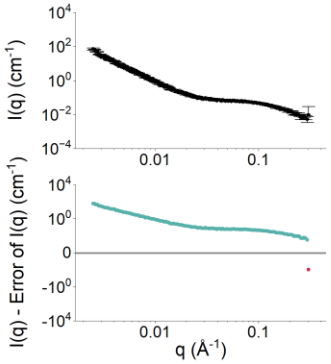
PBI-A 1Py (50) + NDI-M (16)
Classification: Good Scattering



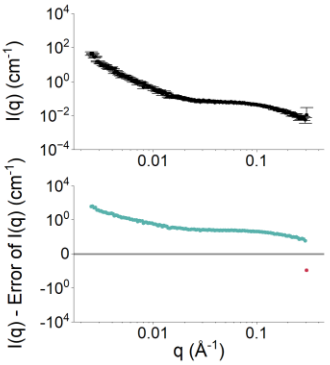
PBI-A 1Py (50) + NDI-V (17)
Classification: Good Scattering



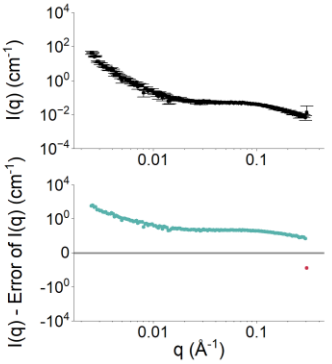
PBI-A 1Py (50) + NDI-W (18)
Classification: Good Scattering



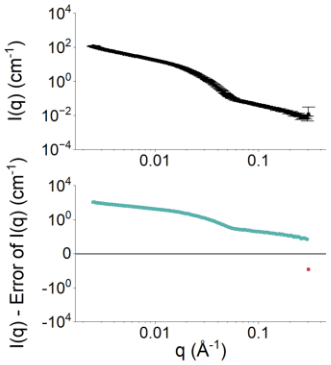
PBI-A 1Py (50) + NDI-Y (19)
Classification: Good Scattering



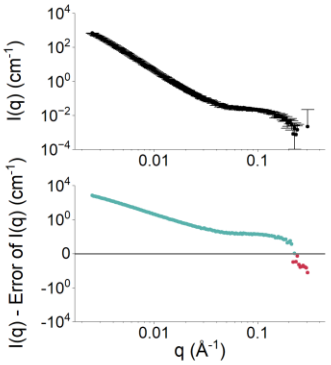
PBI-A 1Py (50) + NDI-YI (49)
Classification: Good Scattering



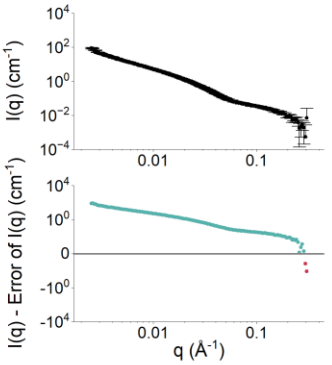
PBI-D (21) + 2Nap-FF (1)
Classification: Good Scattering



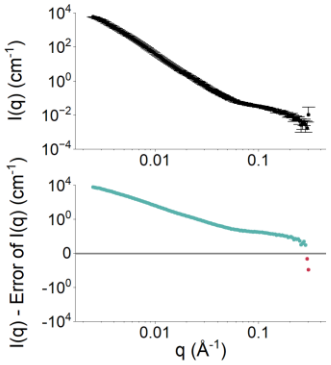
PBI-D (21) + 2Nap-FI (37)
Classification: Good Scattering



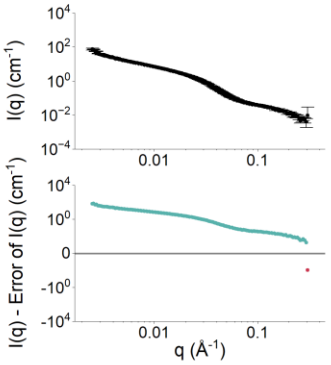
PBI-D (21) + 2Nap-IF (4)
Classification: Good Scattering



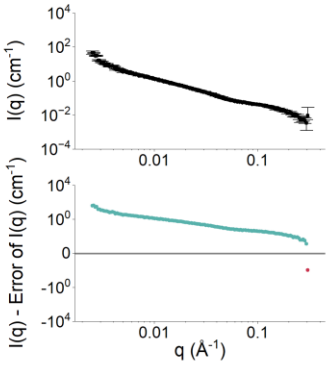
PBI-D (21) + 2Nap-V (2)
Classification: Good Scattering



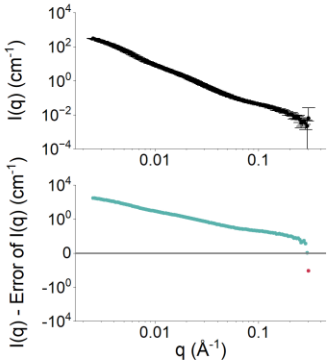
PBI-D (21) + 6MeO2Nap-FI (6)
Classification: Good Scattering



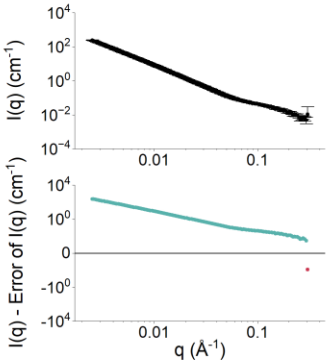
PBI-D (21) + C6-NDI-A (43)
Classification: Good Scattering



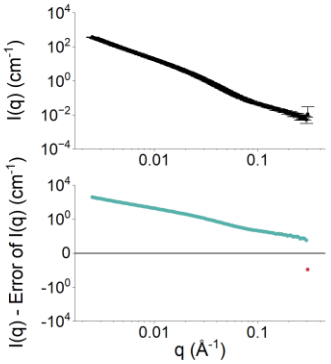
PBI-D (21) + C6-NDI-F (44)
Classification: Good Scattering



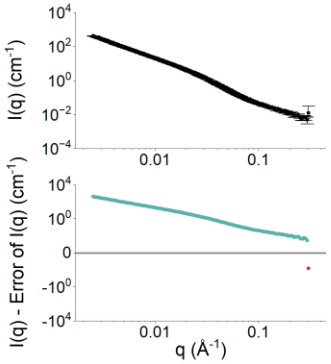
PBI-D (21) + C6-NDI-L (45)
Classification: Good Scattering



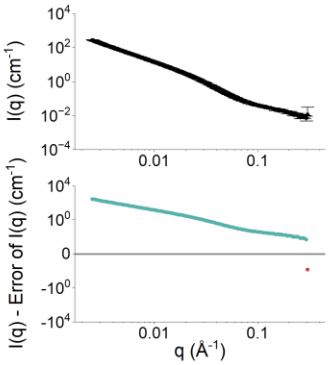
PBI-D (21) + NDI-A (13)
Classification: Good Scattering



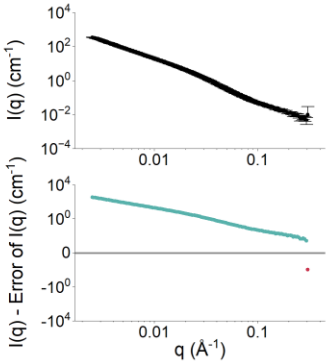
PBI-D (21) + NDI-F (9)
Classification: Good Scattering



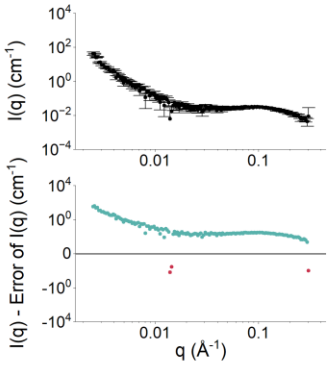
PBI-D (21) + NDI-G (11)
Classification: Good Scattering



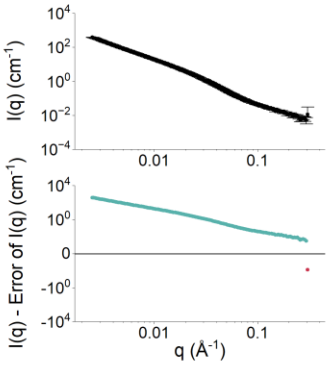
PBI-D (21) + NDI-GF (12)
Classification: Good Scattering



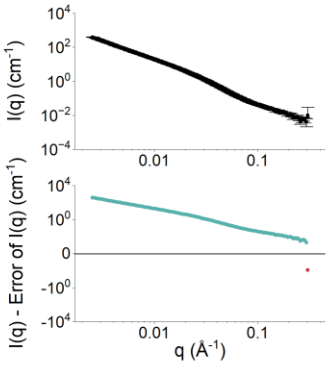
PBI-D (21) + NDI-H (60)
Classification: Good Scattering



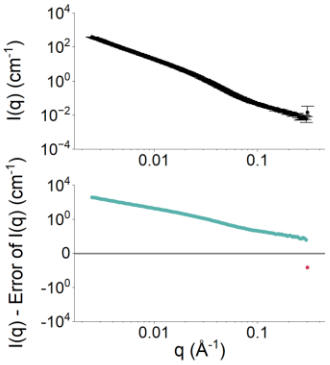
PBI-D (21) + NDI-I (14)
Classification: Good Scattering



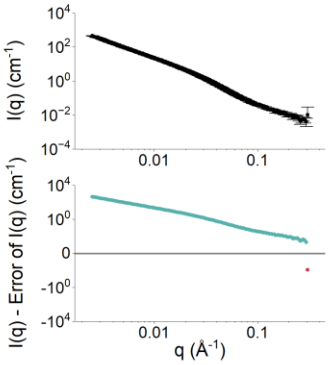
PBI-D (21) + NDI-L (15)
Classification: Good Scattering



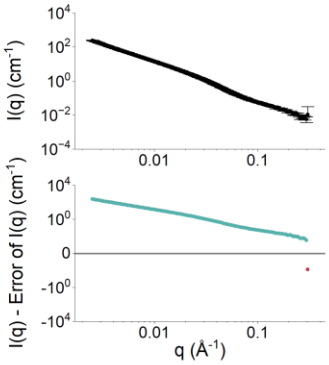
PBI-D (21) + NDI-M (16)
Classification: Good Scattering



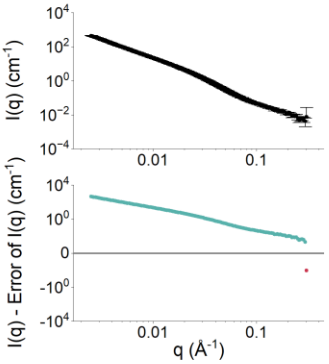
PBI-D (21) + NDI-V (17)
Classification: Good Scattering



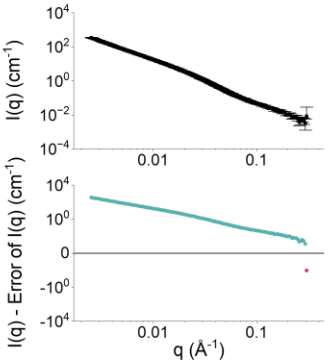
PBI-D (21) + NDI-W (18)
Classification: Good Scattering



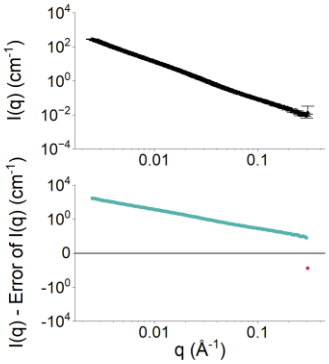
PBI-D (21) + NDI-Y (19)
Classification: Good Scattering



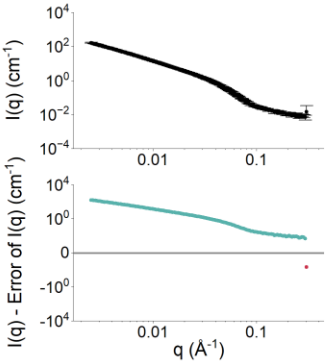
PBI-D (21) + NDI-YI (49)
Classification: Good Scattering



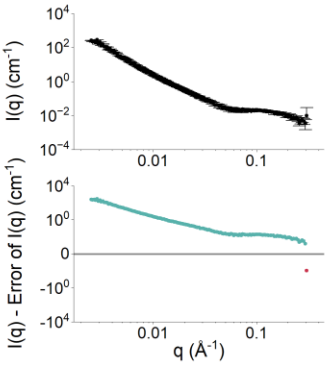
PBI-D (21) + PBI-A 1Py (50)
Classification: Good Scattering



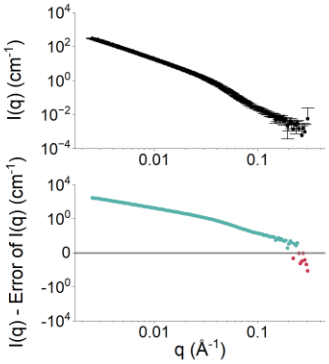
PBI-DOPA (22) + 2Nap-FF (1)
Classification: Good Scattering



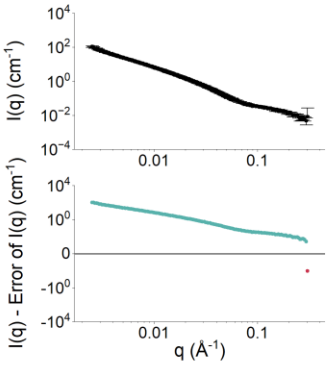
PBI-DOPA (22) + 2Nap-FI (37)
Classification: Good Scattering



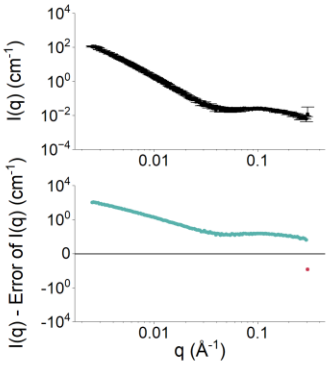
PBI-DOPA (22) + 2Nap-IF (4)
Classification: Good Scattering



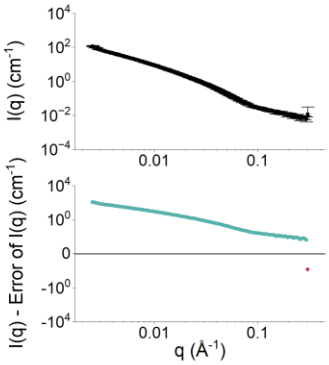
PBI-DOPA (22) + 2Nap-V (2)
Classification: Good Scattering



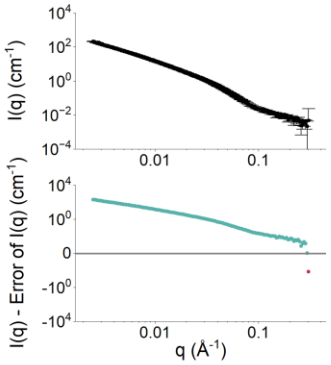
PBI-DOPA (22) + 7MeO2Nap-FV (7)
Classification: Good Scattering



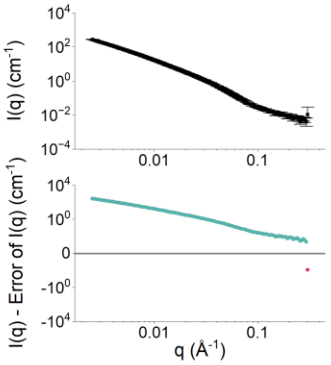
PBI-DOPA (22) + C6-NDI-A (43)
Classification: Good Scattering



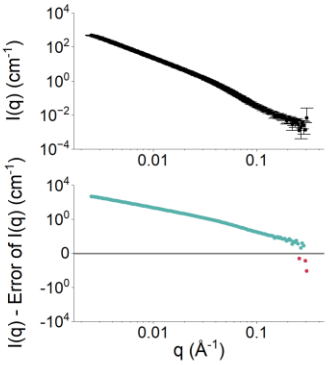
PBI-DOPA (22) + C6-NDI-F (44)
Classification: Good Scattering



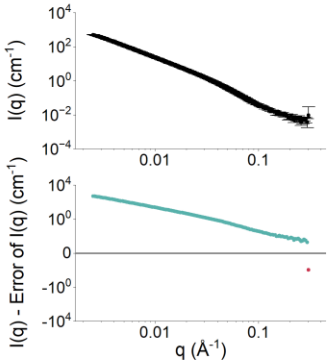
PBI-DOPA (22) + C6-NDI-L (45)
Classification: Good Scattering



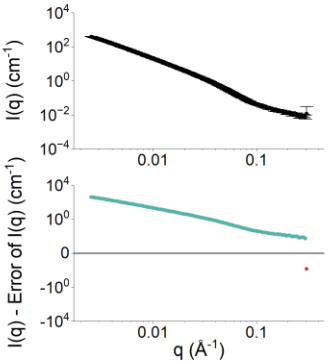
PBI-DOPA (22) + NDI-A (13)
Classification: Good Scattering



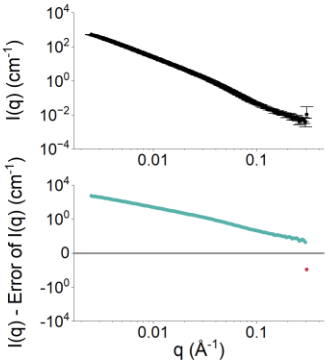
PBI-DOPA (22) + NDI-F (9)
Classification: Good Scattering



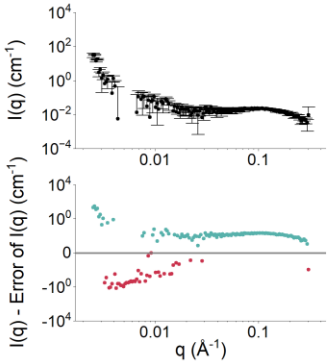
PBI-DOPA (22) + NDI-G (11)
Classification: Good Scattering



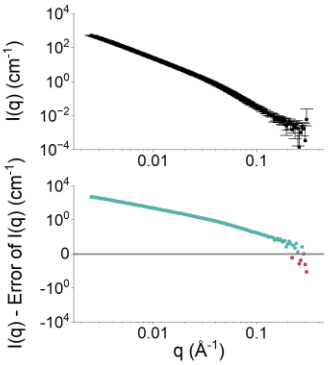
PBI-DOPA (22) + NDI-GF (12)
Classification: Good Scattering



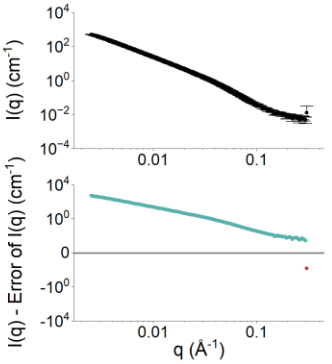
PBI-DOPA (22) + NDI-H (60)
Classification: Bad Scattering



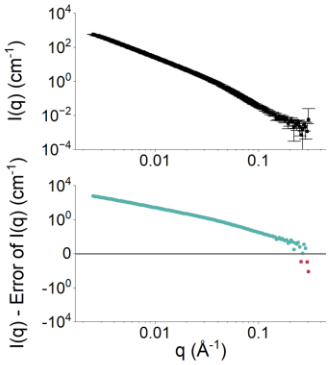
PBI-DOPA (22) + NDI-I (14)
Classification: Good Scattering



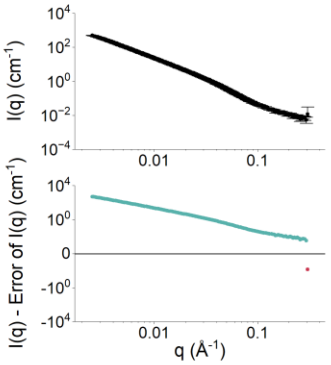
PBI-DOPA (22) + NDI-L (15)
Classification: Good Scattering



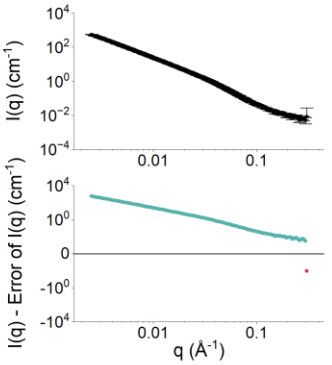
PBI-DOPA (22) + NDI-V (17)
Classification: Good Scattering



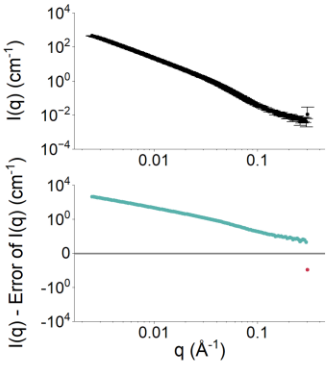
PBI-DOPA (22) + NDI-W (18)
Classification: Good Scattering



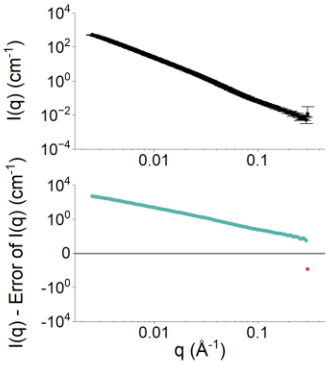
PBI-DOPA (22) + NDI-Y (19)
Classification: Good Scattering



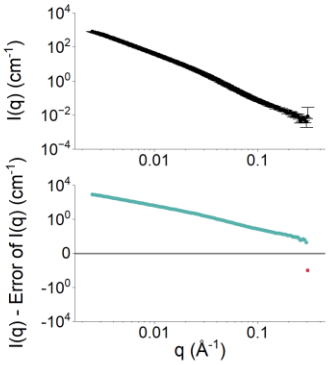
PBI-DOPA (22) + NDI-YI (49)
Classification: Good Scattering



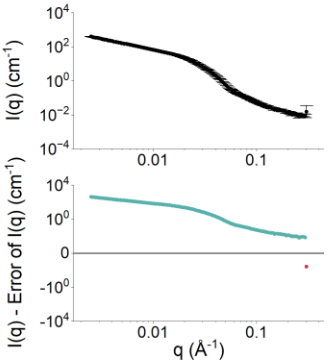
PBI-DOPA (22) + PBI-A 1Py (50)
Classification: Good Scattering



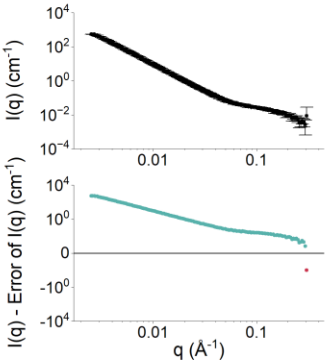
PBI-DOPA (22) + PBI-D (21)
Classification: Good Scattering



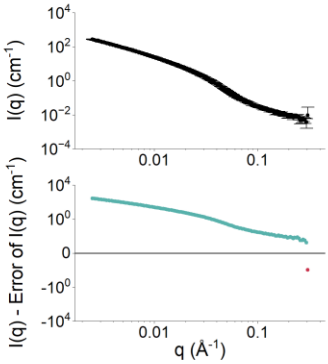
PBI-E (23) + 2Nap-FF (1)
Classification: Good Scattering



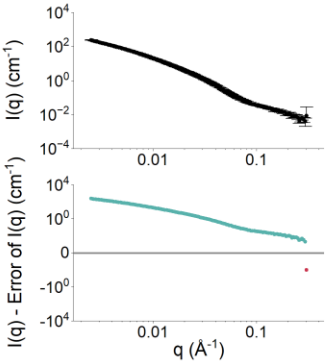
PBI-E (23) + 2Nap-FI (37)
Classification: Good Scattering



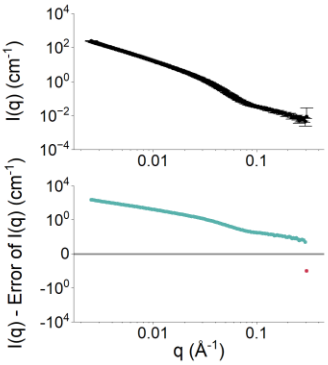
PBI-E (23) + 2Nap-IF (4)
Classification: Good Scattering



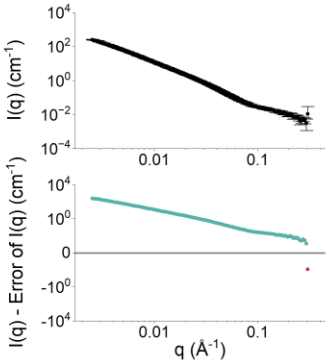
PBI-E (23) + 2Nap-V (2)
Classification: Good Scattering



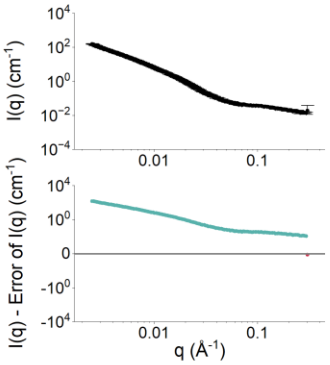
PBI-E (23) + 6MeO2Nap-FI (6)
Classification: Good Scattering



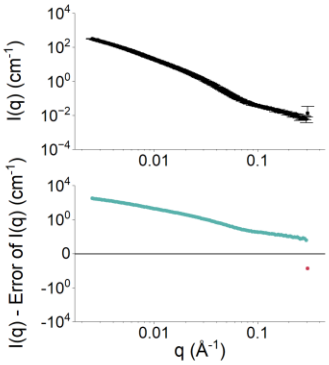
PBI-E (23) + 7MeO2Nap-FV (7)
Classification: Good Scattering



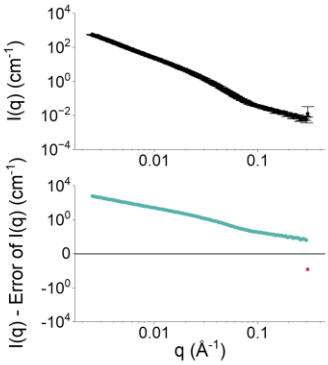
PBI-E (23) + C6-NDI-A (43)
Classification: Good Scattering



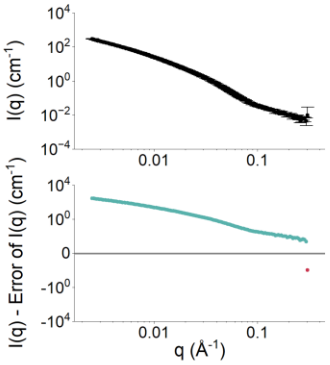
PBI-E (23) + C6-NDI-F (44)
Classification: Good Scattering



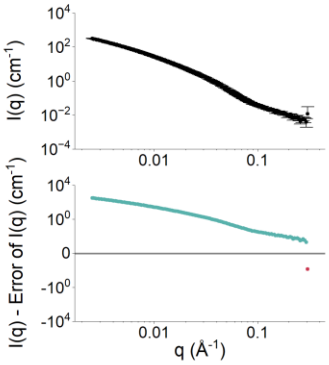
PBI-E (23) + C6-NDI-L (45)
Classification: Good Scattering



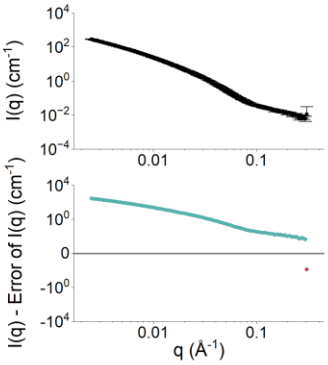
PBI-E (23) + NDI-A (13)
Classification: Good Scattering



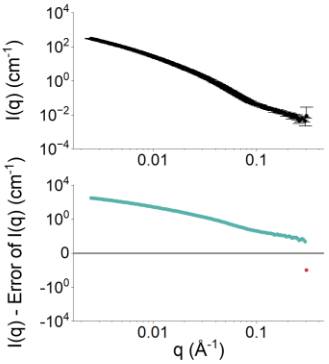
PBI-E (23) + NDI-F (9)
Classification: Good Scattering



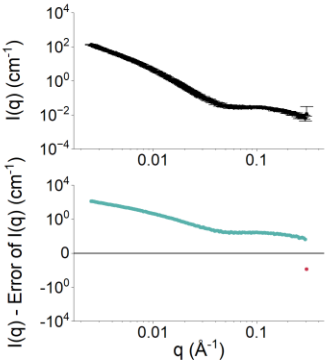
PBI-E (23) + NDI-G (11)
Classification: Good Scattering



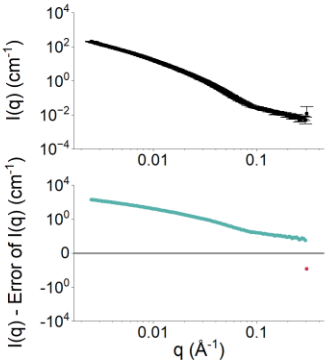
PBI-E (23) + NDI-GF (12)
Classification: Good Scattering



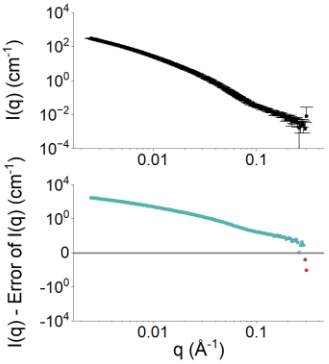
PBI-E (23) + NDI-H (60)
Classification: Good Scattering



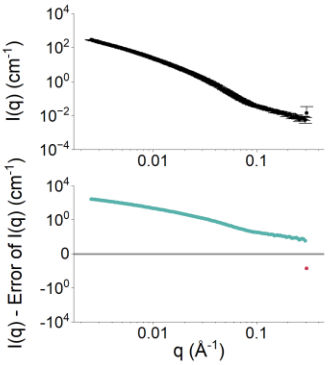
PBI-E (23) + NDI-I (14)
Classification: Good Scattering



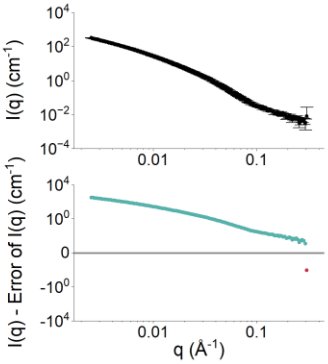
PBI-E (23) + NDI-L (15)
Classification: Good Scattering



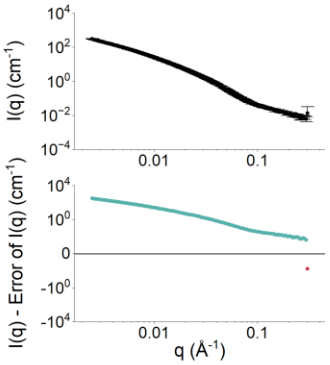
PBI-E (23) + NDI-M (16)
Classification: Good Scattering



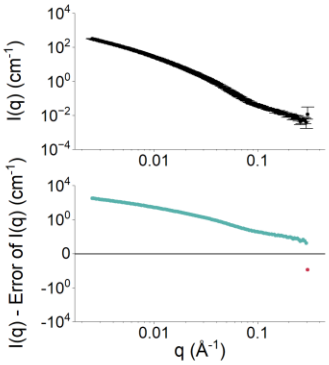
PBI-E (23) + NDI-V (17)
Classification: Good Scattering



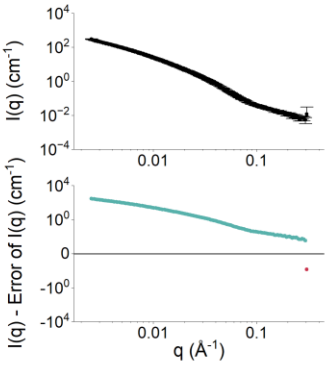
PBI-E (23) + NDI-W (18)
Classification: Good Scattering



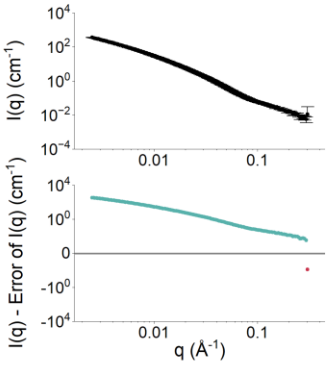
PBI-E (23) + NDI-Y (19)
Classification: Good Scattering



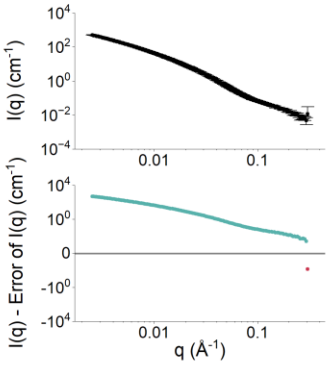
PBI-E (23) + NDI-YI (49)
Classification: Good Scattering



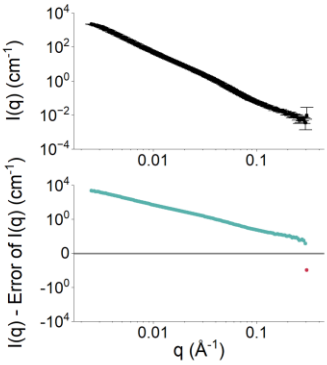
PBI-E (23) + PBI-A 1Py (50)
Classification: Good Scattering



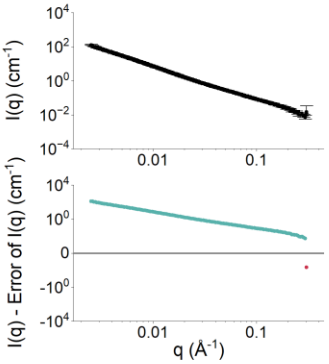
PBI-E (23) + PBI-D (21)
Classification: Good Scattering



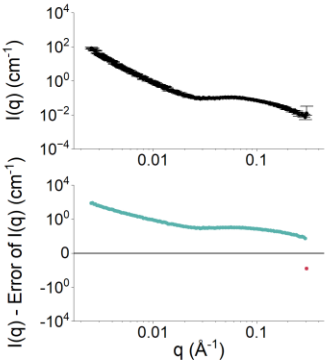
PBI-E (23) + PBI-DOPA (22)
Classification: Good Scattering



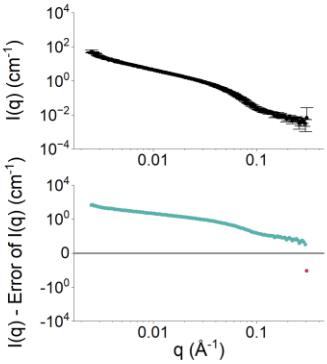
PBI-H (26) + PBI-E (23)
Classification: Good Scattering



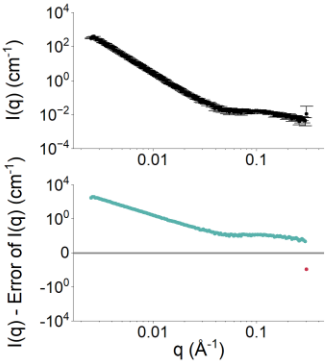
PBI-H (26) + PBI-G (25)
Classification: Good Scattering



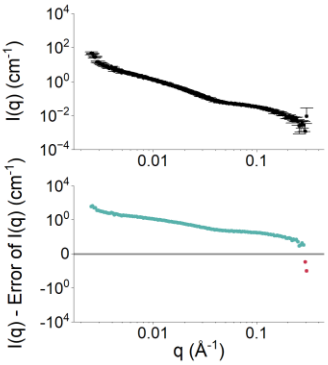
PBI-I (27) + 2Nap-FF (1)
Classification: Good Scattering



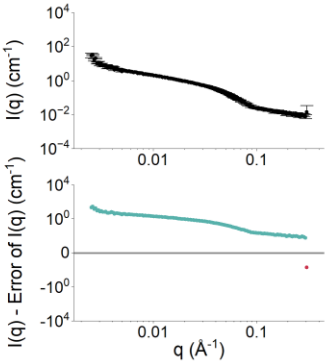
PBI-I (27) + 2Nap-FI (37)
Classification: Good Scattering



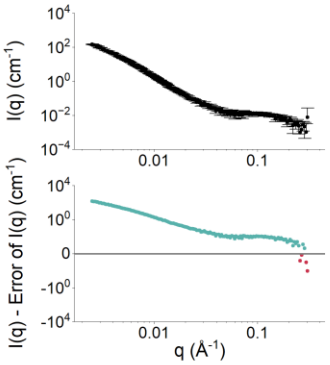
PBI-I (27) + 2Nap-IF (4)
Classification: Good Scattering



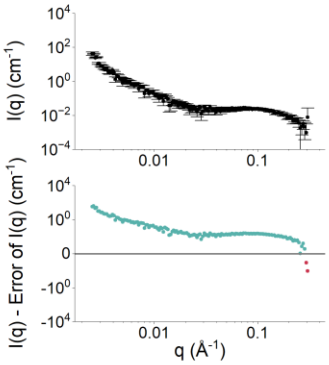
PBI-F (24) + 2Nap-FF (1)
Classification: Good Scattering



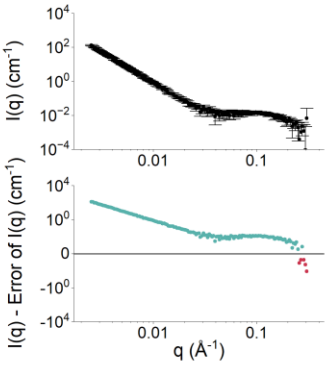
PBI-F (24) + 2Nap-FI (37)
Classification: Good Scattering



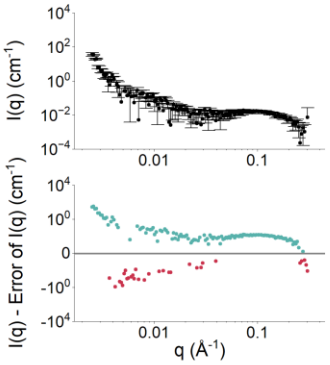
PBI-F (24) + 2Nap-V (2)
Classification: Good Scattering



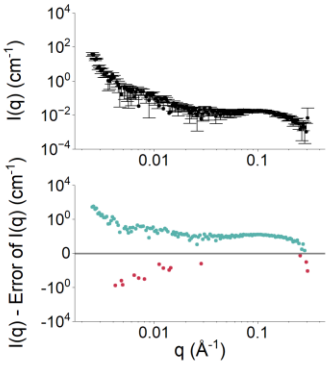
PBI-F (24) + 6MeO2Nap-FI (6)
Classification: Good Scattering



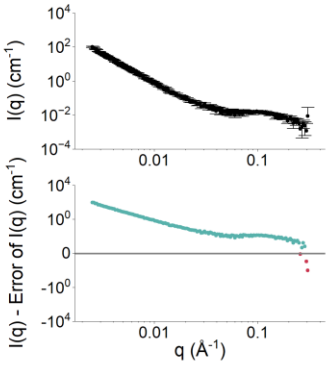
PBI-F (24) + 7MeO2Nap-FV (7)
Classification: Bad Scattering



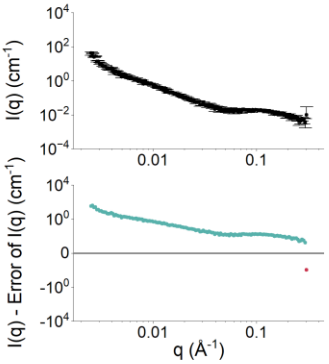
PBI-F (24) + C6-NDI-A (43)
Classification: Bad Scattering



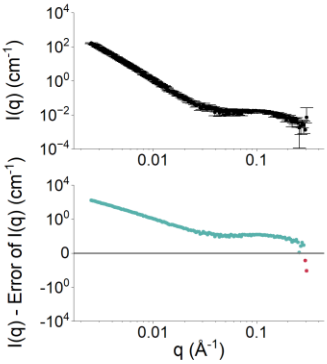
PBI-F (24) + C6-NDI-F (44)
Classification: Good Scattering



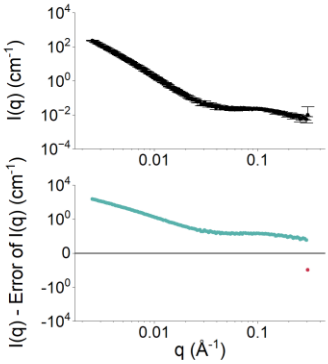
PBI-F (24) + C6-NDI-L (45)
Classification: Good Scattering



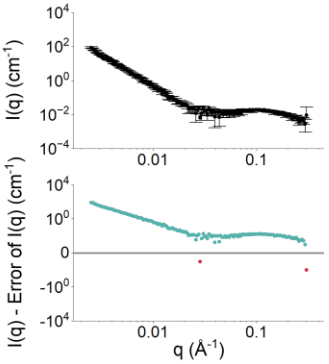
PBI-F (24) + NDI-A (13)
Classification: Good Scattering



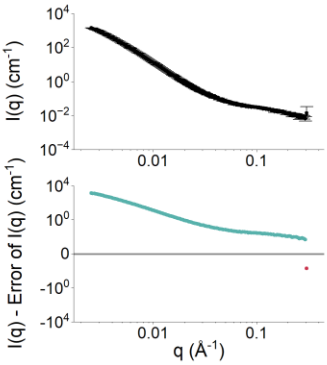
PBI-F (24) + NDI-F (9)
Classification: Good Scattering



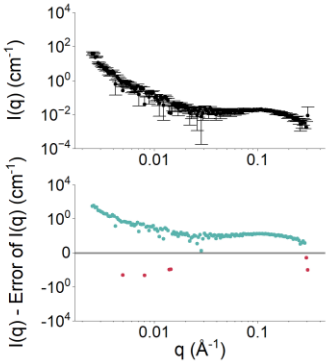
PBI-F (24) + NDI-G (11)
Classification: Good Scattering



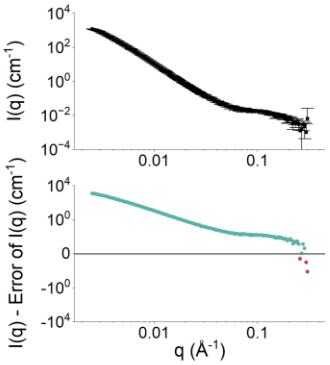
PBI-F (24) + NDI-GF (12)
Classification: Good Scattering



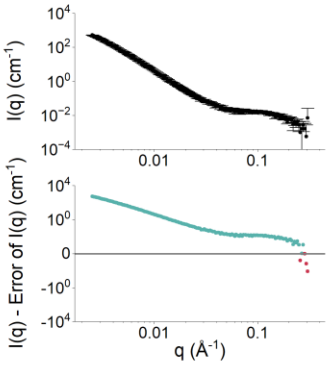
PBI-F (24) + NDI-H (60)
Classification: Good Scattering



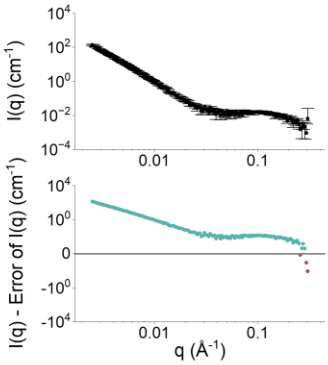
PBI-F (24) + NDI-I (14)
Classification: Good Scattering



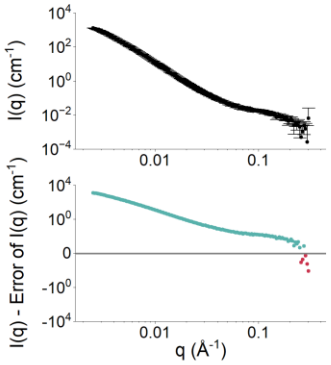
PBI-F (24) + NDI-L (15)
Classification: Good Scattering



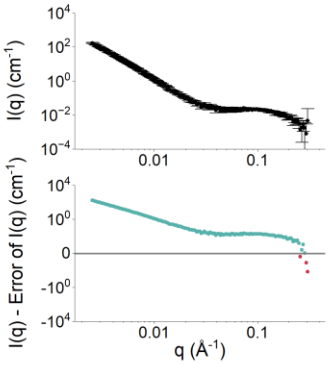
PBI-F (24) + NDI-M (16)
Classification: Good Scattering



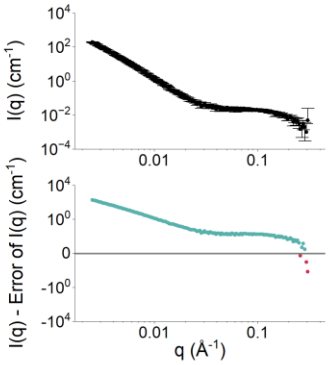
PBI-F (24) + NDI-V (17)
Classification: Good Scattering



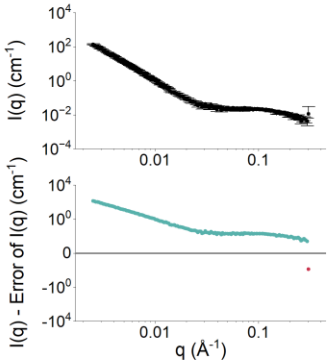
PBI-F (24) + NDI-W (18)
Classification: Good Scattering



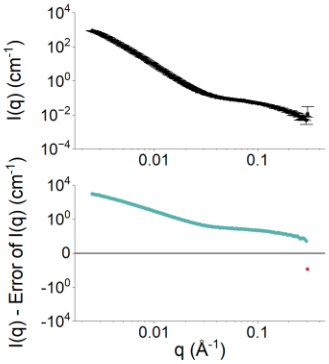
PBI-F (24) + NDI-Y (19)
Classification: Good Scattering



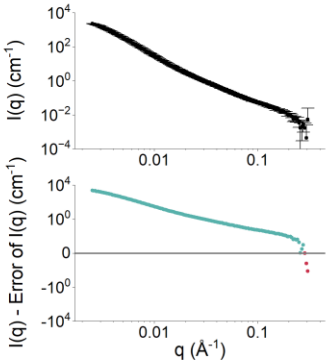
PBI-F (24) + NDI-YI (49)
Classification: Good Scattering



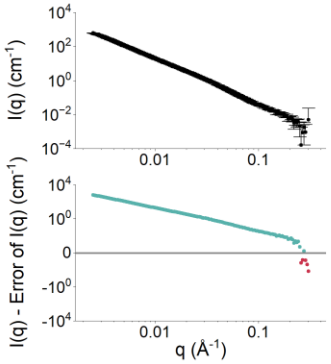
PBI-F (24) + PBI-A 1Py (50)
Classification: Good Scattering



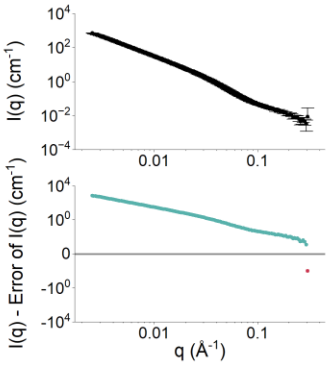
PBI-F (24) + PBI-D (21)
Classification: Good Scattering



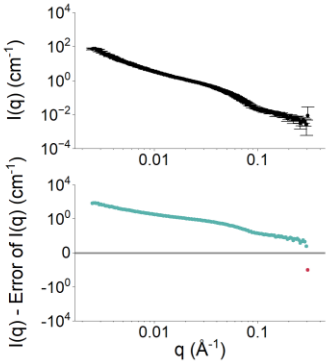
PBI-F (24) + PBI-DOPA (22)
Classification: Good Scattering



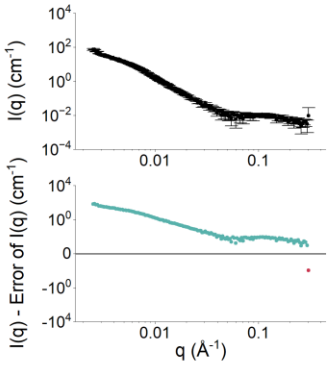
PBI-F (24) + PBI-E (23)
Classification: Good Scattering



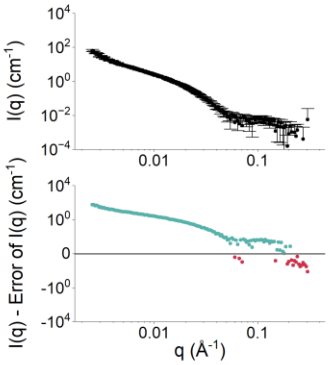
PBI-G (25) + 2Nap-FF (1)
Classification: Good Scattering



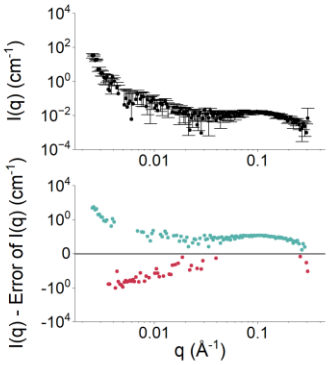
PBI-G (25) + 2Nap-FI (37)
Classification: Good Scattering



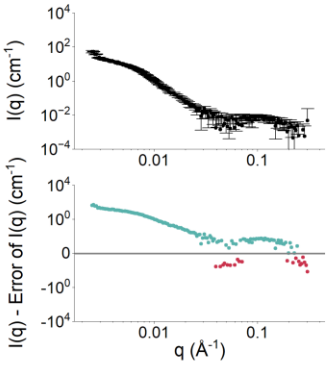
PBI-G (25) + 2Nap-IF (4)
Classification: Good Scattering



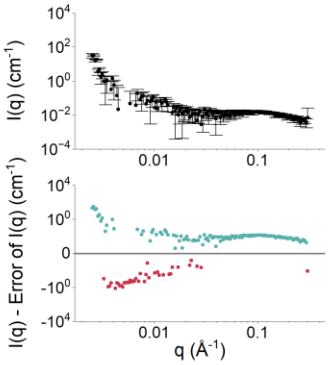
PBI-G (25) + 2Nap-V (2)
Classification: Bad Scattering



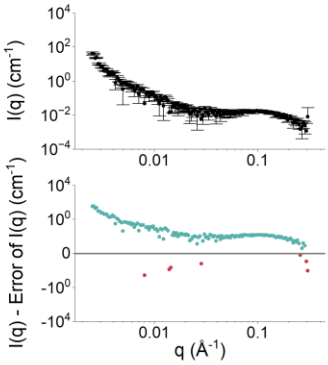
PBI-G (25) + 6MeO2Nap-FI (6)
Classification: Bad Scattering



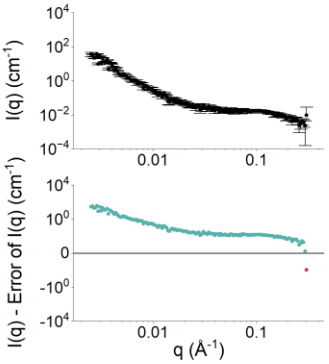
PBI-G (25) + 7MeO2Nap-FV (7)
Classification: Bad Scattering



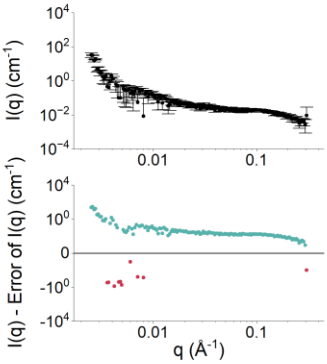
PBI-G (25) + C6-NDI-A (43)
Classification: Good Scattering



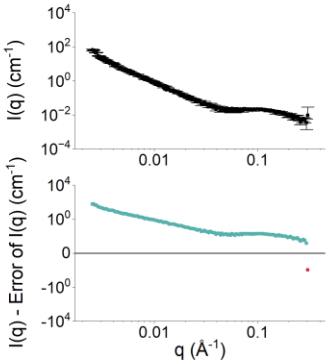
PBI-G (25) + C6-NDI-F (44)
Classification: Good Scattering



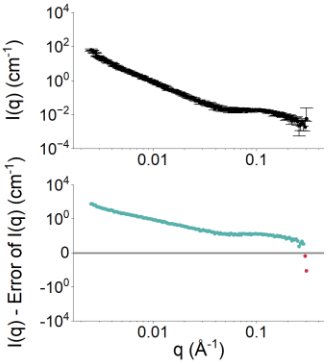
PBI-G (25) + C6-NDI-L (45)
Classification: Bad Scattering



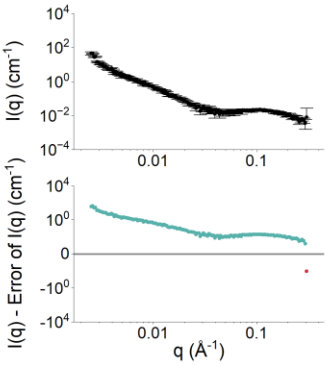
PBI-G (25) + NDI-A (13)
Classification: Good Scattering



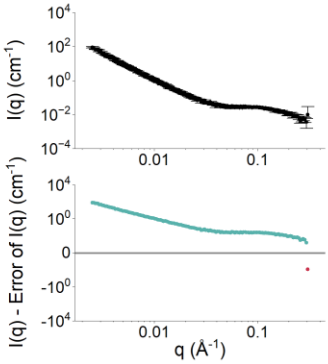
PBI-G (25) + NDI-F (9)
Classification: Good Scattering



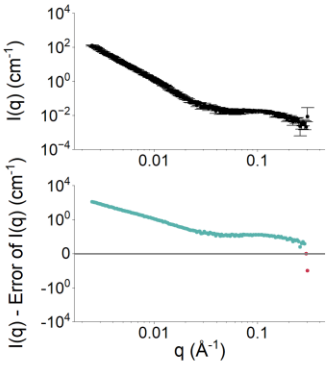
PBI-G (25) + NDI-G (11)
Classification: Good Scattering



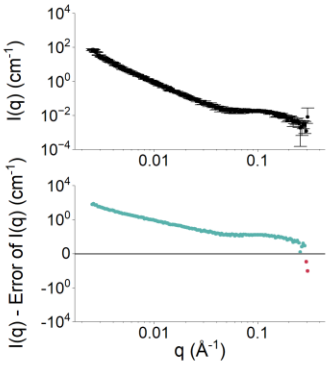
PBI-G (25) + NDI-GF (12)
Classification: Good Scattering



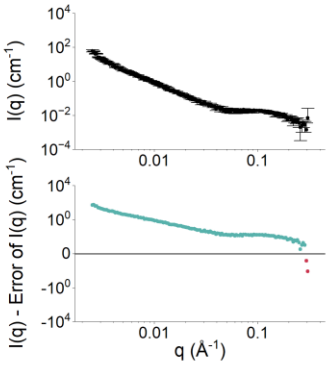
PBI-G (25) + NDI-H (60)
Classification: Good Scattering



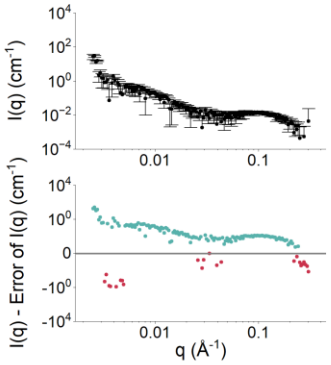
PBI-G (25) + NDI-I (14)
Classification: Good Scattering



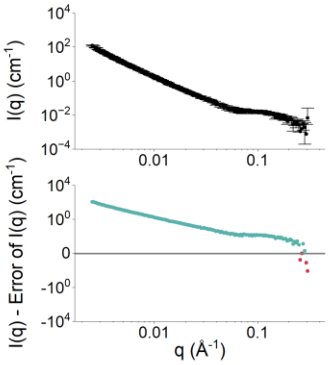
PBI-G (25) + NDI-L (15)
Classification: Good Scattering



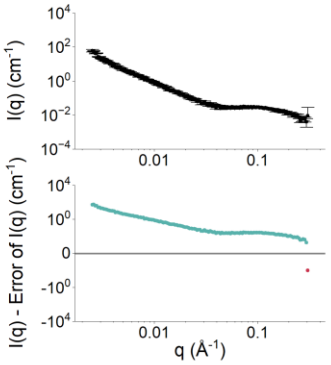
PBI-G (25) + NDI-M (16)
Classification: Bad Scattering



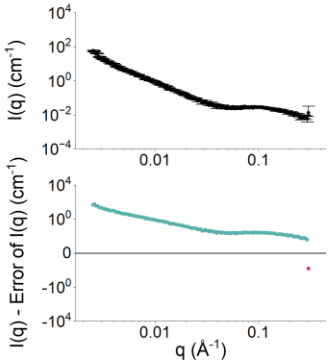
PBI-G (25) + NDI-V (17)
Classification: Good Scattering



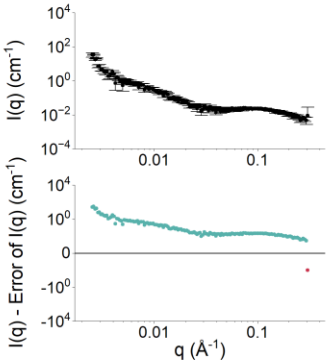
PBI-G (25) + NDI-W (18)
Classification: Good Scattering



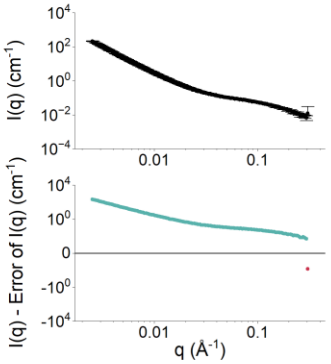
PBI-G (25) + NDI-Y (19)
Classification: Good Scattering



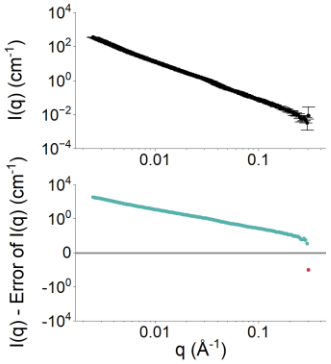
PBI-G (25) + NDI-YI (49)
Classification: Good Scattering



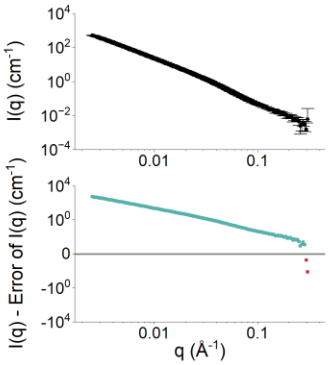
PBI-G (25) + PBI-A 1Py (50)
Classification: Good Scattering



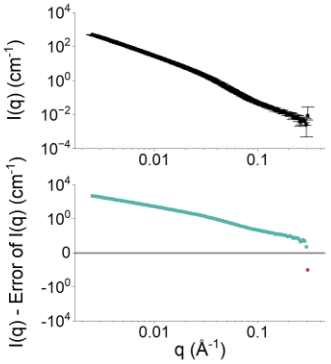
PBI-G (25) + PBI-D (21)
Classification: Good Scattering



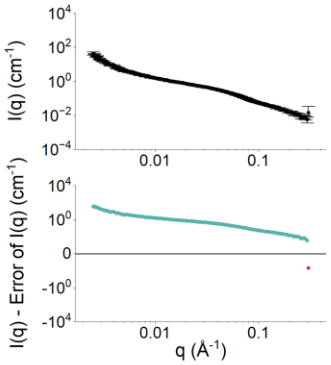
PBI-G (25) + PBI-DOPA (22)
Classification: Good Scattering



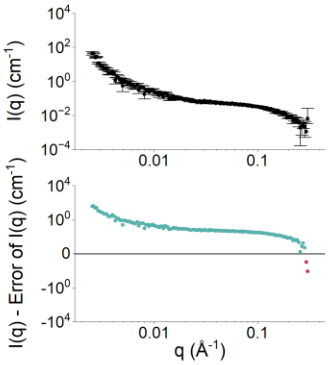
PBI-G (25) + PBI-E (23)
Classification: Good Scattering



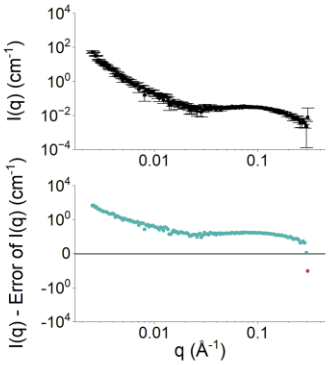
PBI-H (26) + 2Nap-FF (1)
Classification: Good Scattering



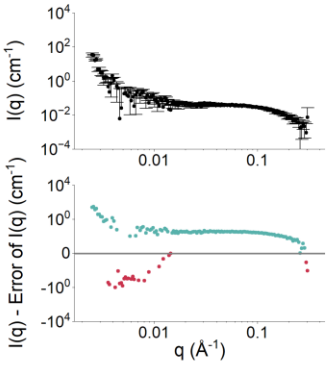
PBI-H (26) + 2Nap-FI (37)
Classification: Good Scattering



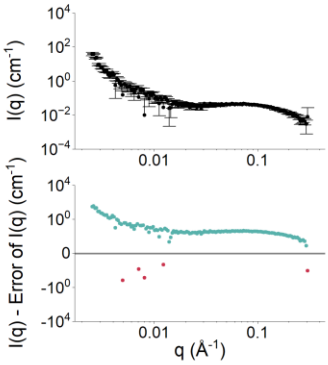
PBI-H (26) + 2Nap-V (2)
Classification: Good Scattering



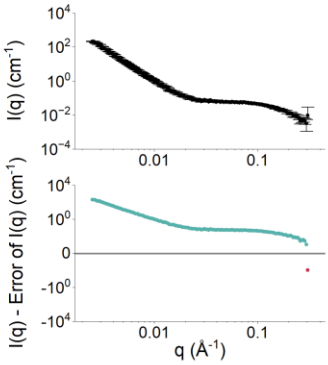
PBI-H (26) + 6MeO2Nap-FI (6)
Classification: Bad Scattering



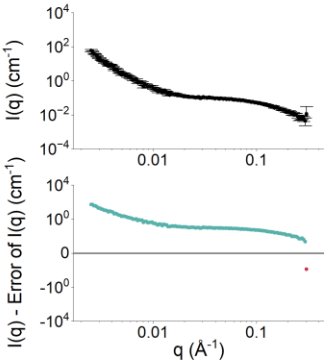
PBI-H (26) + 7MeO2Nap-FV (7)
Classification: Good Scattering



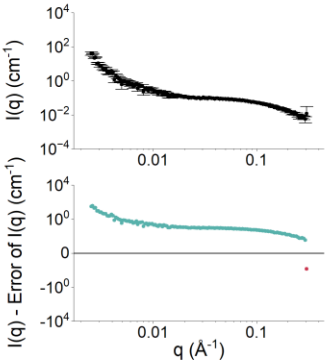
PBI-H (26) + C6-NDI-A (43)
Classification: Good Scattering



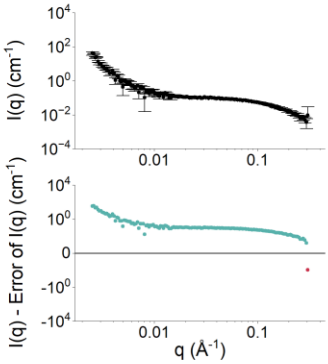
PBI-H (26) + C6-NDI-F (44)
Classification: Good Scattering



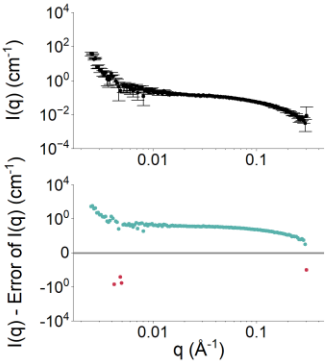
PBI-H (26) + C6-NDI-L (45)
Classification: Good Scattering



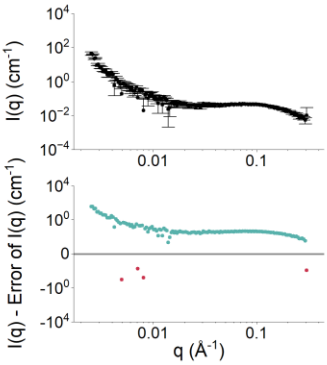
PBI-H (26) + NDI-A (13)
Classification: Good Scattering



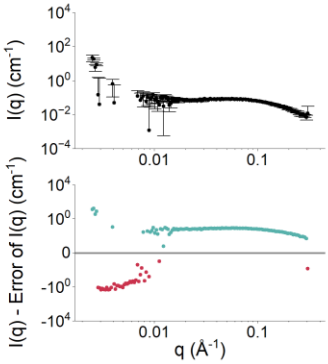
PBI-H (26) + NDI-F (9)
Classification: Good Scattering



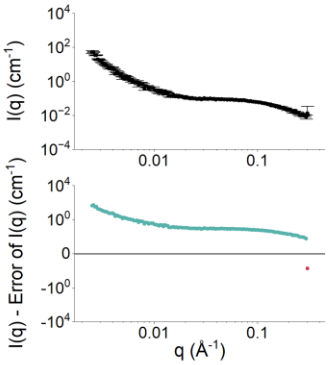
PBI-H (26) + NDI-G (11)
Classification: Good Scattering



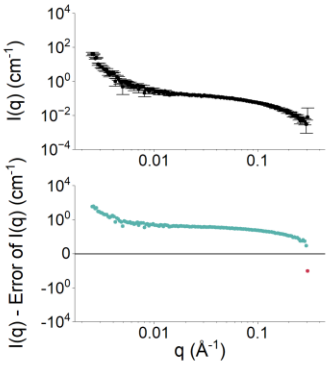
PBI-H (26) + NDI-GF (12)
Classification: Bad Scattering



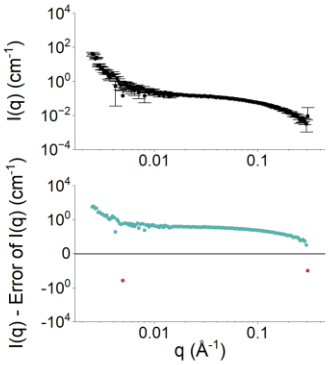
PBI-H (26) + NDI-H (60)
Classification: Good Scattering



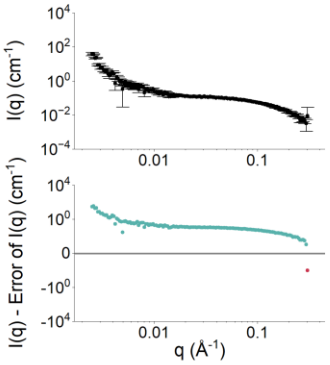
PBI-H (26) + NDI-I (14)
Classification: Good Scattering



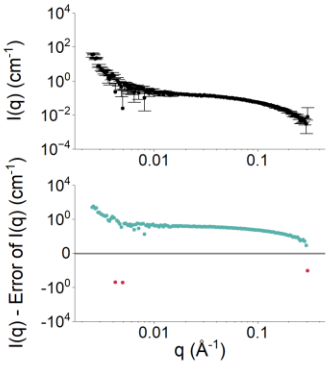
PBI-H (26) + NDI-L (15)
Classification: Good Scattering



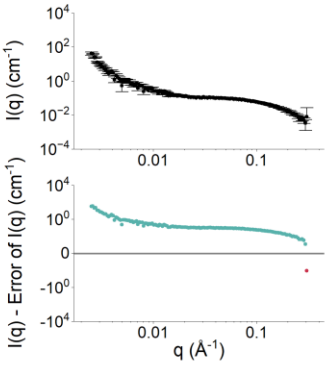
PBI-H (26) + NDI-M (16)
Classification: Good Scattering



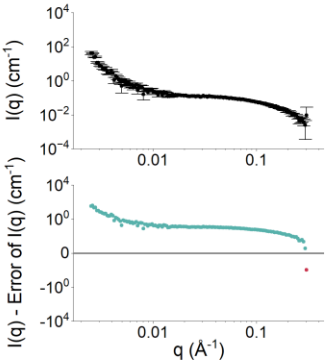
PBI-H (26) + NDI-V (17)
Classification: Good Scattering



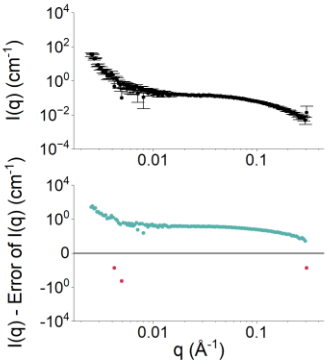
PBI-H (26) + NDI-W (18)
Classification: Good Scattering



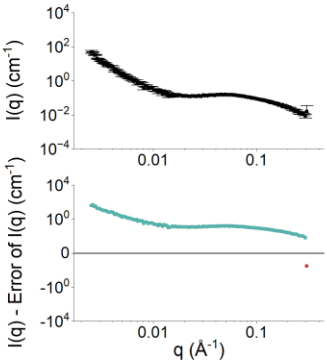
PBI-H (26) + NDI-Y (19)
Classification: Good Scattering



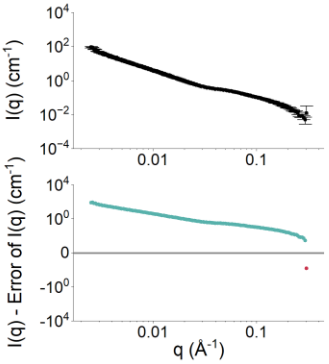
PBI-H (26) + NDI-YI (49)
Classification: Good Scattering



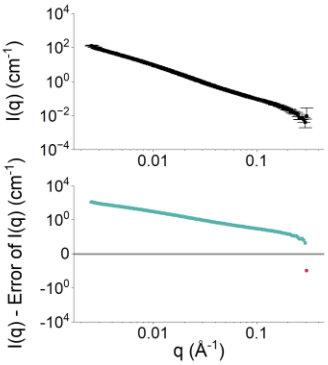
PBI-H (26) + PBI-A 1Py (50)
Classification: Good Scattering



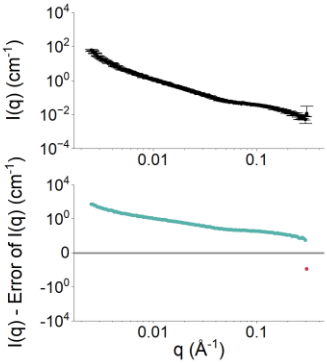
PBI-H (26) + PBI-D (21)
Classification: Good Scattering



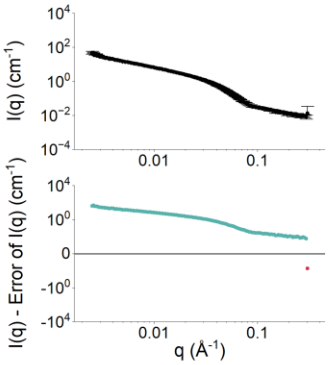
PBI-H (26) + PBI-DOPA (22)
Classification: Good Scattering



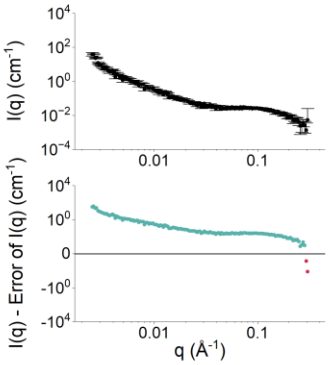
PBI-S (30) + PBI-L (28)
Classification: Good Scattering



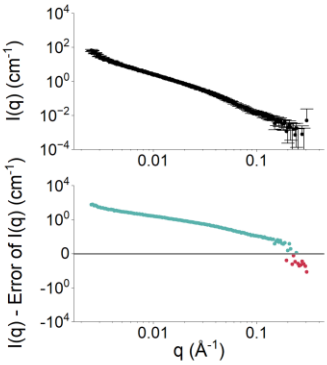
PBI-V (31) + 2Nap-FF (1)
Classification: Good Scattering



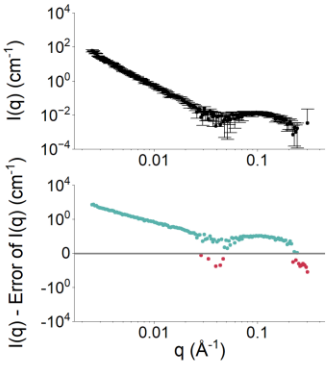
PBI-I (27) + 2Nap-V (2)
Classification: Good Scattering



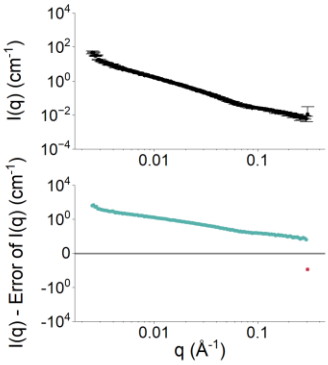
PBI-I (27) + 6MeO2Nap-FI (6)
Classification: Good Scattering



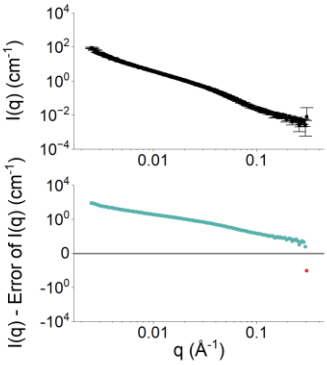
PBI-I (27) + 7MeO2Nap-FV (7)
Classification: Bad Scattering



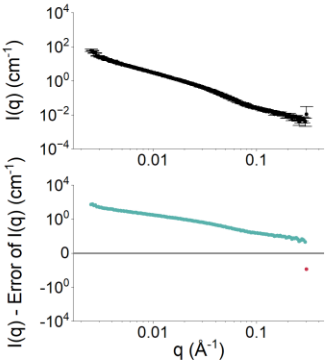
PBI-I (27) + C6-NDI-A (43)
Classification: Good Scattering



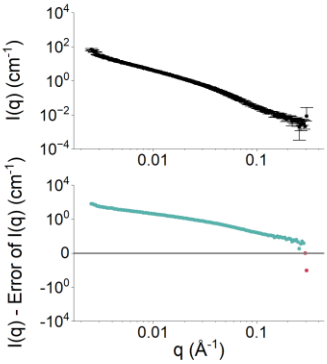
PBI-I (27) + C6-NDI-F (44)
Classification: Good Scattering



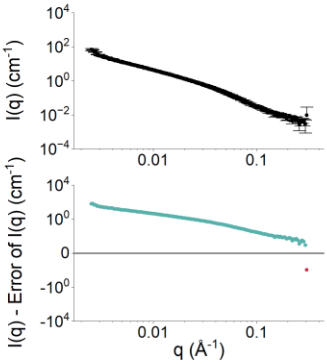
PBI-I (27) + C6-NDI-L (45)
Classification: Good Scattering



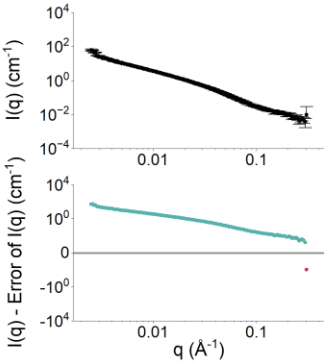
PBI-I (27) + NDI-A (13)
Classification: Good Scattering



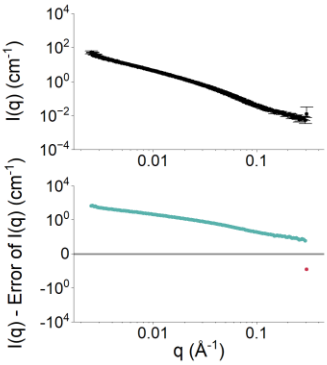
PBI-I (27) + NDI-F (9)
Classification: Good Scattering



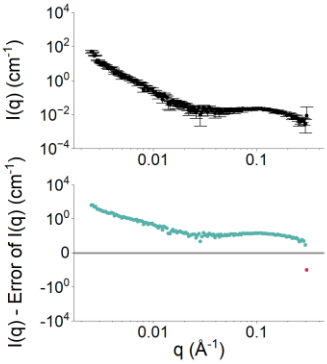
PBI-I (27) + NDI-G (11)
Classification: Good Scattering



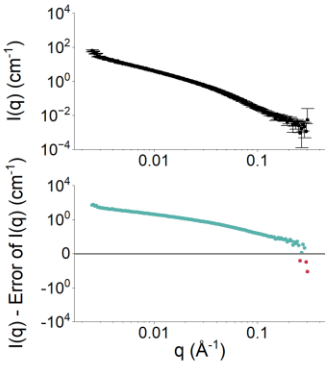
PBI-I (27) + NDI-GF (12)
Classification: Good Scattering



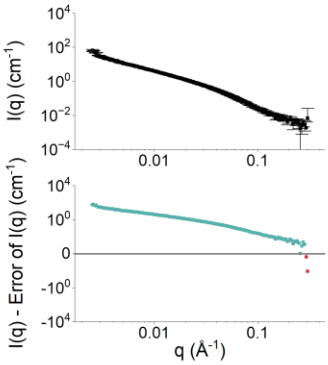
PBI-I (27) + NDI-H (60)
Classification: Good Scattering



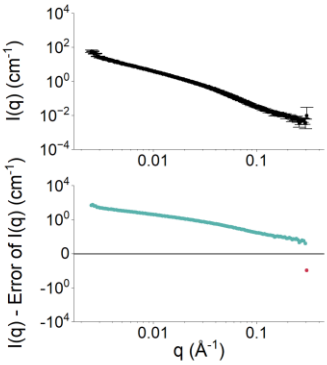
PBI-I (27) + NDI-I (14)
Classification: Good Scattering



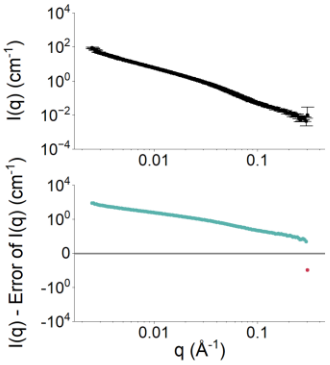
PBI-I (27) + NDI-L (15)
Classification: Good Scattering



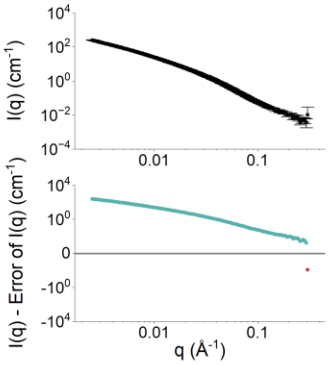
PBI-I (27) + NDI-YI (49)
Classification: Good Scattering



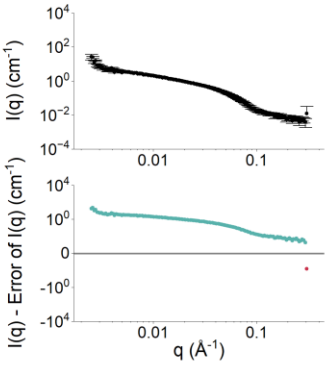
PBI-I (27) + PBI-A 1Py (50)
Classification: Good Scattering



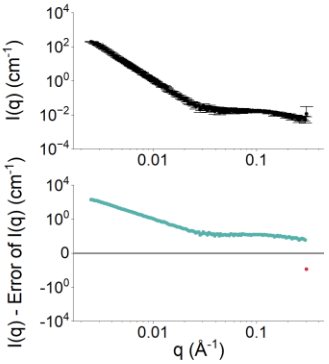
PBI-I (27) + PBI-E (23)
Classification: Good Scattering



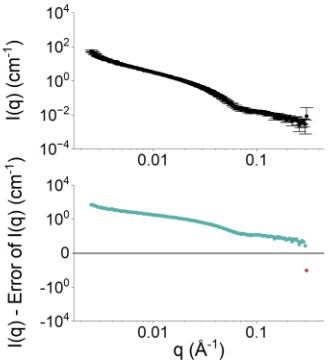
PBI-L (28) + 2Nap-FF (1)
Classification: Good Scattering



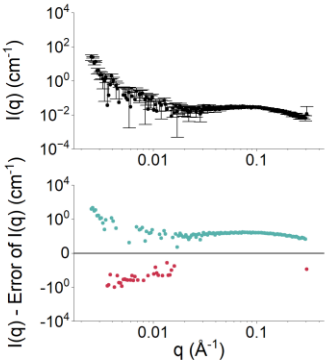
PBI-L (28) + 2Nap-FI (37)
Classification: Good Scattering



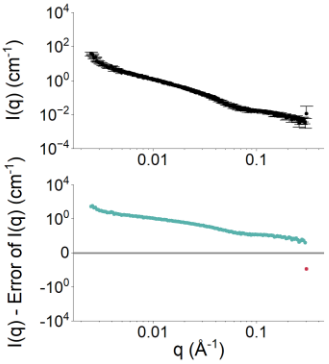
PBI-L (28) + 2Nap-IF (4)
Classification: Good Scattering



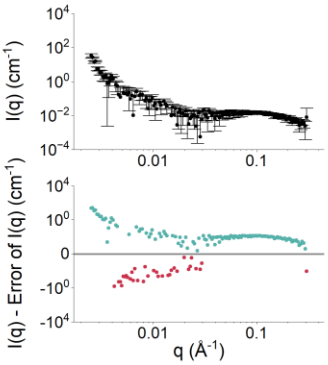
PBI-L (28) + 2Nap-V (2)
Classification: Bad Scattering



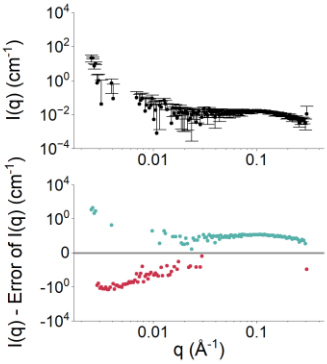
PBI-L (28) + 6MeO2Nap-FI (6)
Classification: Good Scattering



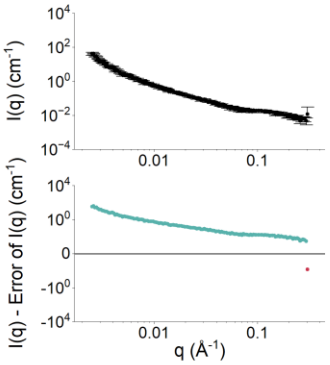
PBI-L (28) + 7MeO2Nap-FV (7)
Classification: Bad Scattering



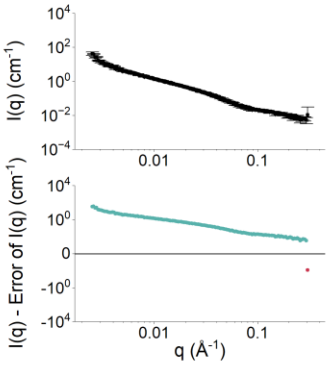
PBI-L (28) + C6-NDI-A (43)
Classification: Bad Scattering



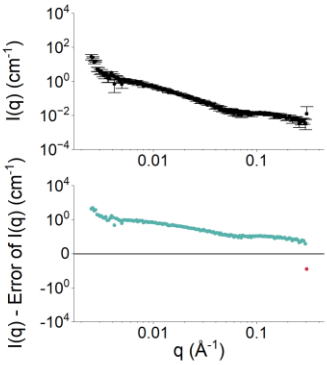
PBI-L (28) + C6-NDI-F (44)
Classification: Good Scattering



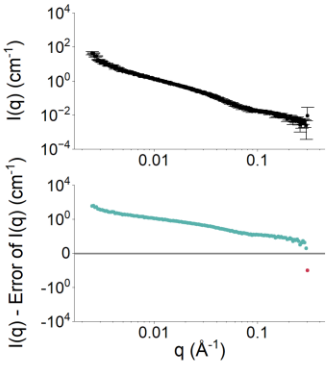
PBI-L (28) + C6-NDI-L (45)
Classification: Good Scattering



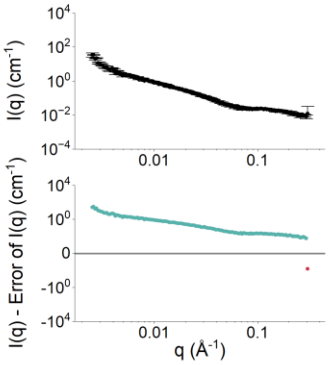
PBI-L (28) + NDI-A (13)
Classification: Good Scattering



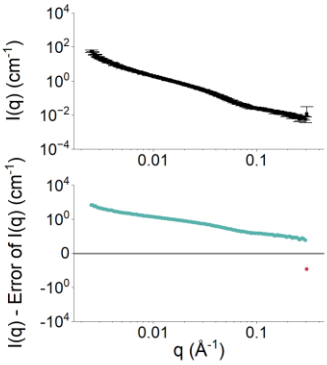
PBI-L (28) + NDI-F (9)
Classification: Good Scattering



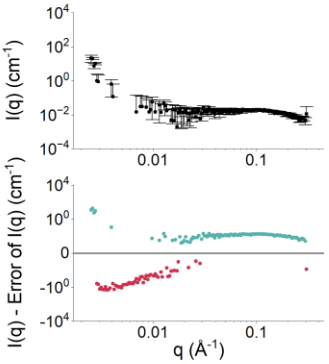
PBI-L (28) + NDI-G (11)
Classification: Good Scattering



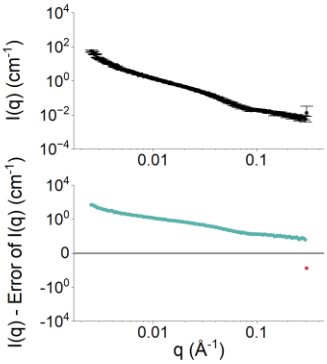
PBI-L (28) + NDI-GF (12)
Classification: Good Scattering



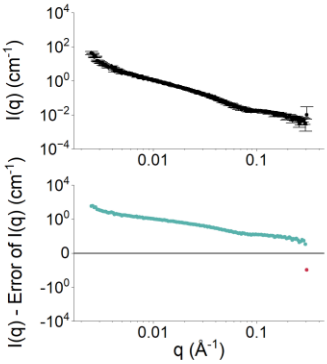
PBI-L (28) + NDI-H (60)
Classification: Bad Scattering



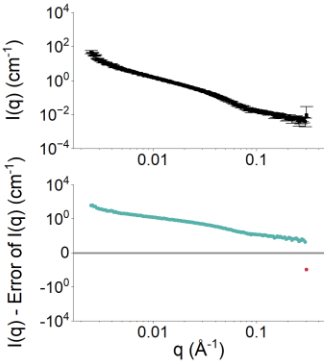
PBI-L (28) + NDI-L (15)
Classification: Good Scattering



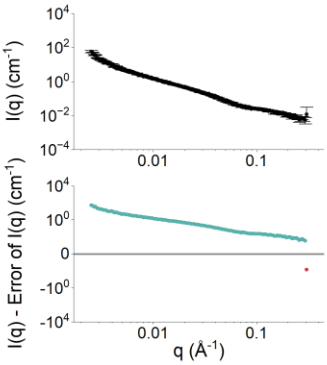
PBI-L (28) + NDI-M (16)
Classification: Good Scattering



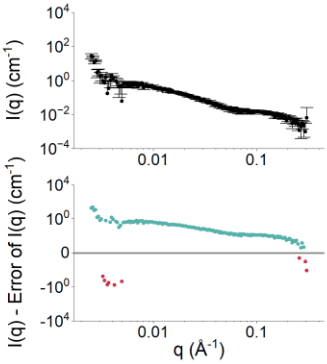
PBI-L (28) + NDI-V (17)
Classification: Good Scattering



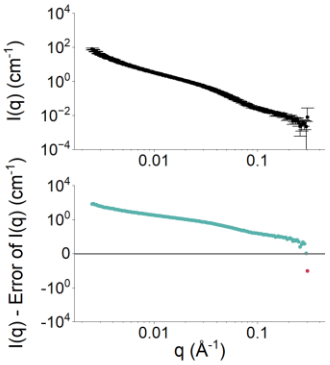
PBI-L (28) + NDI-W (18)
Classification: Good Scattering



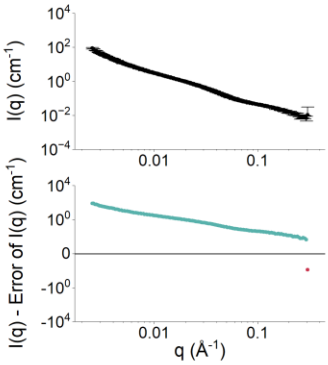
PBI-L (28) + NDI-Y (19)
Classification: Bad Scattering



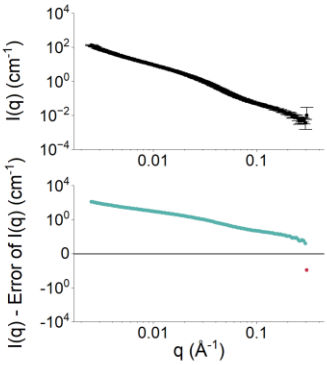
PBI-L (28) + NDI-YI (49)
Classification: Good Scattering



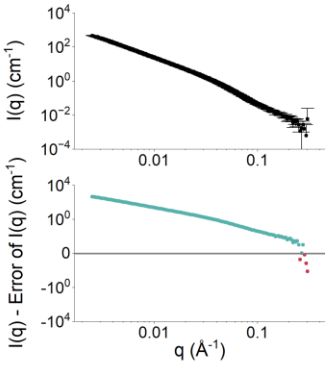
PBI-L (28) + PBI-A 1Py (50)
Classification: Good Scattering



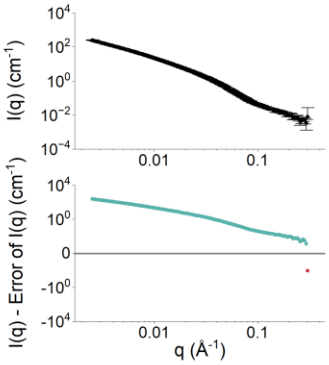
PBI-L (28) + PBI-D (21)
Classification: Good Scattering



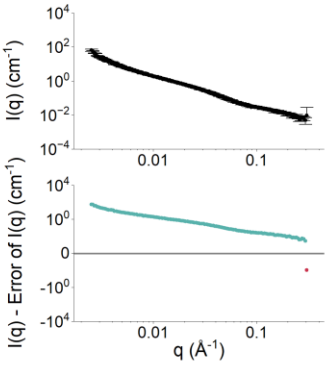
PBI-L (28) + PBI-DOPA (22)
Classification: Good Scattering



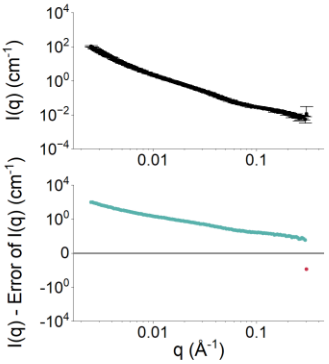
PBI-L (28) + PBI-E (23)
Classification: Good Scattering



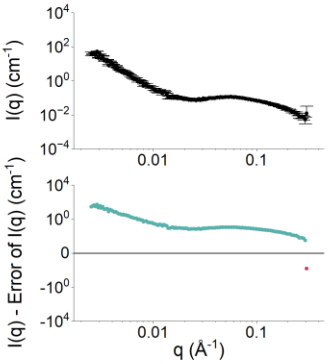
PBI-L (28) + PBI-F (24)
Classification: Good Scattering



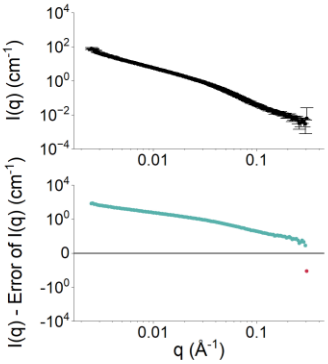
PBI-L (28) + PBI-G (25)
Classification: Good Scattering



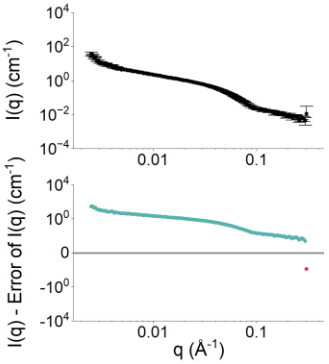
PBI-L (28) + PBI-H (26)
Classification: Good Scattering



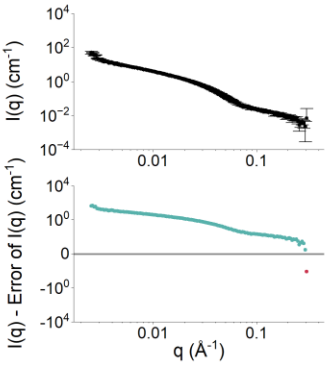
PBI-L (28) + PBI-I (27)
Classification: Good Scattering



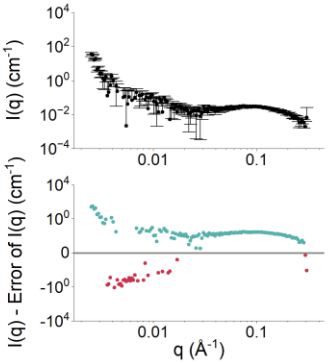
PBI-M (29) + 2Nap-FF (1)
Classification: Good Scattering



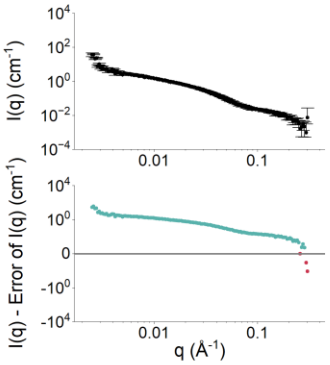
PBI-M (29) + 2Nap-IF (4)
Classification: Good Scattering



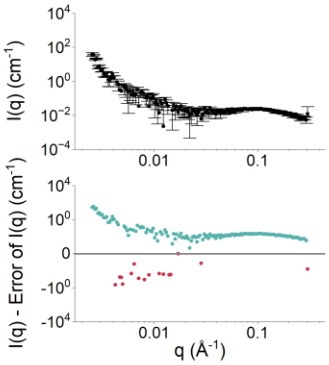
PBI-M (29) + 2Nap-V (2)
Classification: Bad Scattering



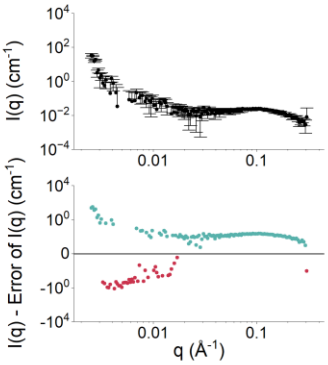
PBI-M (29) + 6MeO2Nap-FI (6)
Classification: Good Scattering



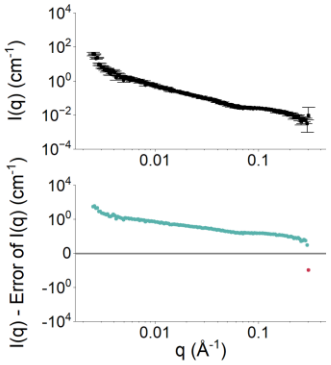
PBI-M (29) + 7MeO2Nap-FV (7)
Classification: Bad Scattering



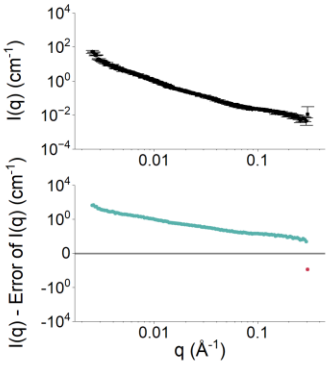
PBI-M (29) + C6-NDI-A (43)
Classification: Bad Scattering



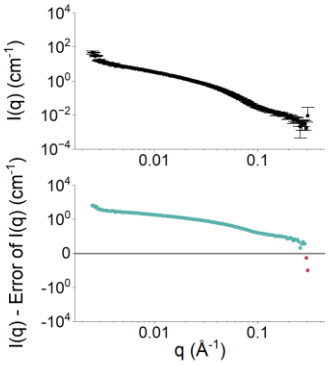
PBI-M (29) + C6-NDI-F (44)
Classification: Good Scattering



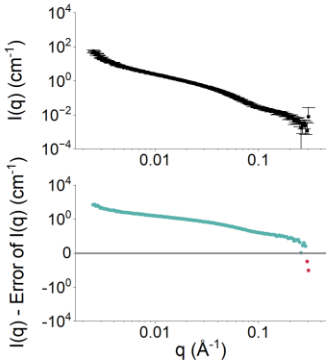
PBI-M (29) + C6-NDI-L (45)
Classification: Good Scattering



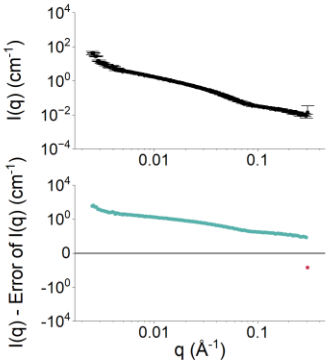
PBI-M (29) + NDI-A (13)
Classification: Good Scattering



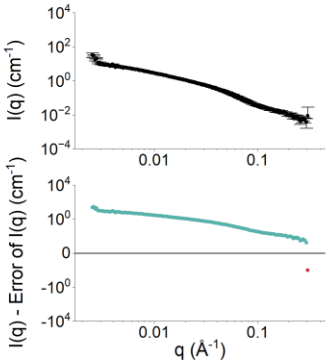
PBI-M (29) + NDI-F (9)
Classification: Good Scattering



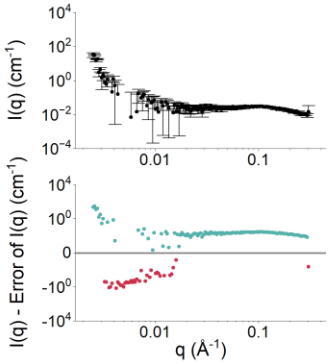
PBI-M (29) + NDI-G (11)
Classification: Good Scattering



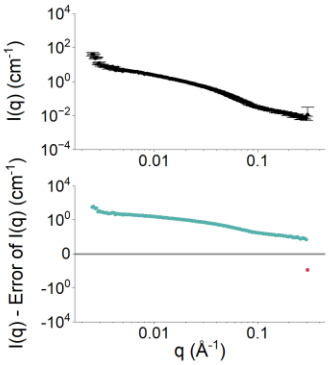
PBI-M (29) + NDI-GF (12)
Classification: Good Scattering



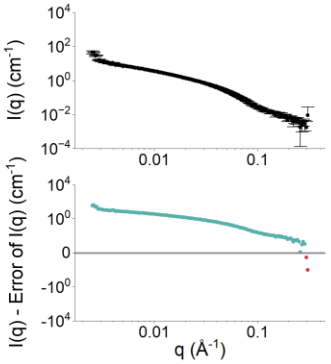
PBI-M (29) + NDI-H (60)
Classification: Bad Scattering



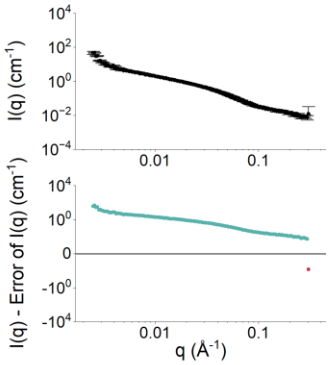
PBI-M (29) + NDI-I (14)
Classification: Good Scattering



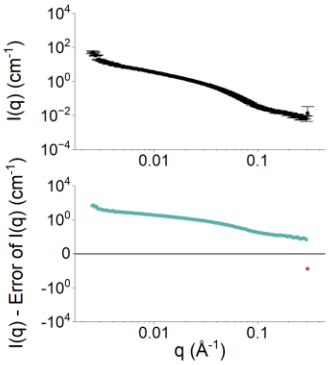
PBI-M (29) + NDI-L (15)
Classification: Good Scattering



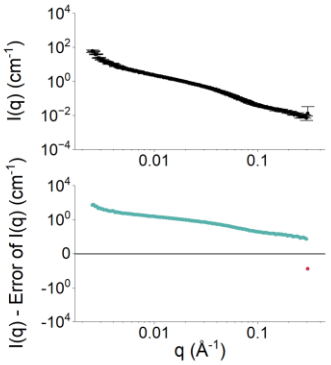
PBI-M (29) + NDI-M (16)
Classification: Good Scattering



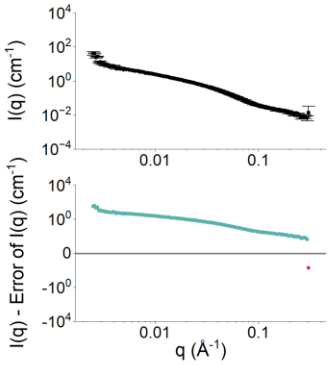
PBI-M (29) + NDI-V (17)
Classification: Good Scattering



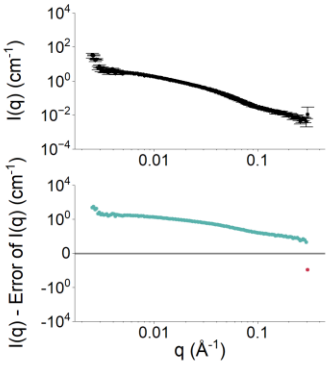
PBI-M (29) + NDI-W (18)
Classification: Good Scattering



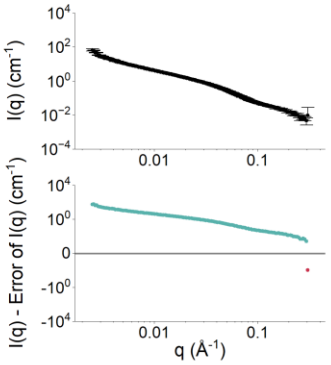
PBI-M (29) + NDI-Y (19)
Classification: Good Scattering



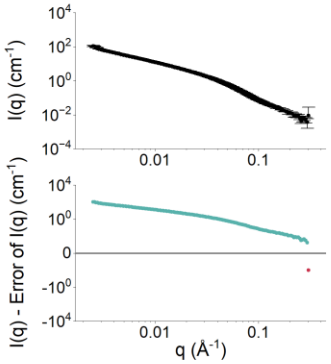
PBI-M (29) + NDI-YI (49)
Classification: Good Scattering



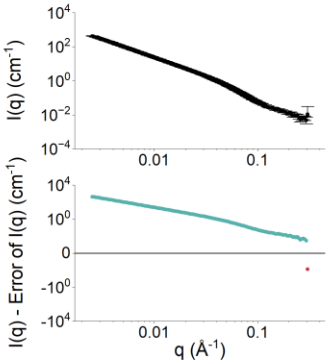
PBI-M (29) + PBI-A 1Py (50)
Classification: Good Scattering



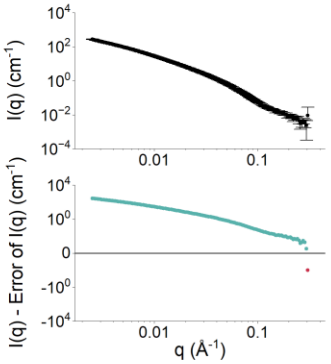
PBI-M (29) + PBI-D (21)
Classification: Good Scattering



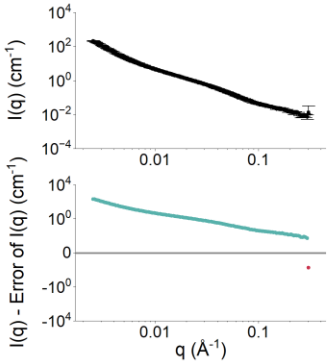
PBI-M (29) + PBI-DOPA (22)
Classification: Good Scattering



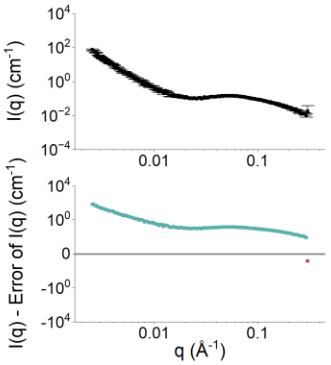
PBI-M (29) + PBI-E (23)
Classification: Good Scattering



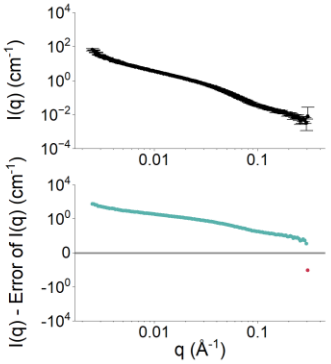
PBI-M (29) + PBI-G (25)
Classification: Good Scattering



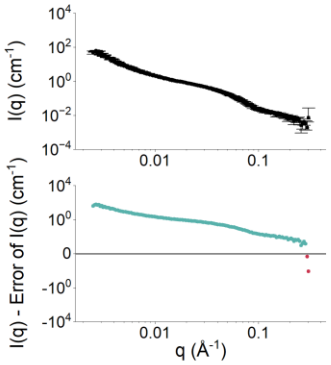
PBI-M (29) + PBI-H (26)
Classification: Good Scattering



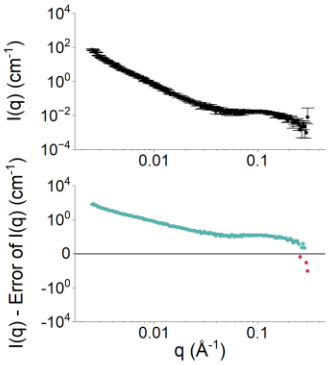
PBI-M (29) + PBI-L (28)
Classification: Good Scattering



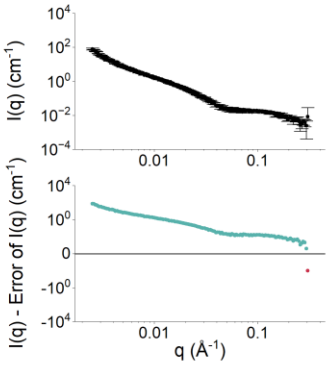
PBI-S (30) + 2Nap-FF (1)
Classification: Good Scattering

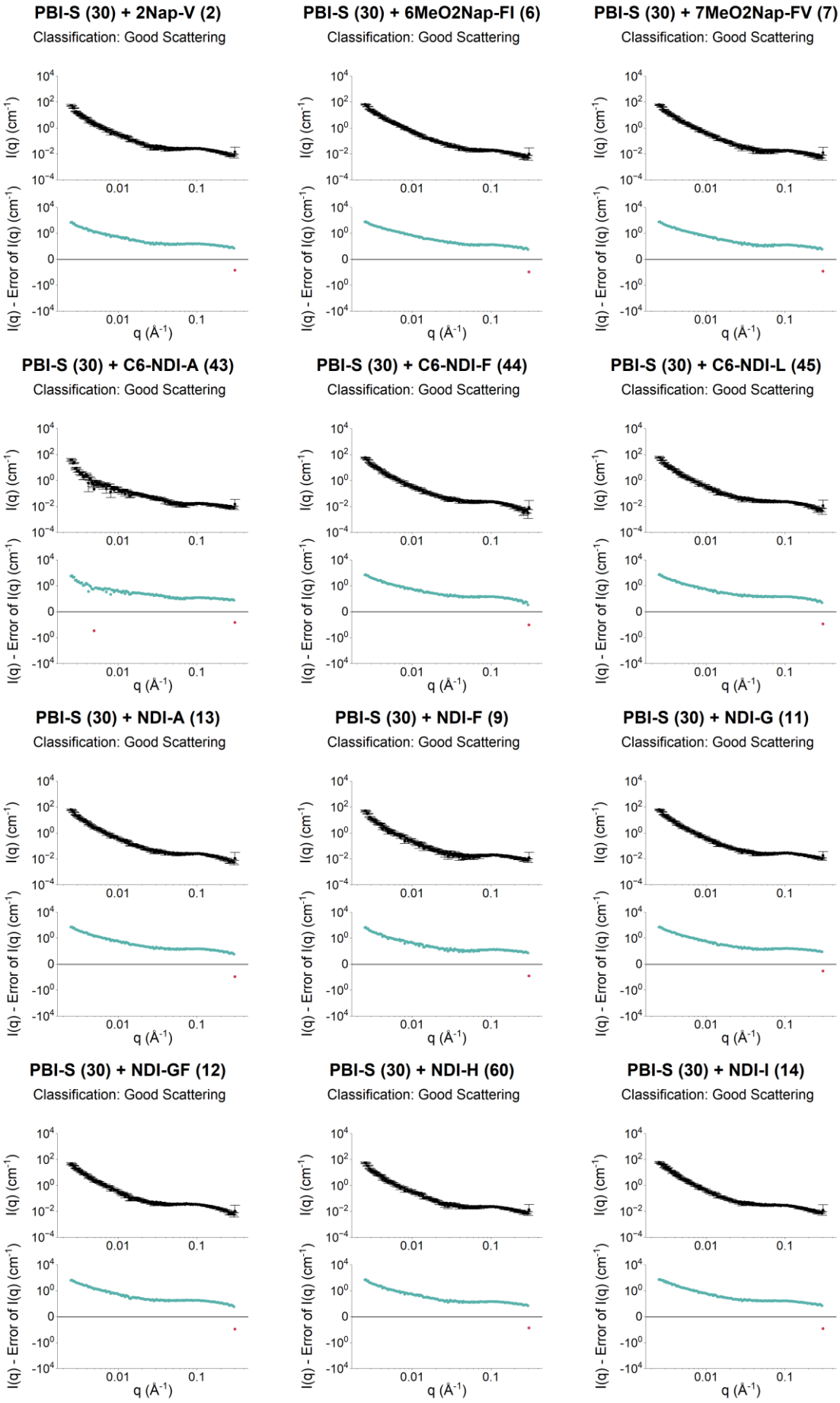


PBI-S (30) + 2Nap-FI (37)
Classification: Good Scattering

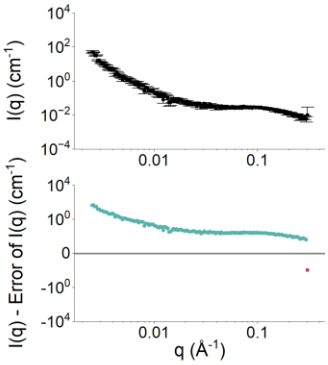


PBI-S (30) + 2Nap-IF (4)
Classification: Good Scattering

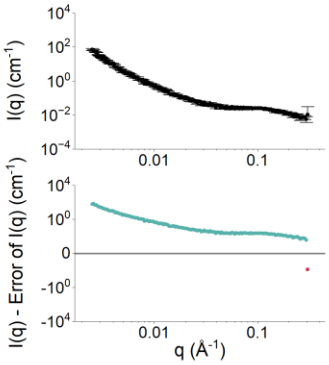



PBI-S (30) + 2Nap-V (2)
Classification: Good Scattering
PBI-S (30) + 6MeO2Nap-FI (6)
Classification: Good Scattering
PBI-S (30) + 7MeO2Nap-FV (7)
Classification: Good Scattering
PBI-S (30) + C6-NDI-A (43)
Classification: Good Scattering
PBI-S (30) + C6-NDI-F (44)
Classification: Good Scattering
PBI-S (30) + C6-NDI-L (45)
Classification: Good Scattering
PBI-S (30) + NDI-A (13)
Classification: Good Scattering
PBI-S (30) + NDI-F (9)
Classification: Good Scattering
PBI-S (30) + NDI-G (11)
Classification: Good Scattering
PBI-S (30) + NDI-GF (12)
Classification: Good Scattering
PBI-S (30) + NDI-H (60)
Classification: Good Scattering
PBI-S (30) + NDI-I (14)
Classification: Good Scattering
I(q) (cm-1)
I(q) - Error of I(q) (cm-1)
q (Å-1)

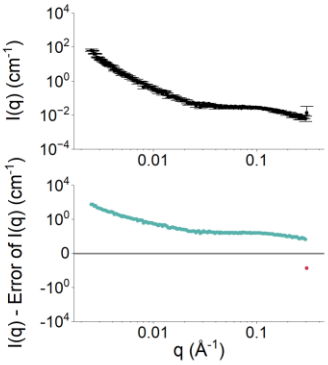
PBI-S (30) + NDI-L (15)
Classification: Good Scattering



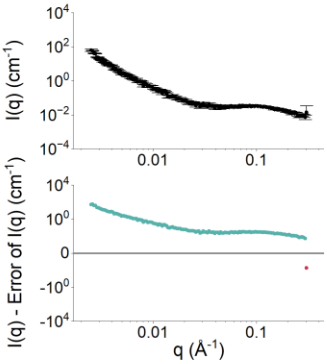
PBI-S (30) + NDI-M (16)
Classification: Good Scattering



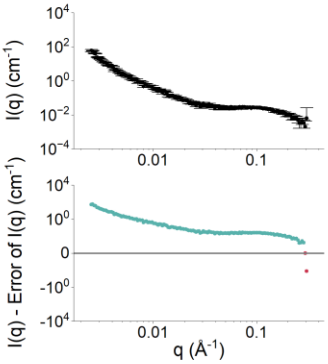
PBI-S (30) + NDI-V (17)
Classification: Good Scattering



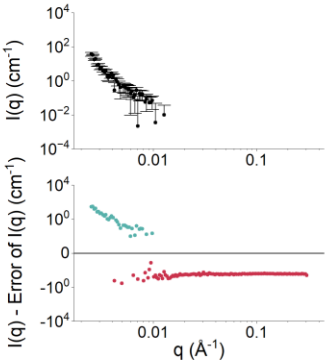
PBI-S (30) + NDI-W (18)
Classification: Good Scattering



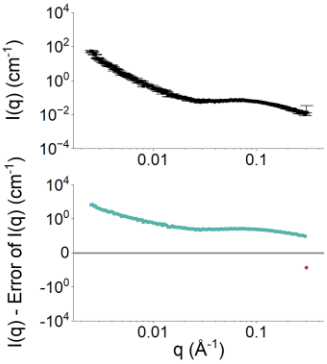
PBI-S (30) + NDI-Y (19)
Classification: Good Scattering



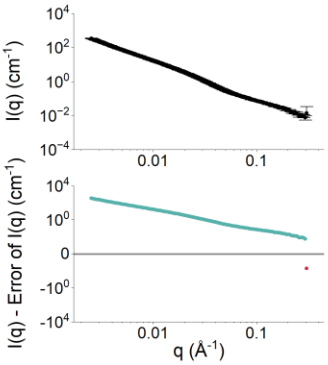
PBI-S (30) + NDI-YI (49)
Classification: Bad Scattering



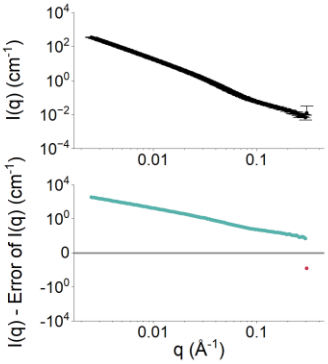
PBI-S (30) + PBI-A 1Py (50)
Classification: Good Scattering



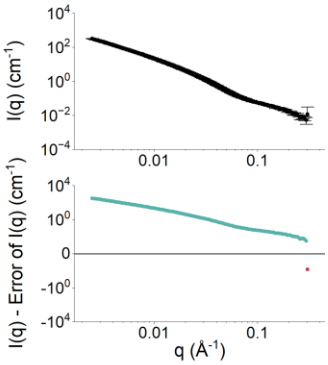
PBI-S (30) + PBI-D (21)
Classification: Good Scattering



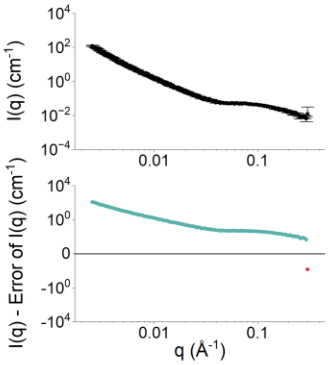
PBI-S (30) + PBI-DOPA (22)
Classification: Good Scattering



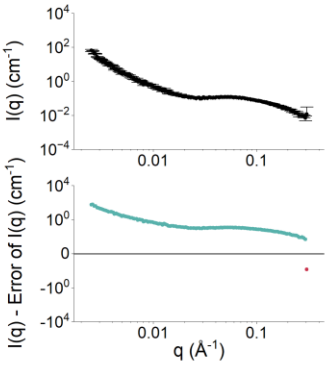
PBI-S (30) + PBI-E (23)
Classification: Good Scattering



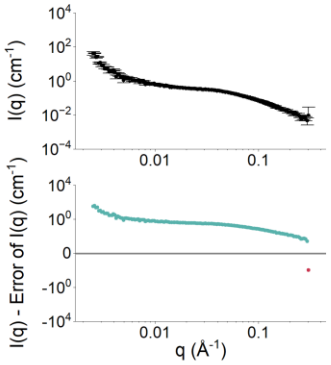
PBI-S (30) + PBI-G (25)
Classification: Good Scattering



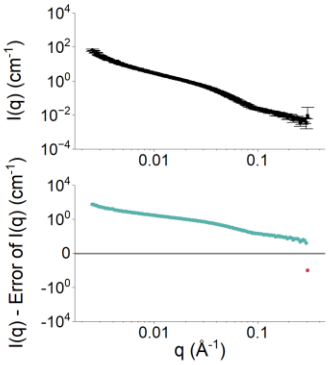
PBI-S (30) + PBI-H (26)
Classification: Good Scattering



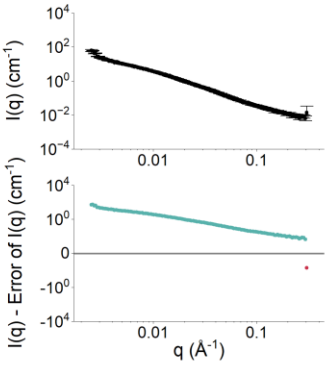
ThNap-FF (34) + PBI-H (26)
Classification: Good Scattering



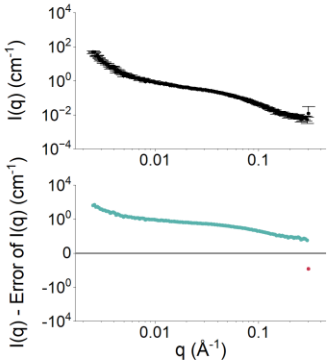
ThNap-FF (34) + PBI-L (28)
Classification: Good Scattering



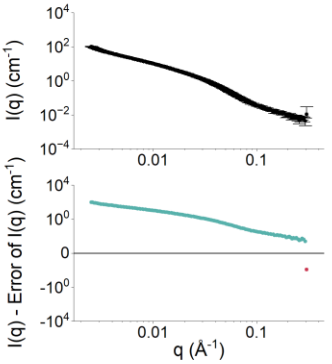
ThNap-FF (34) + PBI-M (29)
Classification: Good Scattering



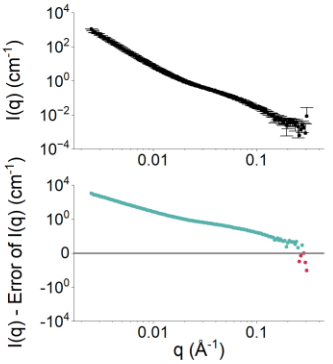
ThNap-FF (34) + PBI-S (30)
Classification: Good Scattering



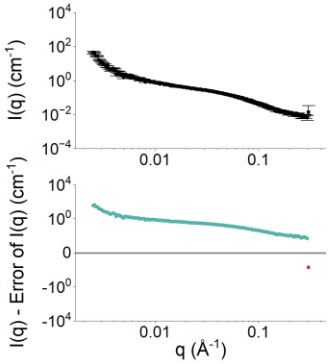
ThNap-FF (34) + PBI-V (31)
Classification: Good Scattering



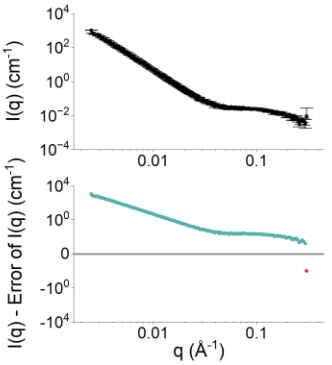
ThNap-FF (34) + PBI-W (32)
Classification: Good Scattering



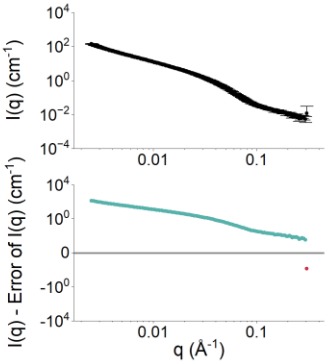
ThNap-FF (34) + PBI-Y (33)
Classification: Good Scattering



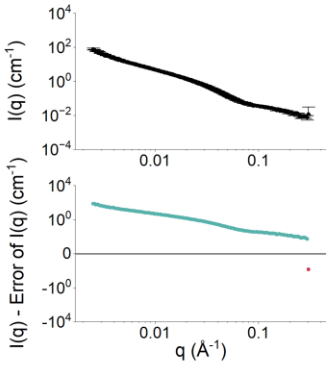
PBI-V (31) + 2Nap-FI (37)
Classification: Good Scattering



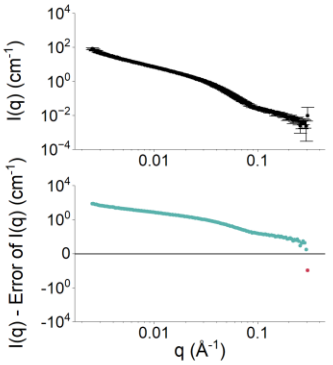
PBI-V (31) + 2Nap-IF (4)
Classification: Good Scattering



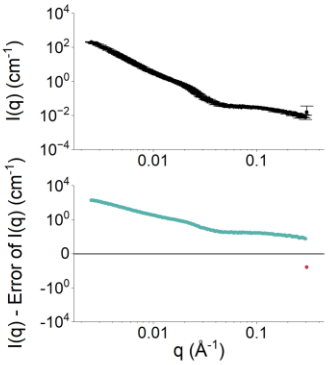
PBI-V (31) + 2Nap-V (2)
Classification: Good Scattering



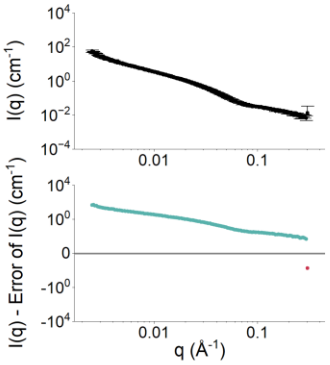
PBI-V (31) + 6MeO2Nap-FI (6)
Classification: Good Scattering



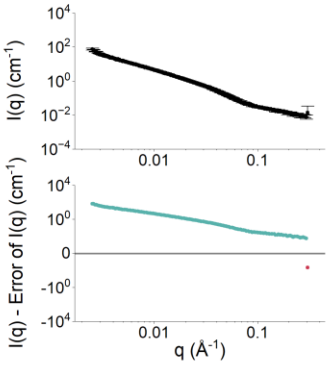
PBI-V (31) + 7MeO2Nap-FV (7)
Classification: Good Scattering



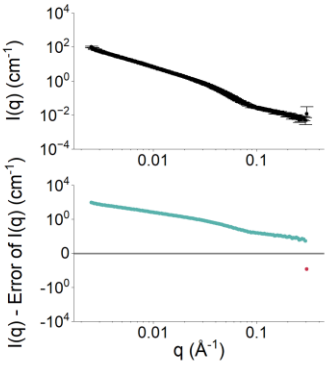
PBI-V (31) + C6-NDI-A (43)
Classification: Good Scattering



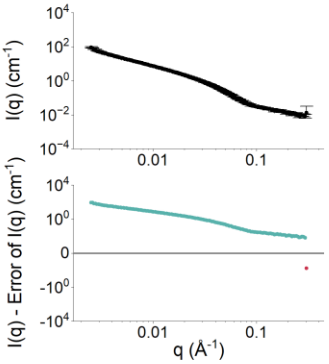
PBI-V (31) + C6-NDI-F (44)
Classification: Good Scattering



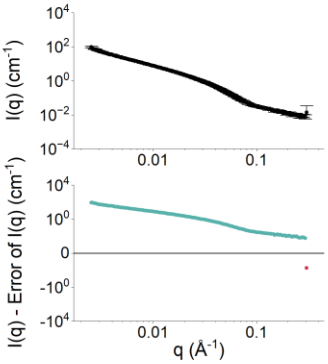
PBI-V (31) + C6-NDI-L (45)
Classification: Good Scattering



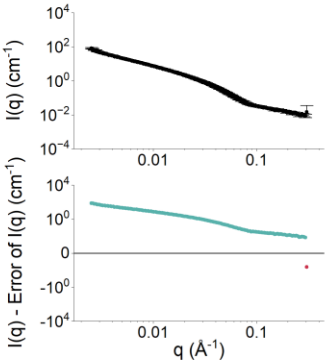
PBI-V (31) + NDI-A (13)
Classification: Good Scattering



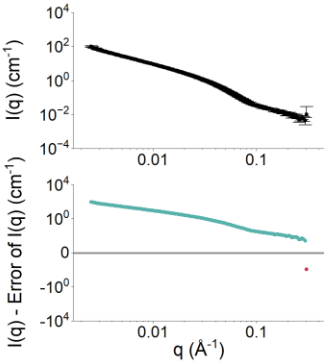
PBI-V (31) + NDI-F (9)
Classification: Good Scattering



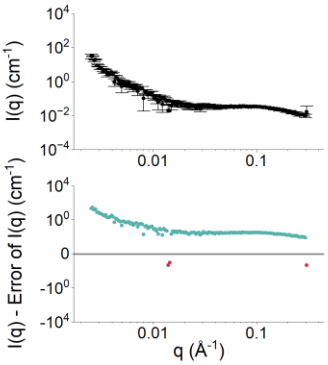
PBI-V (31) + NDI-G (11)
Classification: Good Scattering



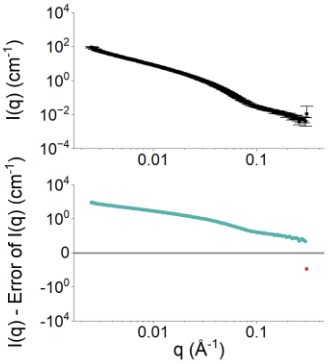
PBI-V (31) + NDI-GF (12)
Classification: Good Scattering



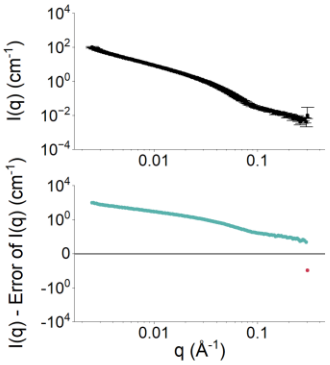
PBI-V (31) + NDI-H (60)
Classification: Good Scattering



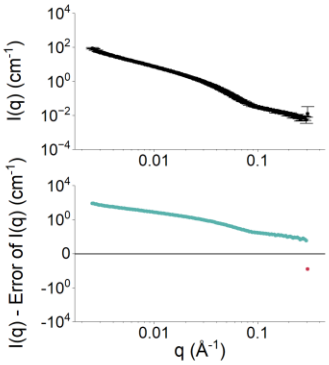
PBI-V (31) + NDI-I (14)
Classification: Good Scattering



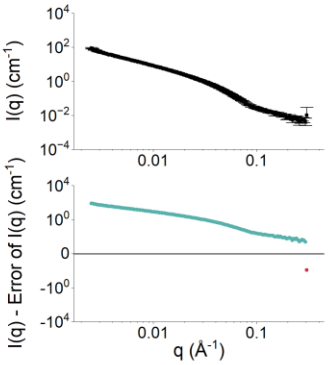
PBI-V (31) + NDI-L (15)
Classification: Good Scattering



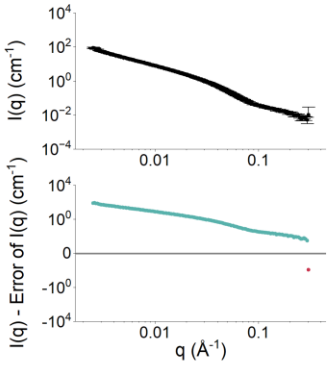
PBI-V (31) + NDI-M (16)
Classification: Good Scattering



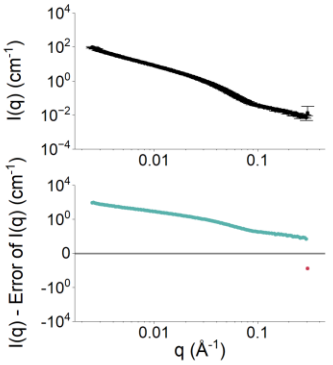
PBI-V (31) + NDI-V (17)
Classification: Good Scattering



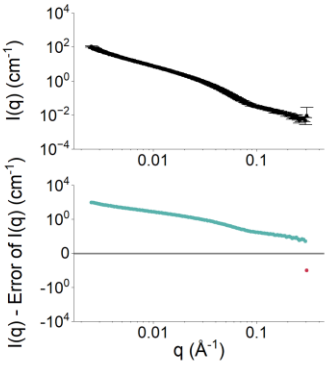
PBI-V (31) + NDI-W (18)
Classification: Good Scattering



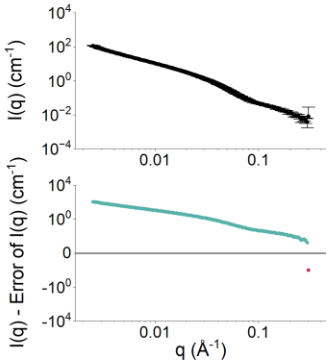
PBI-V (31) + NDI-Y (19)
Classification: Good Scattering



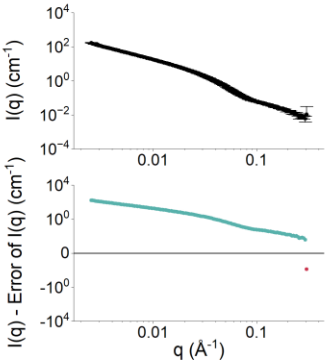
PBI-V (31) + NDI-YI (49)
Classification: Good Scattering



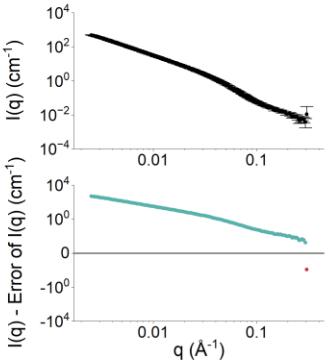
PBI-V (31) + PBI-A 1Py (50)
Classification: Good Scattering



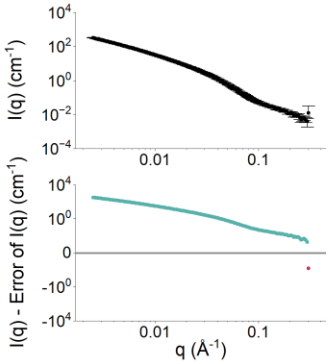
PBI-V (31) + PBI-D (21)
Classification: Good Scattering



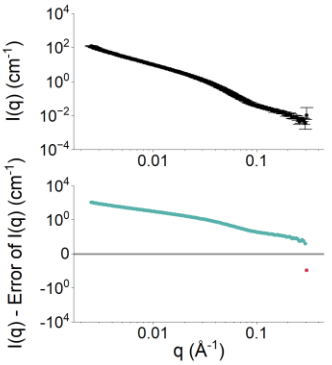
PBI-V (31) + PBI-DOPA (22)
Classification: Good Scattering



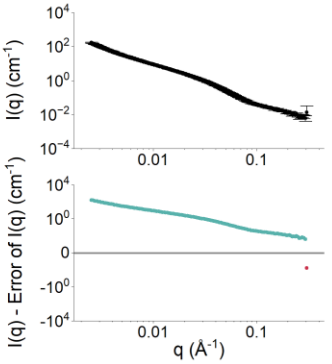
PBI-V (31) + PBI-E (23)
Classification: Good Scattering



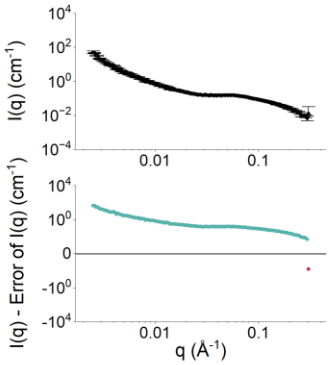
PBI-V (31) + PBI-F (24)
Classification: Good Scattering



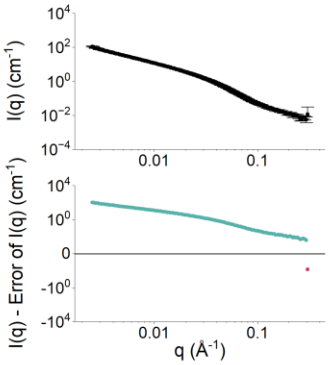
PBI-V (31) + PBI-G (25)
Classification: Good Scattering



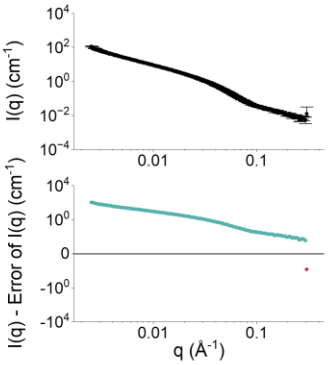
PBI-V (31) + PBI-H (26)
Classification: Good Scattering



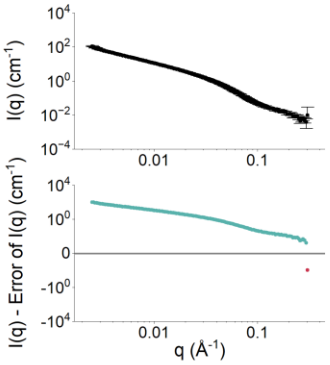
PBI-V (31) + PBI-I (27)
Classification: Good Scattering



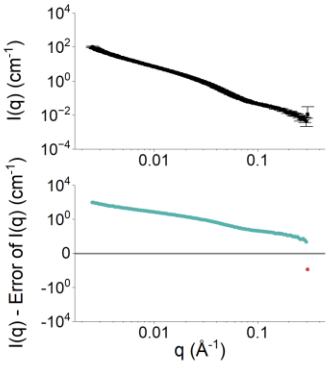
PBI-V (31) + PBI-L (28)
Classification: Good Scattering



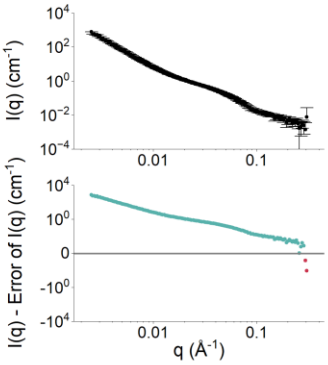
PBI-V (31) + PBI-M (29)
Classification: Good Scattering



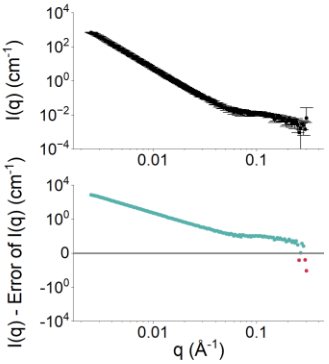
PBI-V (31) + PBI-S (30)
Classification: Good Scattering



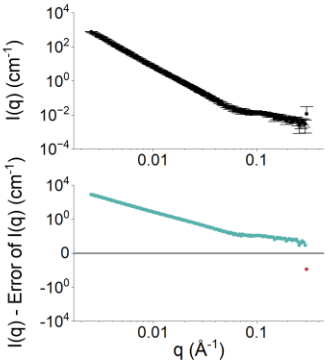
PBI-W (32) + 2Nap-FF (1)
Classification: Good Scattering



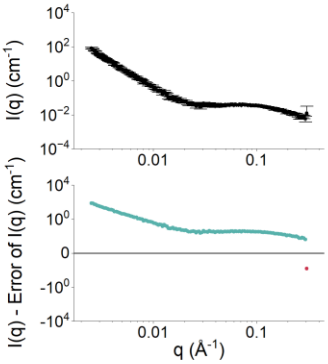
PBI-W (32) + 2Nap-FI (37)
Classification: Good Scattering



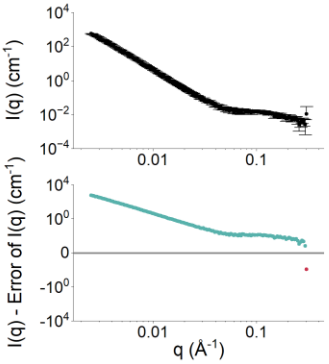
PBI-W (32) + 2Nap-IF (4)
Classification: Good Scattering



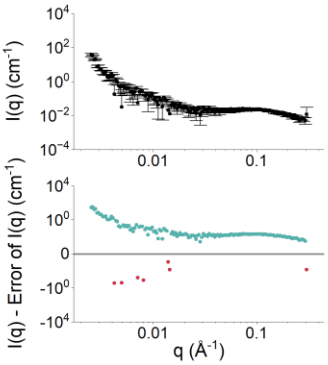
PBI-W (32) + 2Nap-V (2)
Classification: Good Scattering



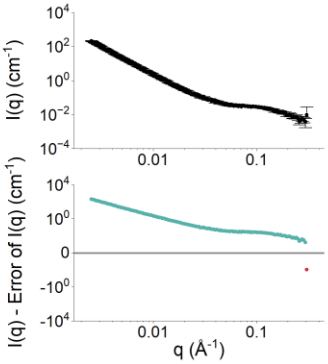
PBI-W (32) + 6MeO2Nap-FI (6)
Classification: Good Scattering



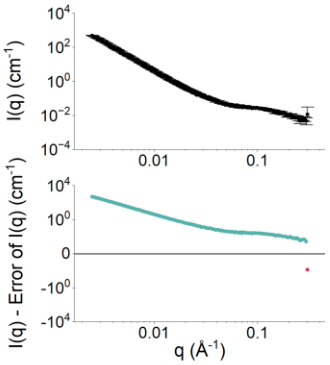
PBI-W (32) + 7MeO2Nap-FV (7)
Classification: Good Scattering



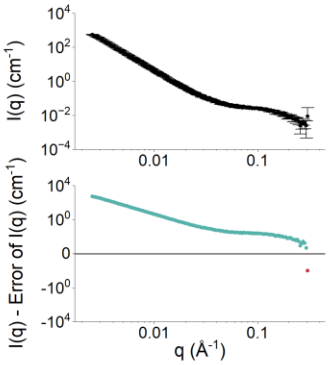
PBI-W (32) + C6-NDI-A (43)
Classification: Good Scattering



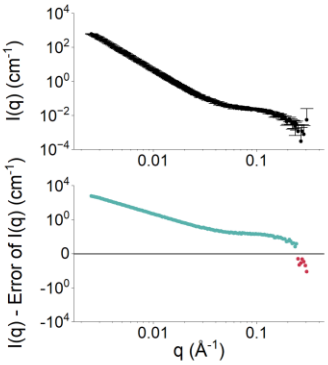
PBI-W (32) + C6-NDI-L (45)
Classification: Good Scattering



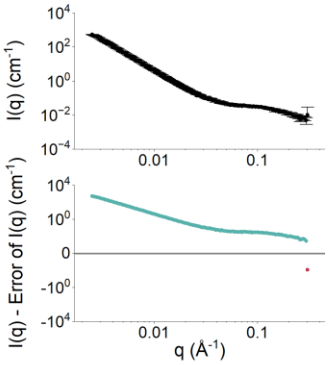
PBI-W (32) + NDI-A (13)
Classification: Good Scattering



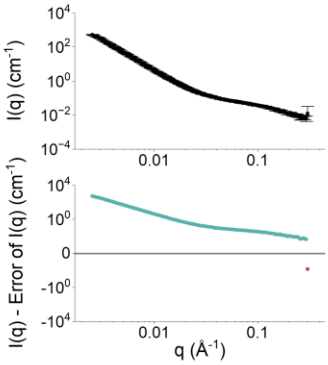
PBI-W (32) + NDI-F (9)
Classification: Good Scattering



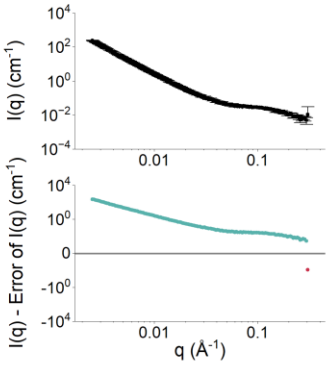
PBI-W (32) + NDI-G (11)
Classification: Good Scattering



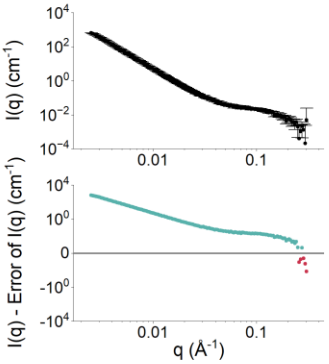
PBI-W (32) + NDI-GF (12)
Classification: Good Scattering



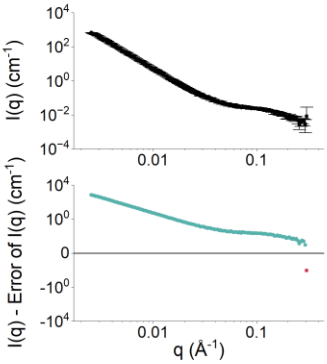
PBI-W (32) + NDI-H (60)
Classification: Good Scattering



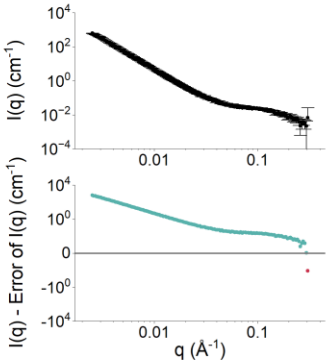
PBI-W (32) + NDI-I (14)
Classification: Good Scattering



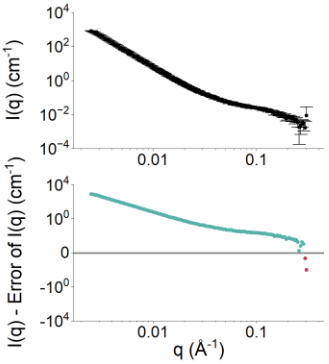
PBI-W (32) + NDI-L (15)
Classification: Good Scattering



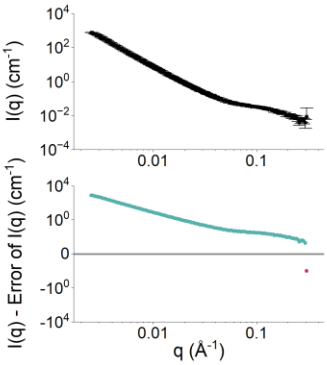
PBI-W (32) + NDI-M (16)
Classification: Good Scattering



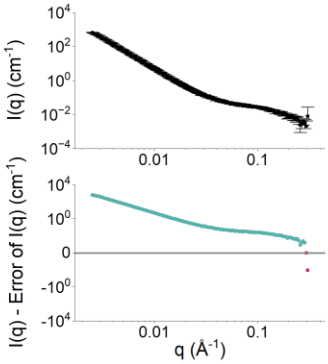
PBI-W (32) + NDI-V (17)
Classification: Good Scattering



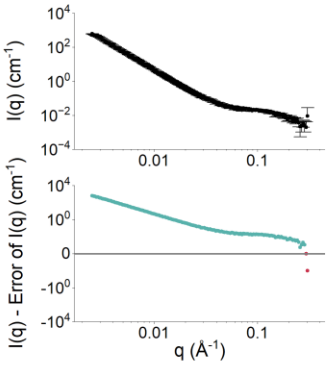
PBI-W (32) + NDI-W (18)
Classification: Good Scattering



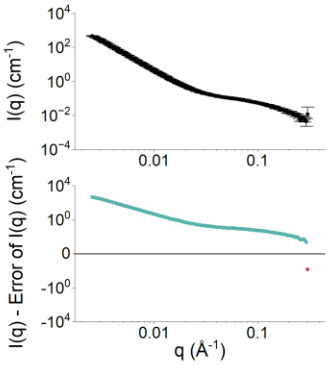
PBI-W (32) + NDI-Y (19)
Classification: Good Scattering



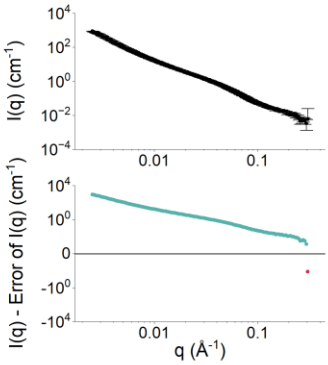
PBI-W (32) + NDI-YI (49)
Classification: Good Scattering



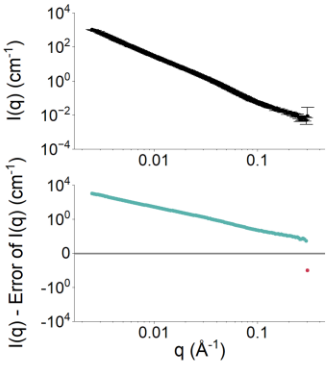
PBI-W (32) + PBI-A 1Py (50)
Classification: Good Scattering



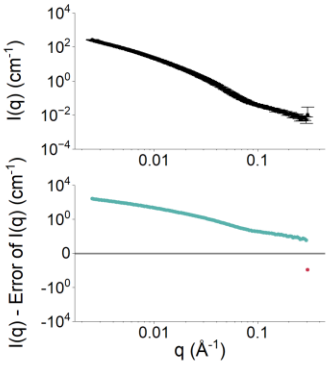
PBI-W (32) + PBI-D (21)
Classification: Good Scattering



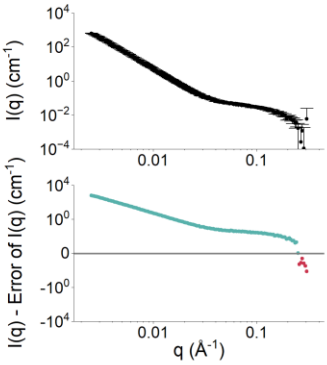
PBI-W (32) + PBI-DOPA (22)
Classification: Good Scattering



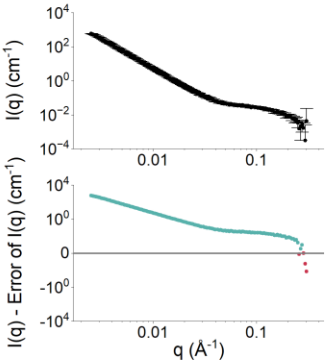
PBI-W (32) + PBI-E (23)
Classification: Good Scattering



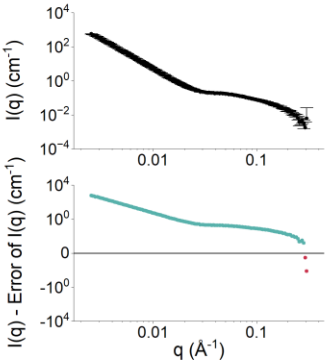
PBI-W (32) + PBI-F (24)
Classification: Good Scattering



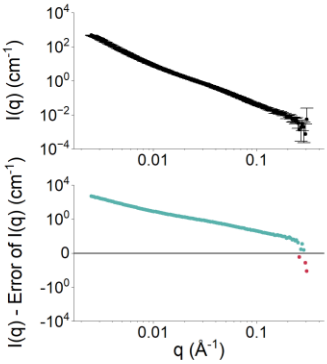
PBI-W (32) + PBI-G (25)
Classification: Good Scattering



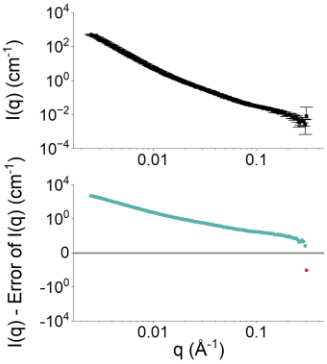
PBI-W (32) + PBI-H (26)
Classification: Good Scattering



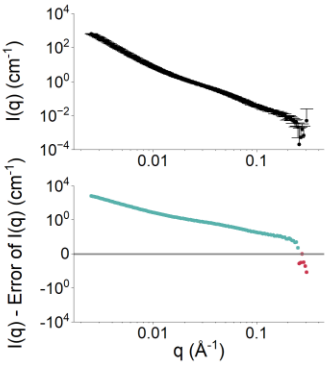
PBI-W (32) + PBI-I (27)
Classification: Good Scattering



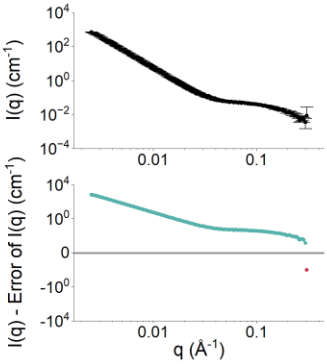
PBI-W (32) + PBI-L (28)
Classification: Good Scattering



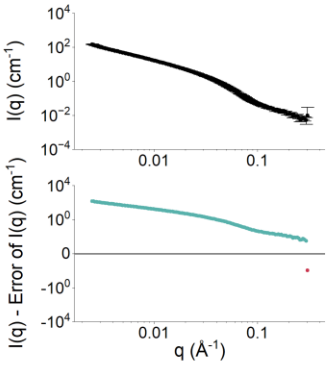
PBI-W (32) + PBI-M (29)
Classification: Good Scattering



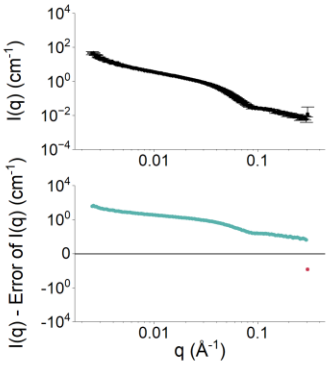
PBI-W (32) + PBI-S (30)
Classification: Good Scattering



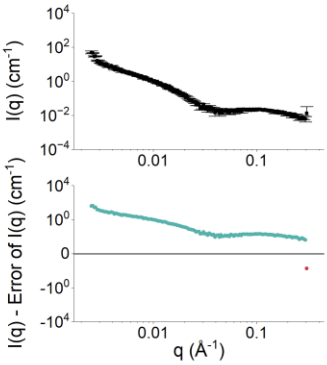
PBI-W (32) + PBI-V (31)
Classification: Good Scattering

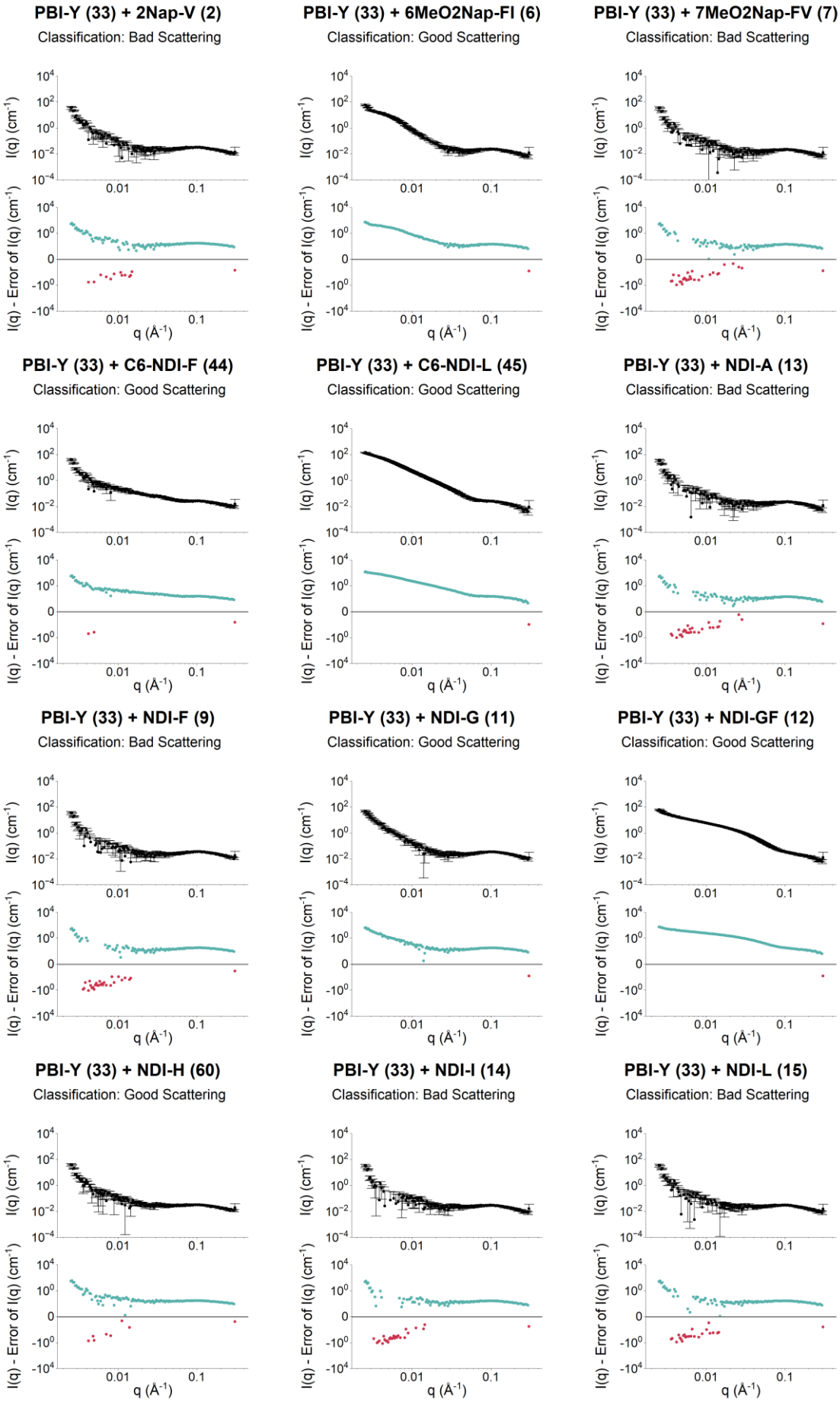


PBI-Y (33) + 2Nap-FF (1)
Classification: Good Scattering

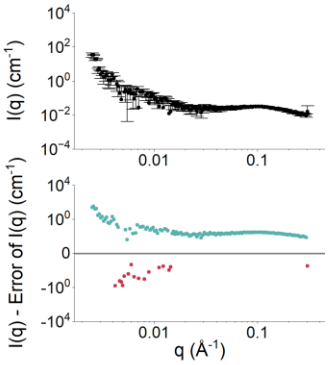


PBI-Y (33) + 2Nap-IF (4)
Classification: Good Scattering

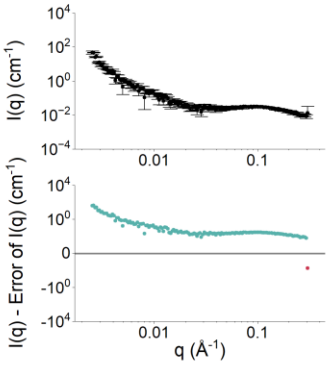



PBI-Y (33) + 2Nap-V (2)
Classification: Bad Scattering
PBI-Y (33) + 6MeO2Nap-FI (6)
Classification: Good Scattering
PBI-Y (33) + 7MeO2Nap-FV (7)
Classification: Bad Scattering
PBI-Y (33) + C6-NDI-F (44)
Classification: Good Scattering
PBI-Y (33) + C6-NDI-L (45)
Classification: Good Scattering
PBI-Y (33) + NDI-A (13)
Classification: Bad Scattering
PBI-Y (33) + NDI-F (9)
Classification: Bad Scattering
PBI-Y (33) + NDI-G (11)
Classification: Good Scattering
PBI-Y (33) + NDI-GF (12)
Classification: Good Scattering
PBI-Y (33) + NDI-H (60)
Classification: Good Scattering
PBI-Y (33) + NDI-I (14)
Classification: Bad Scattering
PBI-Y (33) + NDI-L (15)
Classification: Bad Scattering
I(q) (cm-1)
I(q) - Error of I(q) (cm-1)
q (Å-1)

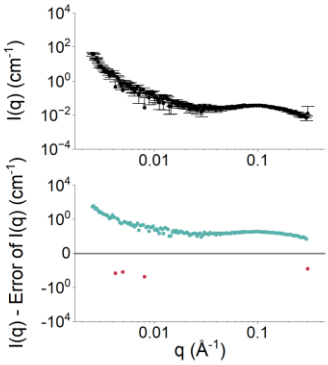
PBI-Y (33) + NDI-M (16)
Classification: Bad Scattering



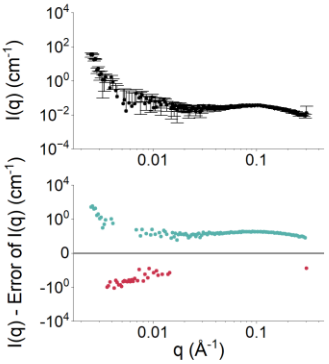
PBI-Y (33) + NDI-V (17)
Classification: Good Scattering



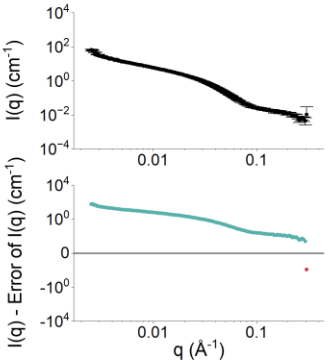
PBI-Y (33) + NDI-W (18)
Classification: Good Scattering



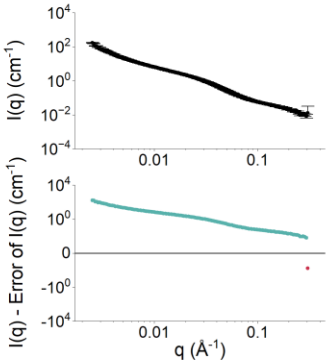
PBI-Y (33) + NDI-Y (19)
Classification: Bad Scattering



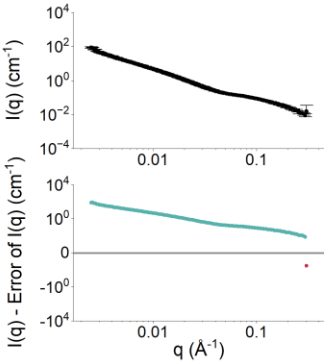
PBI-Y (33) + NDI-YI (49)
Classification: Good Scattering



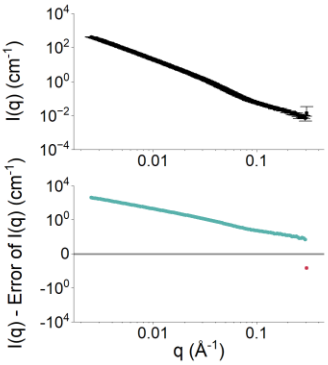
PBI-Y (33) + PBI-A 1Py (50)
Classification: Good Scattering



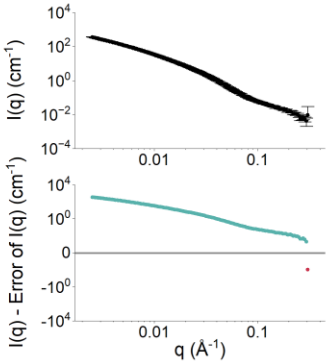
PBI-Y (33) + PBI-D (21)
Classification: Good Scattering



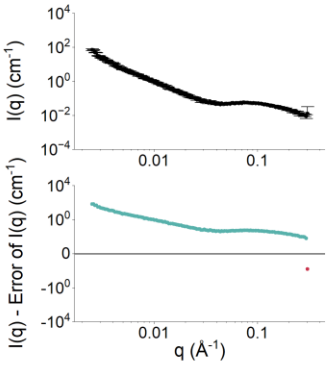
PBI-Y (33) + PBI-DOPA (22)
Classification: Good Scattering



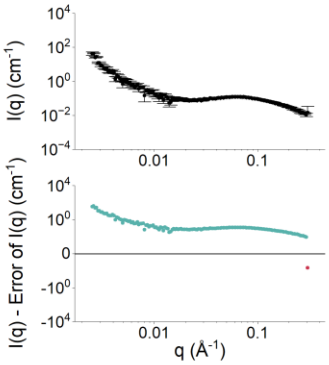
PBI-Y (33) + PBI-E (23)
Classification: Good Scattering



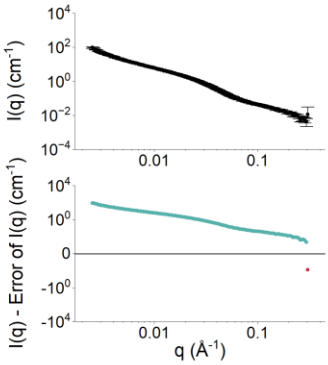
PBI-Y (33) + PBI-G (25)
Classification: Good Scattering



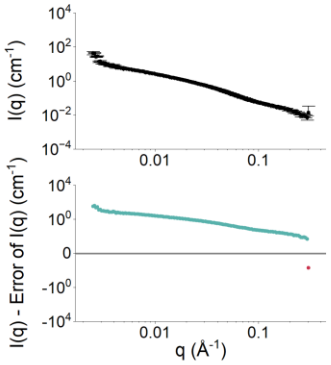
PBI-Y (33) + PBI-H (26)
Classification: Good Scattering



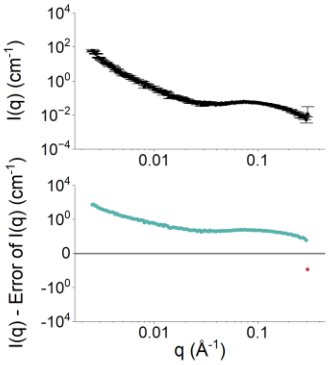
PBI-Y (33) + PBI-L (28)
Classification: Good Scattering



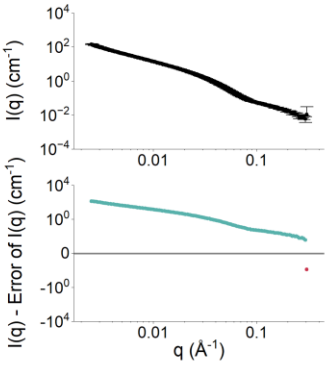
PBI-Y (33) + PBI-M (29)
Classification: Good Scattering



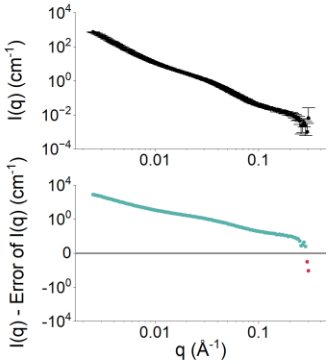
PBI-Y (33) + PBI-S (30)
Classification: Good Scattering



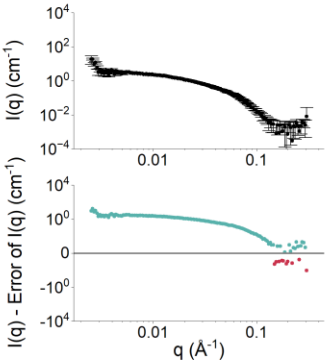
PBI-Y (33) + PBI-V (31)
Classification: Good Scattering



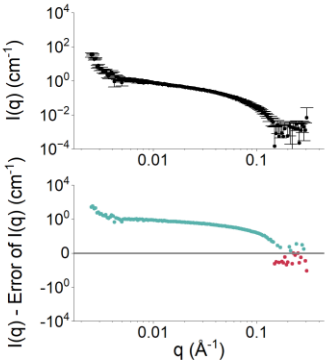
PBI-Y (33) + PBI-W (32)
Classification: Good Scattering



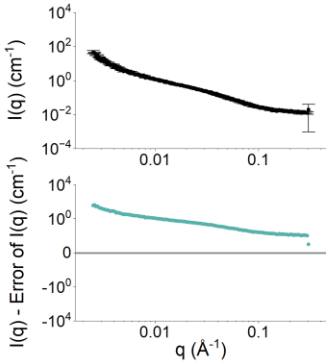
ThNap-FF (34) + 2Nap-FF (1)
Classification: Good Scattering



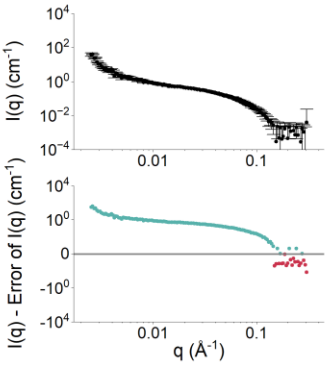
ThNap-FF (34) + C6-NDI-L (45)
Classification: Good Scattering



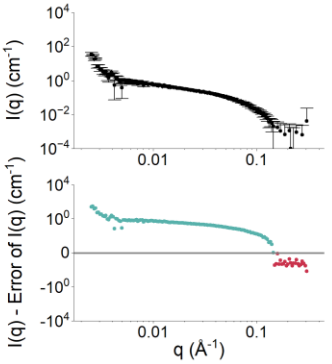
ThNap-FF (34) + NDI-GF (12)
Classification: Good Scattering



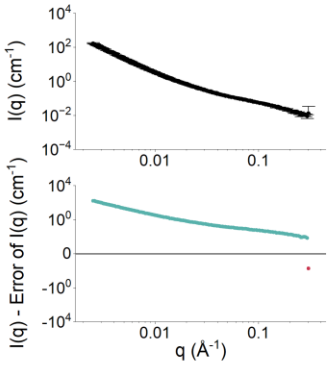
ThNap-FF (34) + NDI-I (14)
Classification: Good Scattering



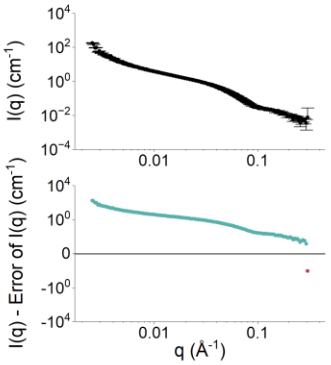
ThNap-FF (34) + NDI-YI (49)
Classification: Bad Scattering



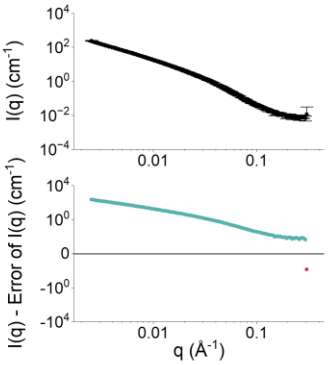
ThNap-FF (34) + PBI-A 1Py (50)
Classification: Good Scattering

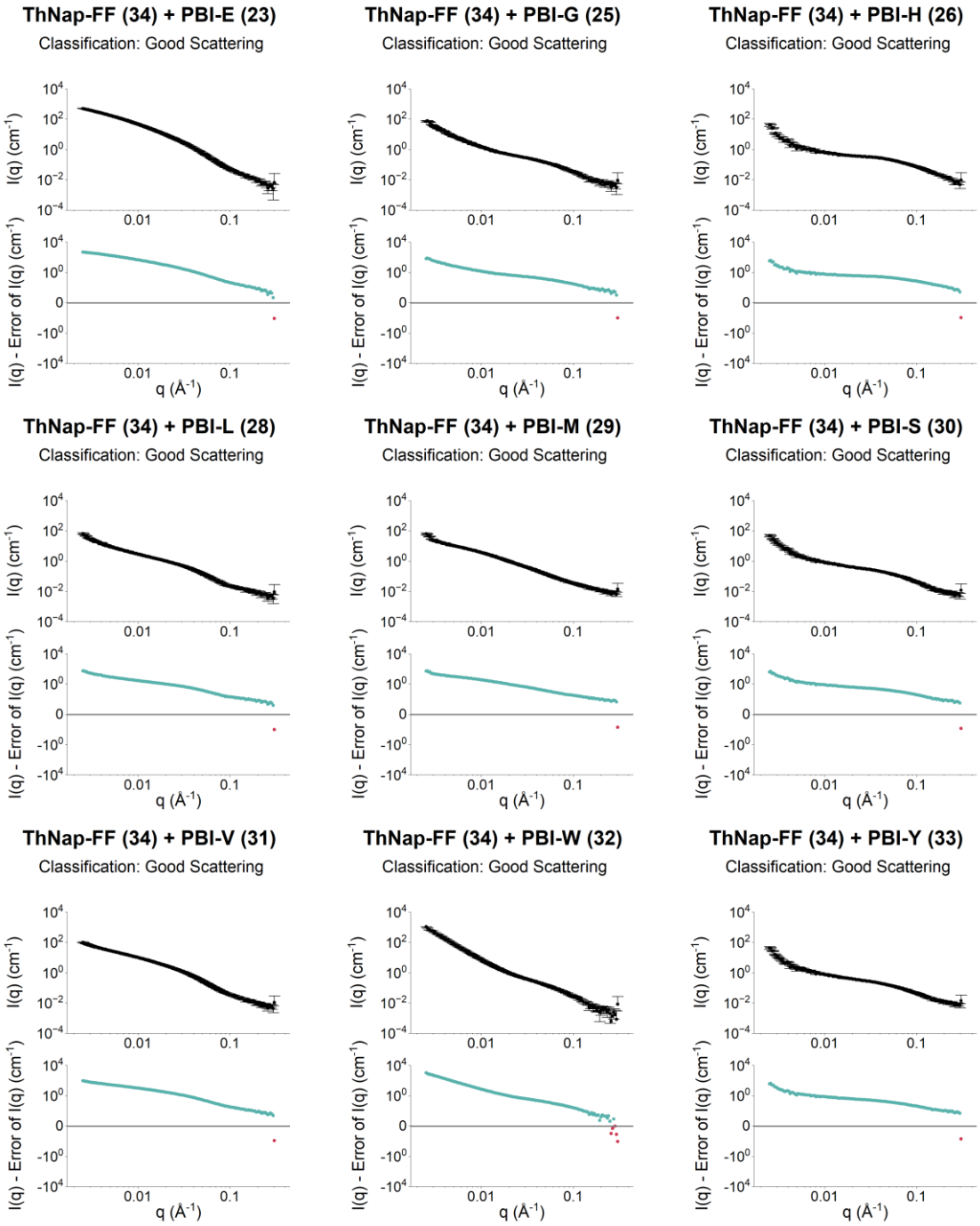


ThNap-FF (34) + PBI-D (21)
Classification: Good Scattering



ThNap-FF (34) + PBI-DOPA (22)
Classification: Good Scattering

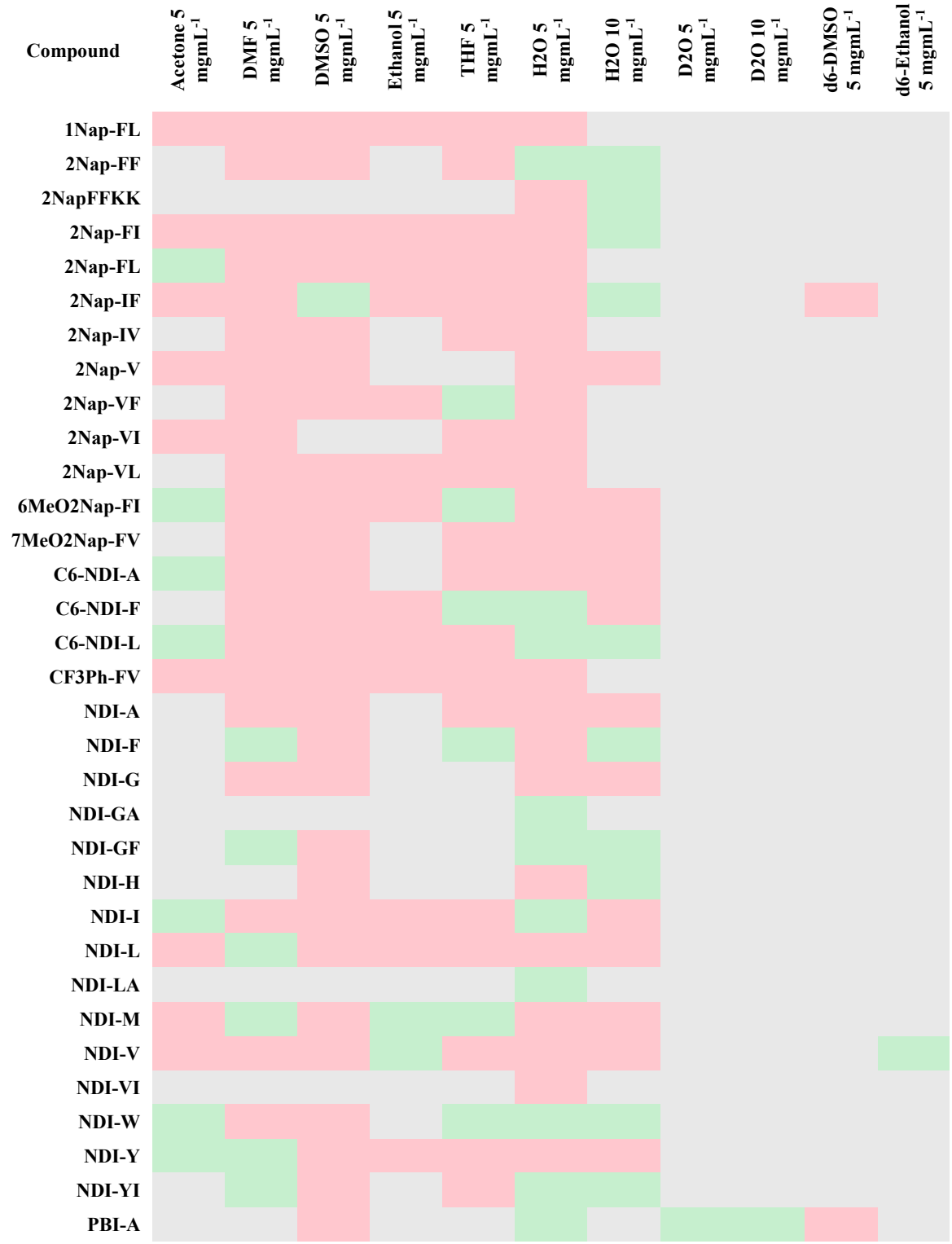


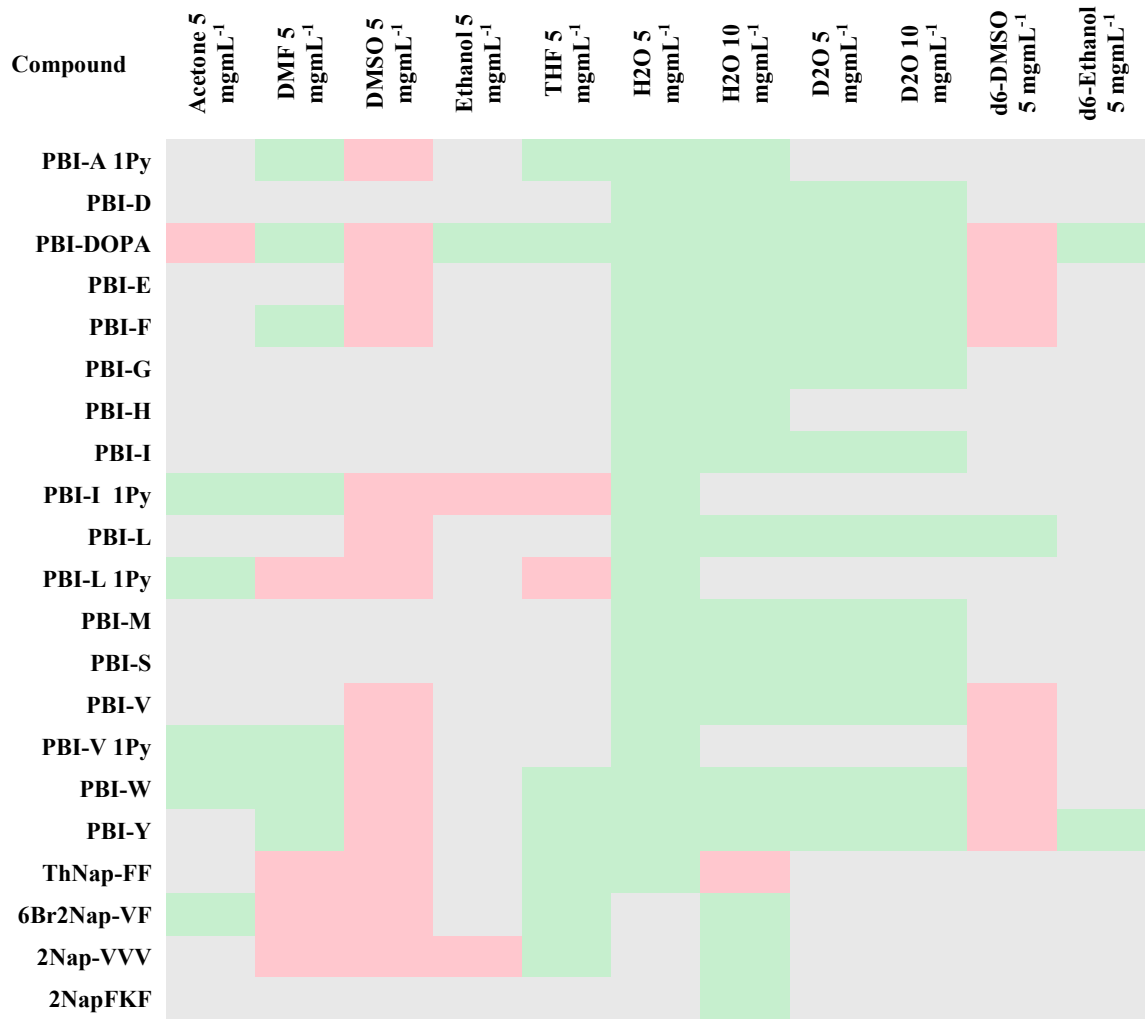

ThNap-FF (34) + PBI-E (23)
Classification: Good Scattering
ThNap-FF (34) + PBI-G (25)
Classification: Good Scattering
ThNap-FF (34) + PBI-H (26)
Classification: Good Scattering
ThNap-FF (34) + PBI-L (28)
Classification: Good Scattering
ThNap-FF (34) + PBI-M (29)
Classification: Good Scattering
ThNap-FF (34) + PBI-S (30)
Classification: Good Scattering
ThNap-FF (34) + PBI-V (31)
Classification: Good Scattering
ThNap-FF (34) + PBI-W (32)
Classification: Good Scattering
ThNap-FF (34) + PBI-Y (33)
Classification: Good Scattering
I(q) (cm-1)
I(q) - Error of I(q) (cm-1)
q (Å-1)

A 4.3 Classification Summaries

A 4.3.1 Single Component Summary

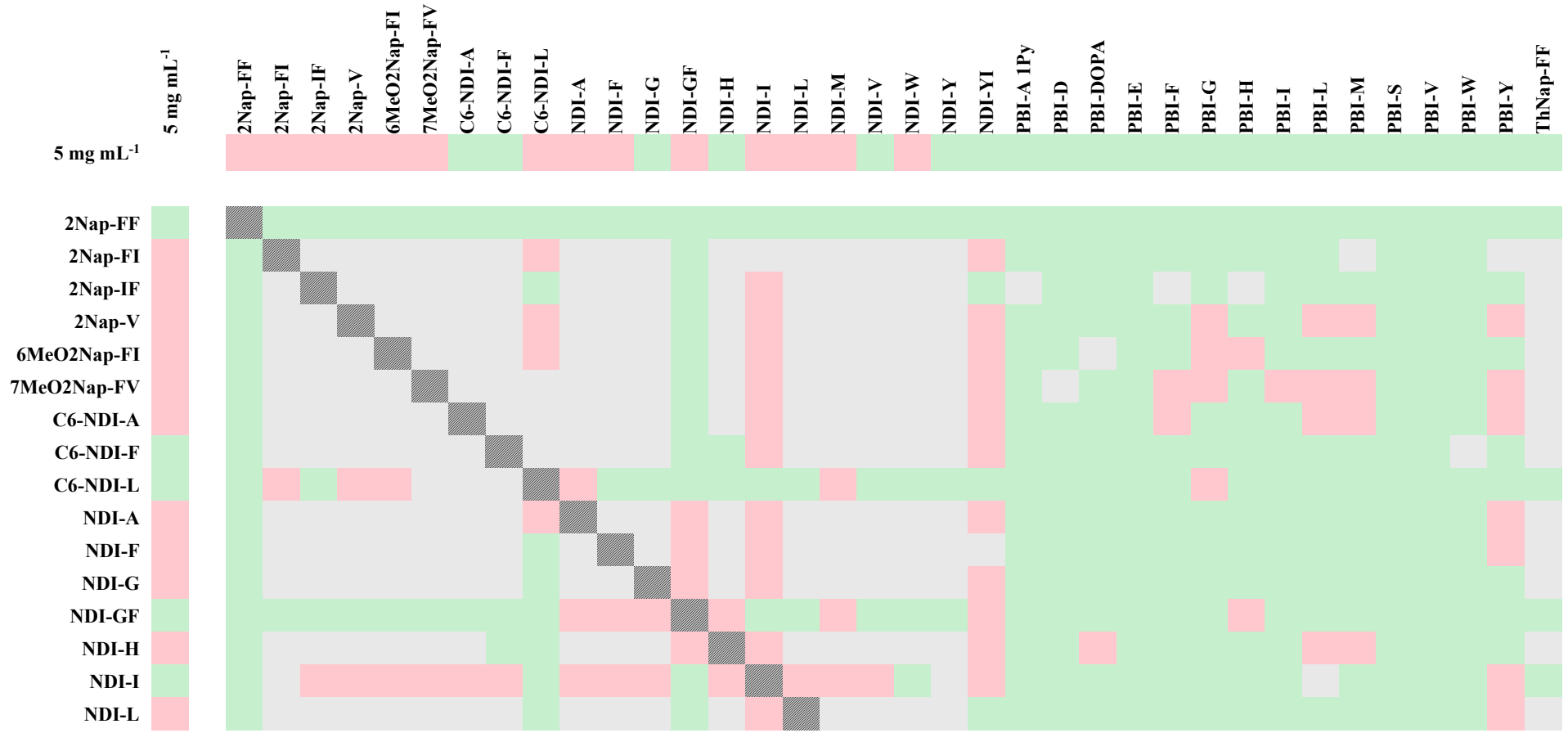
The classification of compounds across the varying solvents and concentrations. Green indicates good scattering and red indicates bad scattering as described in Section [#]. Grey indicates samples that were not characterised.

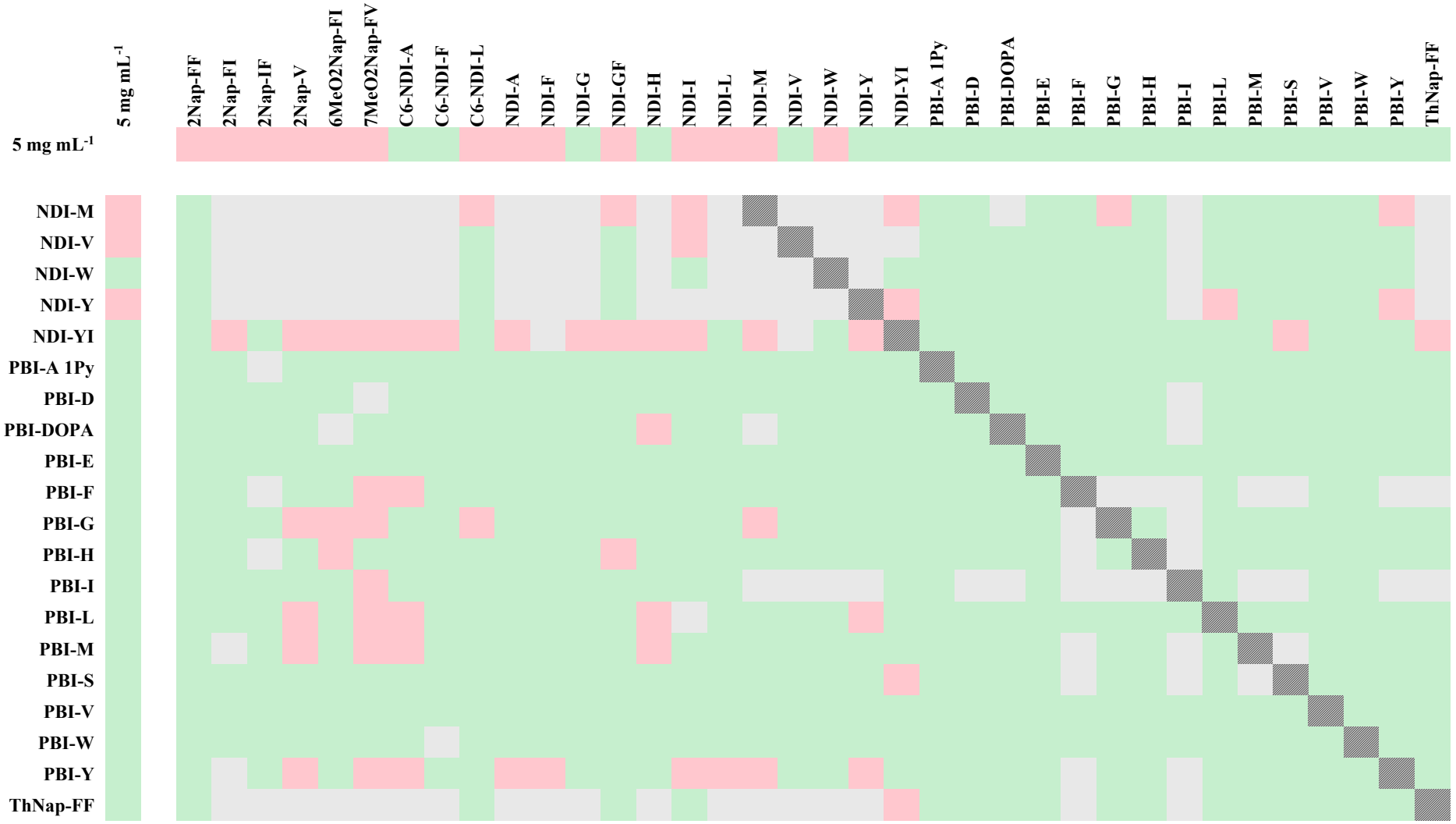



Compound
Acetone 5 mgmL-1
DMF 5 mgmL-1
DMSO 5 mgmL-1
Ethanol 5 mgmL-1
THF 5 mgmL-1
H2O 5 mgmL-1
H2O 10 mgmL-1
D2O 5 mgmL-1
D2O 10 mgmL-1
d6-DMSO 5 mgmL-1
d6-Ethanol 5 mgmL-1
PBI-A 1Py
PBI-D
PBI-DOPA
PBI-E
PBI-F
PBI-G
PBI-H
PBI-I
PBI-I 1Py
PBI-L
PBI-L 1Py
PBI-M
PBI-S
PBI-V
PBI-V 1Py
PBI-W
PBI-Y
ThNap-FF
6Br2Nap-VF
2Nap-VVV
2NapFKF

A 4.3.2 Multicomponent Scattering Classification

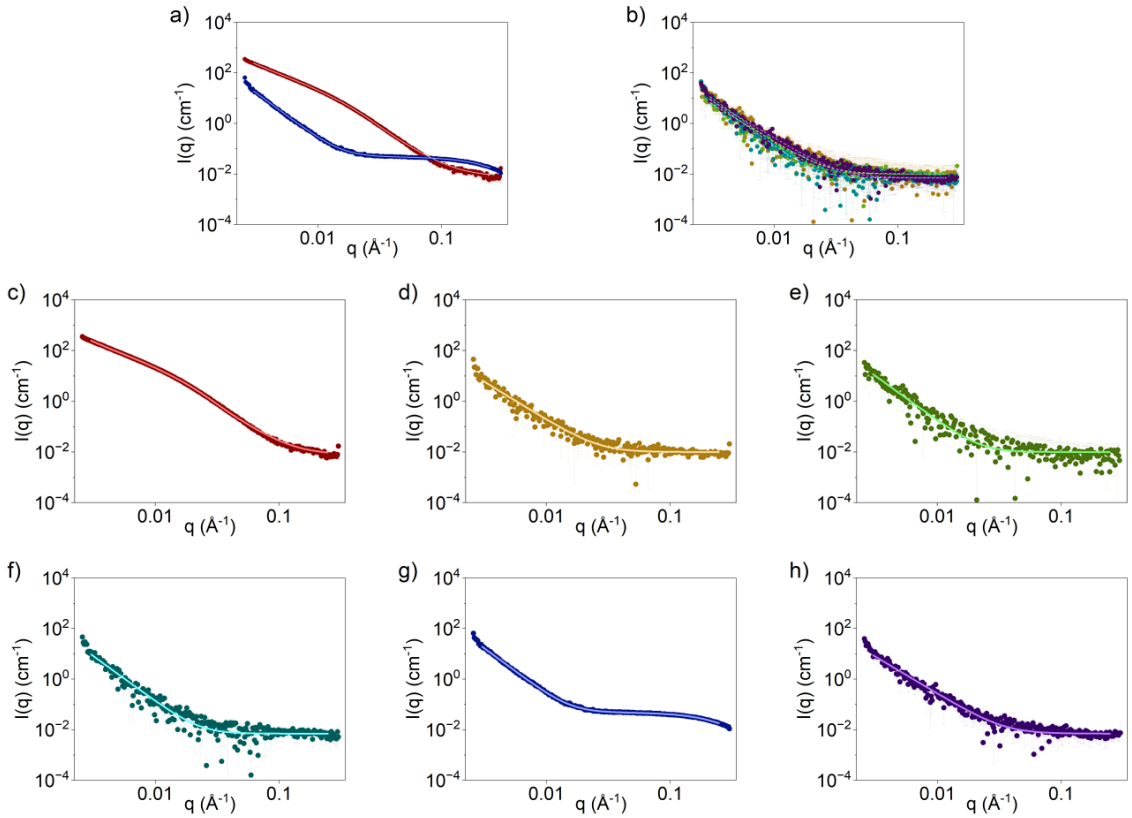
The classification of multicomponent combinations aqueous solutions. Green indicates good scattering and red indicates bad scattering as described in Section [#]. Grey indicates samples that were not characterised. The single component scattering of aqueous 5 mgmL^{-1} solutions is also provided.




5 mg mL-1
2Nap-FF
2Nap-FI
2Nap-IF
2Nap-V
6MeO2Nap-FI
7MeO2Nap-FV
C6-NDI-A
C6-NDI-F
C6-NDI-L
NDI-A
NDI-F
NDI-G
NDI-GF
NDI-H
NDI-I
NDI-L
NDI-M
NDI-V
NDI-W
NDI-Y
NDI-YI
PBI-A 1Py
PBI-D
PBI-DOPA
PBI-E
PBI-F
PBI-G
PBI-H
PBI-I
PBI-L
PBI-M
PBI-S
PBI-V
PBI-W
PBI-Y
ThNap-FF

A 4.4 Solvent Isotope Scattering Comparisons

A 4.4.1 Example across Solvents



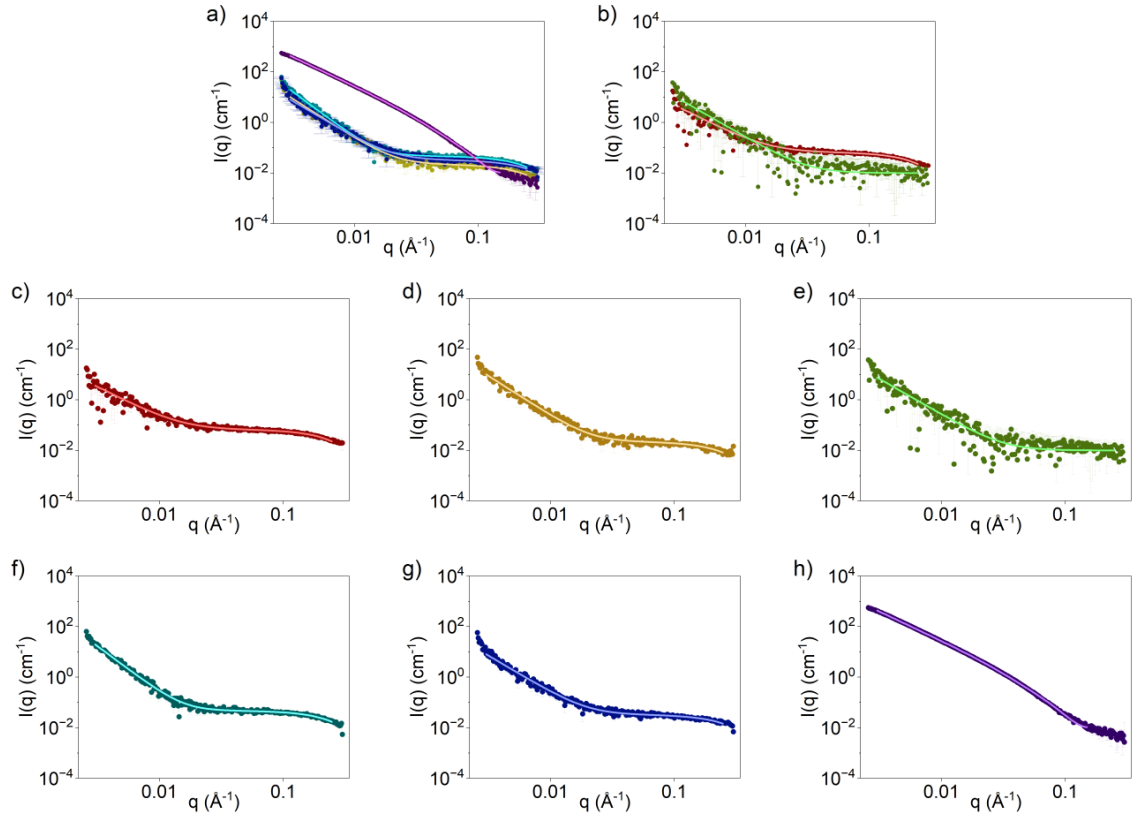
SAXS data (filled circles) and fits (line) for a solution of 6MeO2Nap-FI (**6**) at a concentration of 5 mgmL^{-1}. a) represents the good scatterers. b) represents the bad scatterers. c – h) are the individual scattering profiles for acetone, DMF, DMSO, ethanol, THF, and water, respectively.

The model fit parameters generated for 6MeO2Nap-FI (**6**) by fitting models to the SAXS data in SasView for acetone, DMF, and DMSO.

	Acetone	DMF	DMSO
Scattering Quality Classification	Good	Bad	Bad
Model	Flexible Cylinder + Power Law	Power Law	Power Law
Background (cm^{-1})	0.008	0.01	0.01
Kuhn Length (Å)	167.5 ± 4.2	-	-
Length (Å)	> 2000	-	-
Radius (Å)	60.4 ± 0.3	-	-
Polydispersity	0.5		
Scale	$4.1\times10^{-4} \pm 4.4\times10^{-6}$	-	-
Power Law	2.7 ± 0.0	3.0 ± 0.1	3.5 ± 0.2
Scale	$2.4\times10^{-5} \pm 6.9\times10^{-7}$	$1.9\times10^{-7} \pm 5.2\times10^{-8}$	$1.4\times10^{-8} \pm 1.2\times10^{-8}$
χ^2	2.170	0.8436	0.4399

The model fit parameters generated for 6MeO2Nap-FI (**6**) by fitting models to the SAXS data in SasView for ethanol, THF, and water.

	Ethanol	THF	Water
Scattering Quality Classification	Bad	Good	Bad
Model	Power Law	Cylinder + Power Law	Power Law
Background (cm^{-1})	0.01	0.01	0.007
Kuhn Length (Å)	-	-	-
Length (Å)	-	> 2000	-
Radius (Å)	-	30.4 ± 0.1	-
Polydispersity	-	0.3	-
Scale	-	$5.7\times10^{-4} \pm 4.7\times10^{-6}$	-
Power Law	3.6 ± 0.1	2.4 ± 0.0	2.9 ± 0.0
Scale	$9.5\times10^{-9} \pm 3.7\times10^{-9}$	$4.0\times10^{-5} \pm 2.9\times10^{-6}$	$3.7\times10^{-7} \pm 8.3\times10^{-5}$
χ^2	0.9119	1.9210	0.8579



SAXS data (filled circles) and fits (line) for a solution of PBI-DOPA (**22**) at a concentration of 5 mgmL^{-1}. a) represents the good scatterers. b) represents the bad scatterers. c – h) are the individual scattering profiles for acetone, DMF, DMSO, ethanol, THF, and water, respectively.

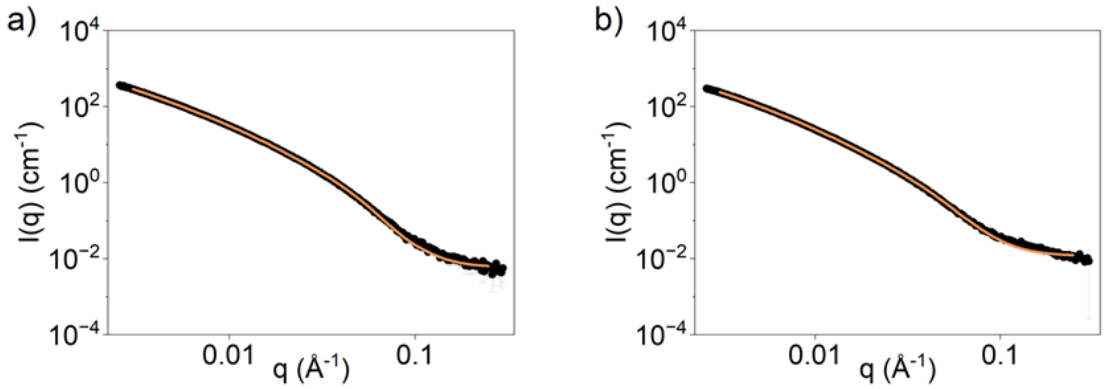
The model fit parameters generated for PBI-DOPA (**22**) by fitting models to the SAXS data in SasView for acetone, DMF, and DMSO.

	Acetone	DMF	DMSO
Scattering Quality Classification	Bad	Good	Bad
Model	Sphere + Power Law	Sphere + Power Law	Sphere + Power Law
Background (cm^{-1})	0.001	0.001	0.01
Radius (Å)	9.2 ± 0.1	9.0 ± 0.3	-
Scale	$4.7\times10^{-3} \pm 1.3\times10^{-4}$	$1.6\times10^{-3} \pm 1.4\times10^{-4}$	-
Power Law	2.5 ± 0.1	3.1 ± 0.0	2.8 ± 0.1
Scale	$2.2\times10^{-6} \pm 6.6\times10^{-7}$	$1.3\times10^{-7} \pm 3.0\times10^{-8}$	$6.4\times10^{-7} \pm 3.7\times10^{-7}$
χ^2	0.8456	0.7618	0.5220

The model fit parameters generated for PBI-DOPA (**22**) by fitting models to the SAXS data in SasView for ethanol, THF, and water.

	Ethanol	THF	Water
Scattering Quality Classification	Good	Good	Good
Model	Sphere + Power Law	Sphere + Power Law	Flexible Cylinder + Power Law
Background (cm^{-1})	0.001	0.001	0.001
Kuhn Length (Å)	-	-	54.7 ± 0.0
Length (Å)	-	-	218.64 ± 5.0
Radius (Å)	8.9 ± 0.1	8.0 ± 0.1	11.1 ± 0.2
Scale	$3.6\times10^{-3} \pm 1.5\times10^{-4}$	$3.3\times10^{-3} \pm 2.2\times10^{-4}$	$5.4\times10^{-3} \pm 1.5\times10^{-4}$
Power Law	3.6 ± 0.0	2.9 ± 0.0	2.3 ± 0.0
Scale	$1.3\times10^{-8} \pm 2.7\times10^{-9}$	$3.9\times10^{-7} \pm 8.6\times10^{-8}$	$6.8\times10^{-4} \pm 6.8\times10^{-6}$
χ^2	0.8569	0.8224	2.9245

A 4.4.2 Examples in H_2O and D_2O

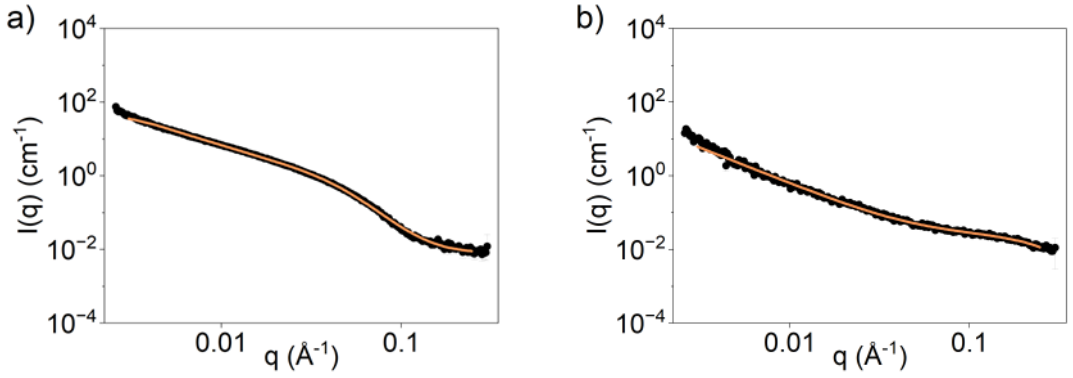


SAXS data (filled circles) and fits (line) for a solution of PBI-E (**23**) at a concentration of 5 mgmL^{-1}. a) represents the sample prepared in H_2O. b) represents the sample prepared in D_2O.

The model fit parameters generated for PBI-E (**23**) by fitting models to the SAXS data in SasView.

	H_2O	D_2O
Scattering Quality Classification	Good	Good
Model	Flexible Elliptical Cylinder	Flexible Elliptical Cylinder
Background (cm^{-1})	0.006	0.012
Kuhn Length (Å)	264.6 ± 10.5	311.6 ± 26.8
Length (Å)	> 2000 [a]	> 2000 [a]
Radius (Å)	36.4 ± 0.1	34.2 ± 0.2
Polydispersity	0.3	0.5
Axis Ratio	3.7 ± 0.1	4.0 ± 0.2
Scale	0.00958 ± 1.4×10^{-5}	0.00788 ± 3.2×10^{-5}
χ^2	1.9621	2.4548

a Beyond the limit of quantification determined by minimum q = 0.003 Å^{-1} = 2000 Å



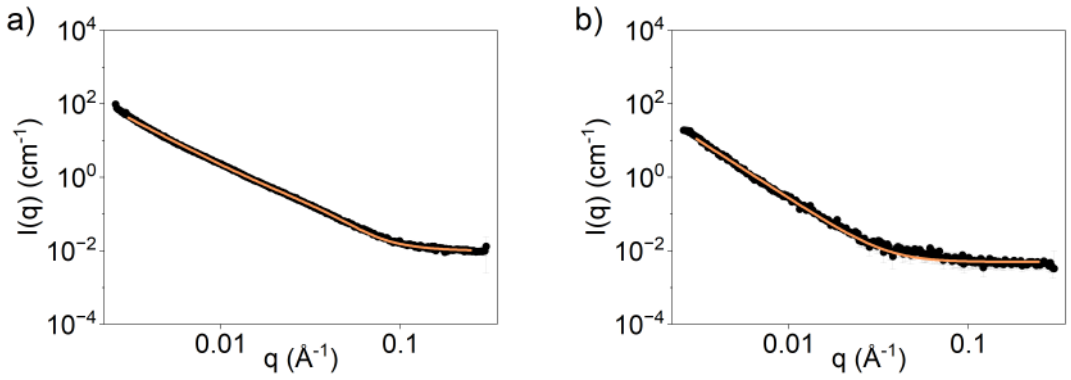
SAXS data (filled circles) and fits (line) for a solution of PBI-M (**29**) at a concentration of 5 mgmL^{-1}. a) represents the sample prepared in H_2O. b) represents the sample prepared in D_2O.

The model fit parameters generated for PBI-M (**23**) by fitting models to the SAXS data in SasView.

	H_2O	D_2O
Scattering Quality Classification	Good	Good
Model	Flexible Elliptical Cylinder	Sphere + Power Law
Background (cm-1)	0.009	0.012
Kuhn Length (Å)	373.0 ± 6.5	-
Length (Å)	> 2000 [a]	-
Radius (Å)	20.9 ± 0.2	8.5 ± 0.3
Polydispersity	0.3	-
Axis Ratio	2.5 ± 0.4	-
Scale	0.00109 ± 1.4×10^{-5}	0.005 ± 3.5×10^{-4}
Power Law	-	1.95 ± 0.02
Scale	-	7.6×10^{-5} ± 6.0×10^{-6}
χ^2	1.6030	1.1053

a Beyond the limit of quantification determined by minimum q = 0.003 Å^{-1} = 2000 Å

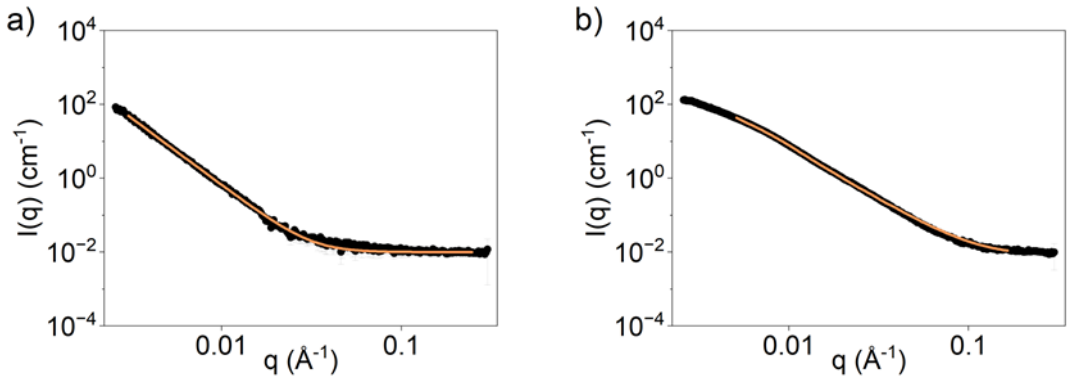
A 4.4.3 Examples in DMSO and d_6-DMSO



SAXS data (filled circles) and fits (line) for a solution of 2Nap-IF (**4**) at a concentration of 5 mgmL^{-1}. a) represents the sample prepared in DMSO. b) represents the sample prepared in d_6-DMSO.

The model fit parameters generated for 2Nap-IF (**4**) by fitting models to the SAXS data in SasView.

	DMSO	d_6-DMSO
Scattering Quality Classification	Good	Bad
Model	Cylinder + Power Law	Power Law
Background (cm^{-1})	0.01	0.005
Length (Å)	381.2 ± 30.5	
Radius (Å)	37.2 ± 0.6	-
Scale	$2.0\times10^{-4} \pm 7.1\times10^{-6}$	-
Power Law	2.6 ± 0.0	3.1 ± 0.0
Scale	$1.1\times10^{-5} \pm 6.0\times10^{-7}$	$1.4\times10^{-7} \pm 1.1\times10^{-8}$
χ^2	1.9205	0.9450

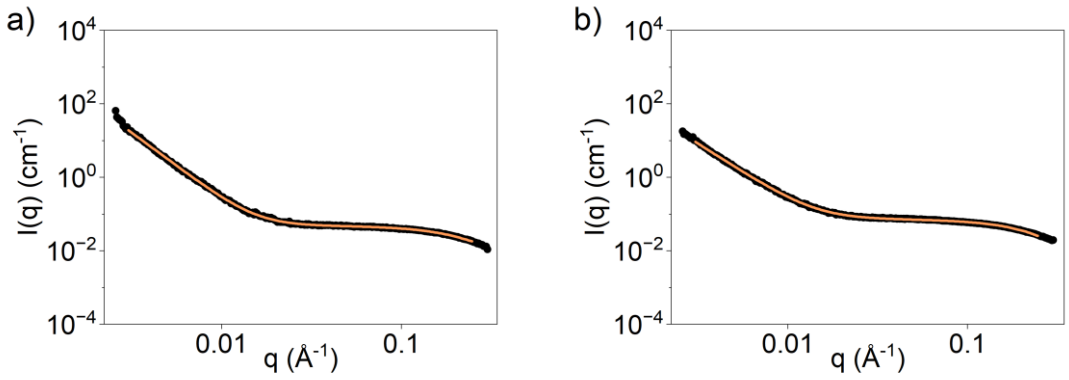


SAXS data (filled circles) and fits (line) for a solution of PBI-L (**28**) at a concentration of 5 mgmL^{-1}. a) represents the sample prepared in DMSO. b) represents the sample prepared in d_6-DMSO.

The model fit parameters generated for PBI-L (**28**) by fitting models to the SAXS data in SasView.

	DMSO	d_6-DMSO
Scattering Quality Classification	Bad	Good
Model	Power Law	Cylinder + Power Law
Background (cm^{-1})	0.01	0.008
Length (Å)	-	574.5 ± 5.2
Radius (Å)	-	144.6 ± 0.6
Polydispersity	-	0.5
Scale	-	$1.4\times10^{-4} \pm 2.1\times10^{-6}$
Power Law	3.6 ± 0.0	2.7 ± 0.0
Scale	$3.9\times10^{-8} \pm 1.5\times10^{-9}$	$2.3\times10^{-5} \pm 3.0\times10^{-7}$
χ^2	1.7185	4.9039

A 4.4.4 Examples in Ethanol and d_6-Ethanol



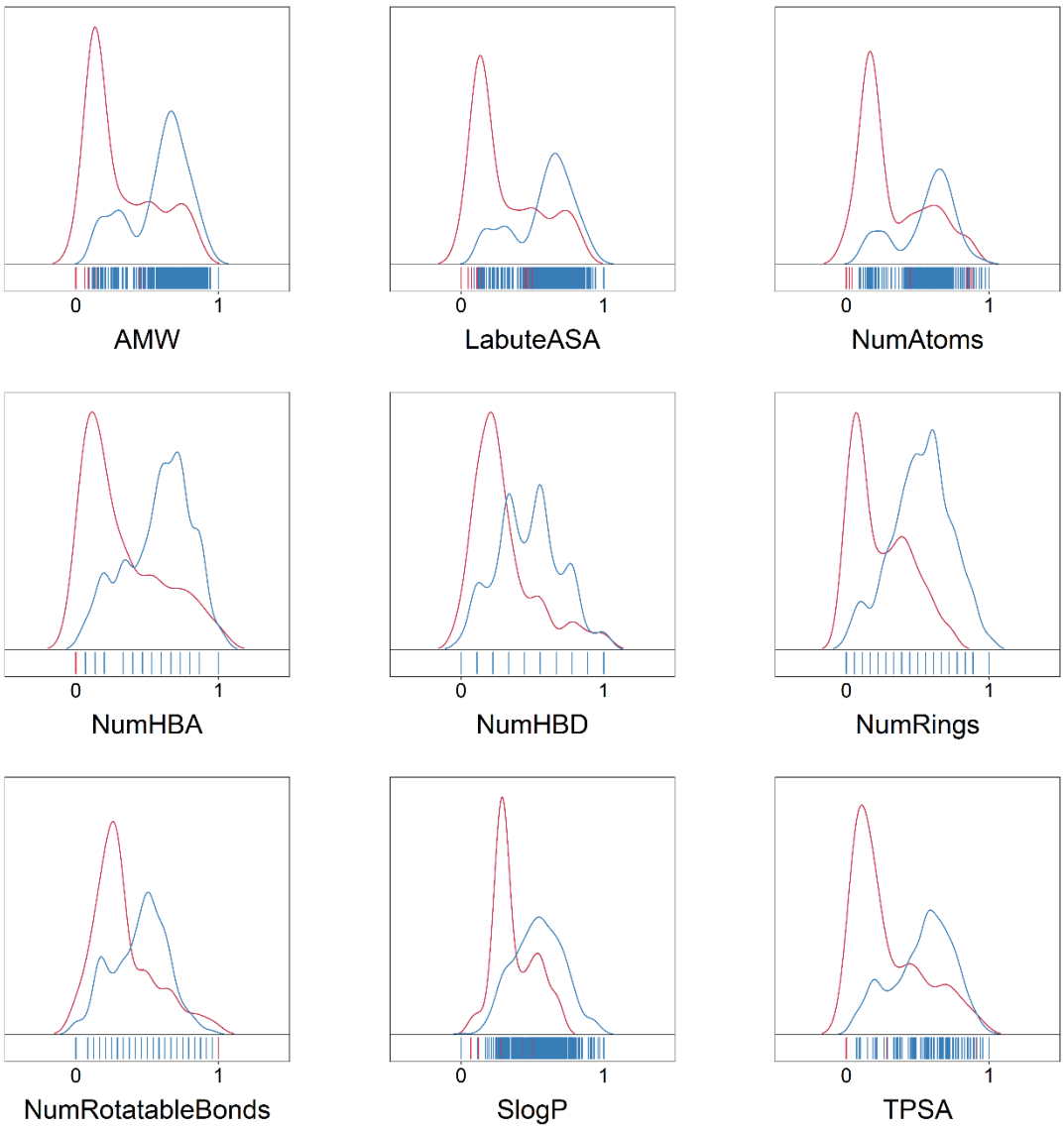
SAXS data (filled circles) and fits (line) for a solution of PBI-DOPA (**22**) at a concentration of 5 mgmL^{-1}. a) represents the sample prepared in ethanol. b) represents the sample prepared in d_6-ethanol.

The model fit parameters generated for PBI-DOPA (**22**) by fitting models to the SAXS data in SasView.

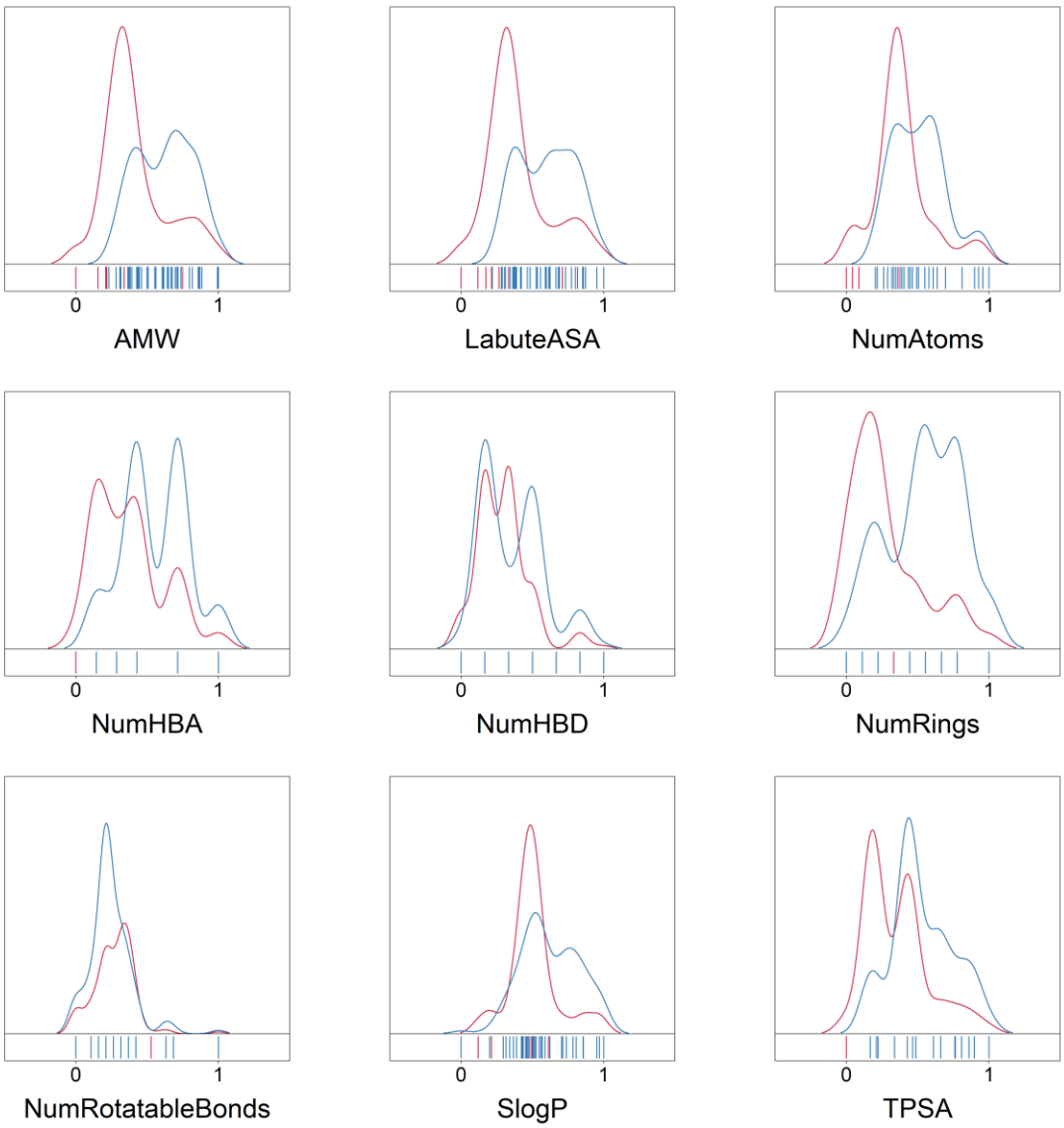
	Ethanol	d_6-ethanol
Scattering Quality Classification	Good	Good
Model	Sphere + Power Law	Sphere + Power Law
Background (cm^{-1})	0.01	0.02
Radius (Å)	10.6 ± 0.0	11.83 ± 0.0
Scale	$2.0\times10^{-3} \pm 1.5\times10^{-5}$	$2.1\times10^{-3} \pm 6.7\times10^{-6}$
Power Law	3.6 ± 0.0	3.2 ± 0.0
Scale	$1.2\times10^{-8} \pm 4.9\times10^{-10}$	$9.7\times10^{-8} \pm 3.7\times10^{-9}$
χ^2	1.4803	2.4528

A 4.5 Physicochemical Descriptor Distribution

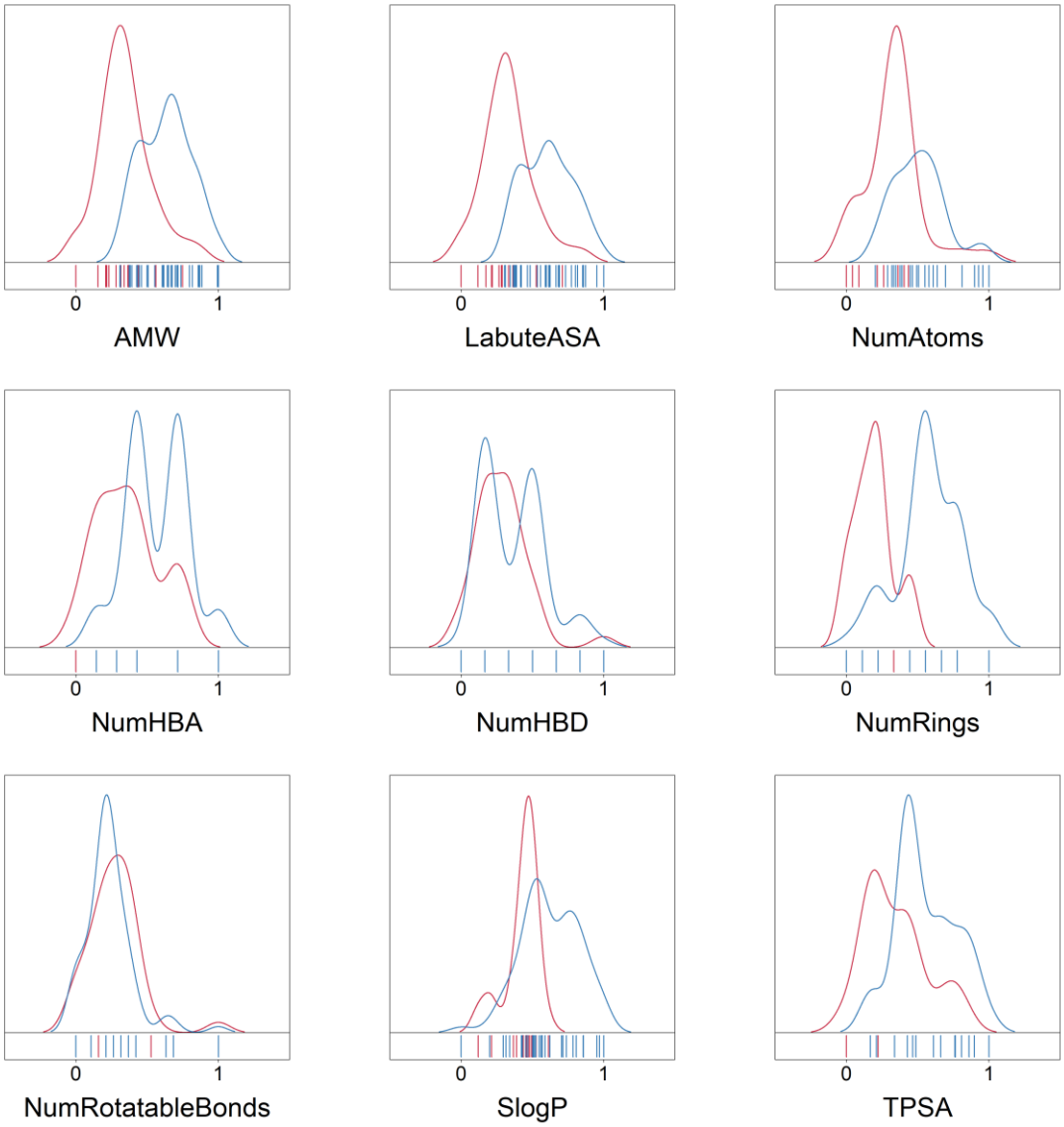
Plot of 9 physicochemical descriptors calculated and their distribution, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the "all solutions" QSPR model. The value of each molecule's descriptor is depicted by | marker. Values are normalised from 0 to 1.



Plot of 9 physicochemical descriptors calculated and their distribution, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the "single component" QSPR model. The value of each molecule's descriptor is depicted by | marker. Values are normalised from 0 to 1.

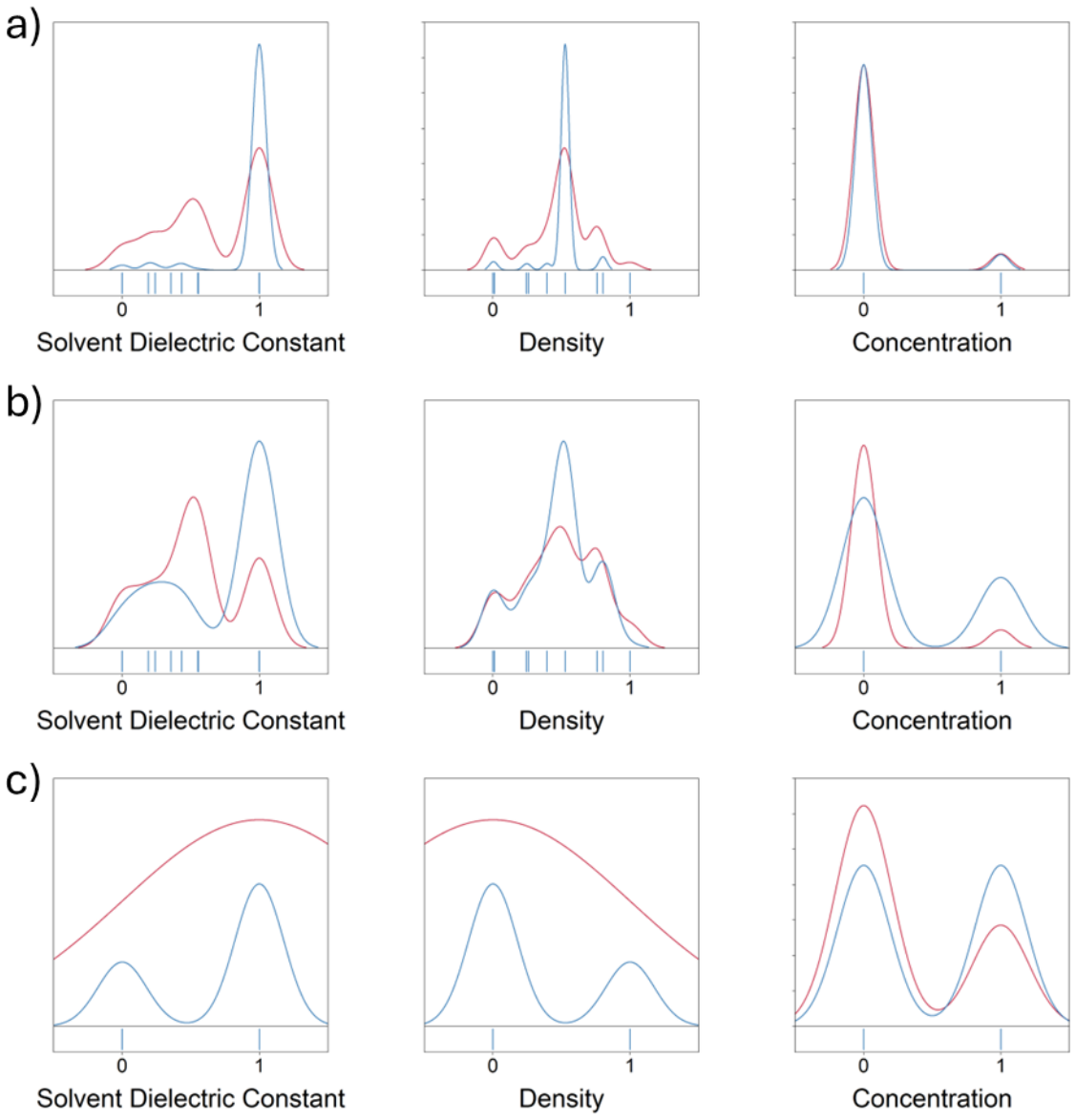


Plot of 9 physicochemical descriptors calculated and their distribution, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the "aqueous" QSPR model. The value of each molecule's descriptor is depicted by | marker. Values are normalised from 0 to 1.



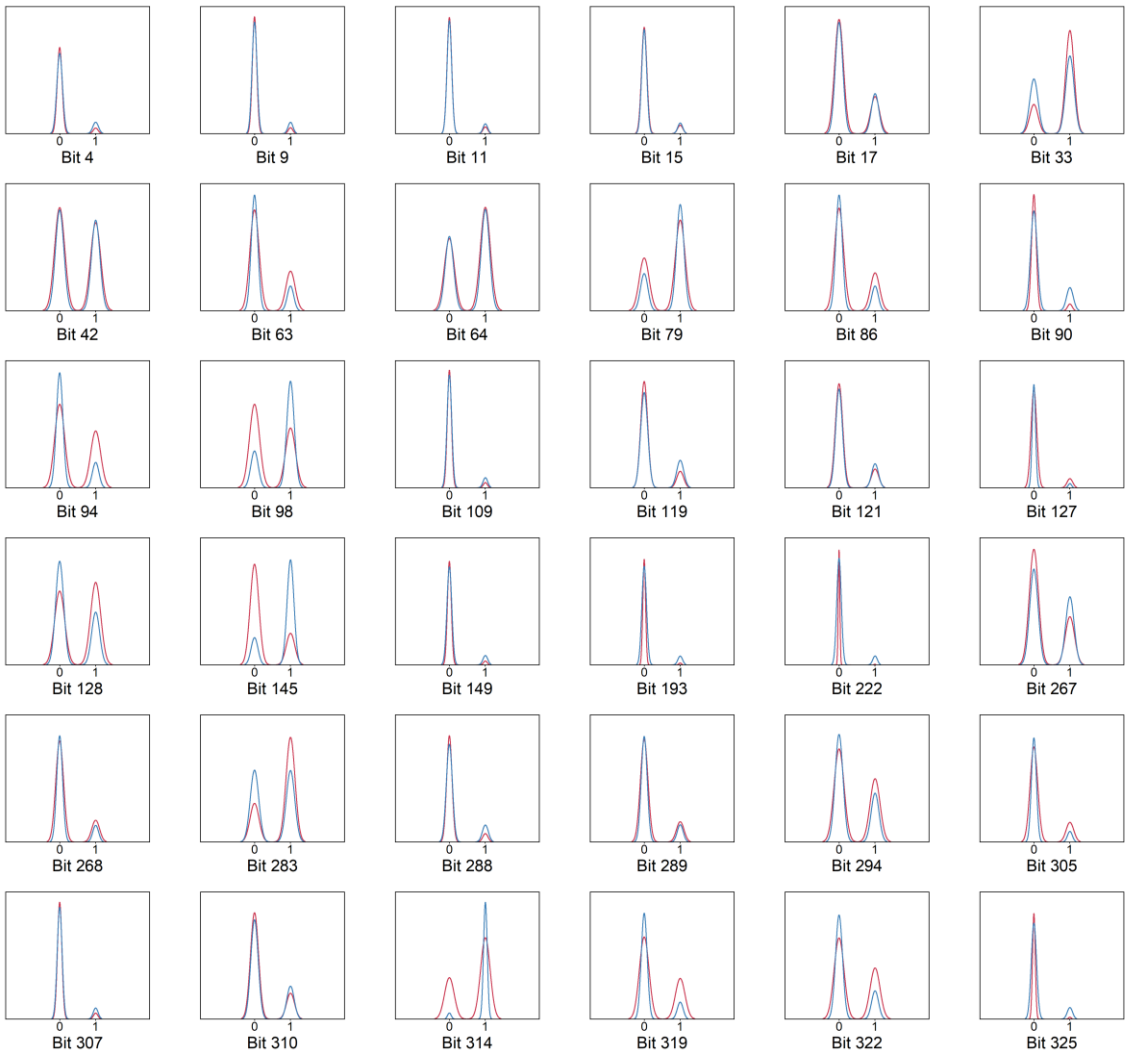
A 4.6 Solvent Descriptor Distribution

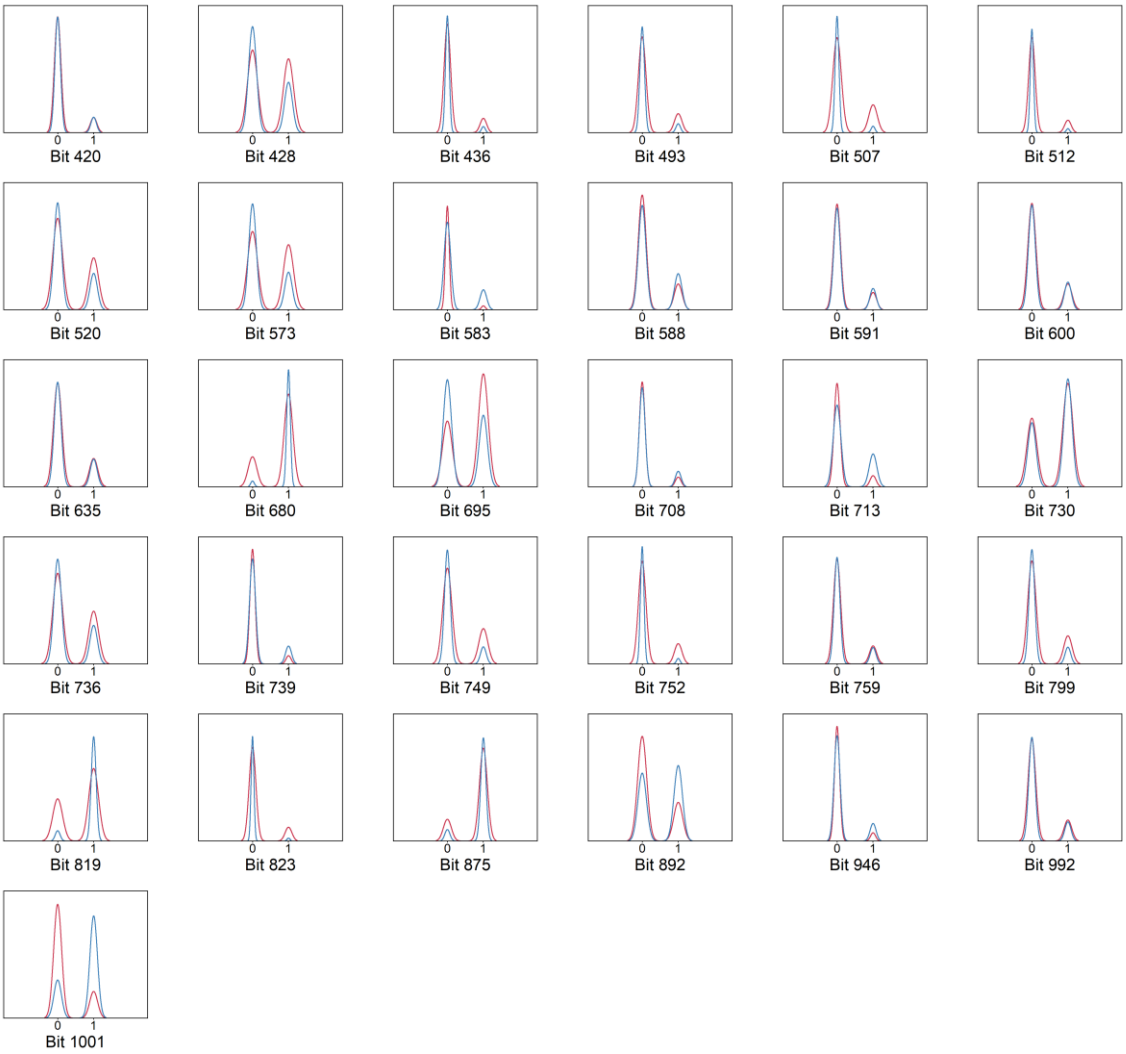
Plot of 3 solvent descriptors calculated and their distribution, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the a) “all solutions”, b) “single component”, and c) “aqueous” QSPR models. The value of each molecule’s descriptor is depicted by | marker. Values are normalised from 0 to 1.



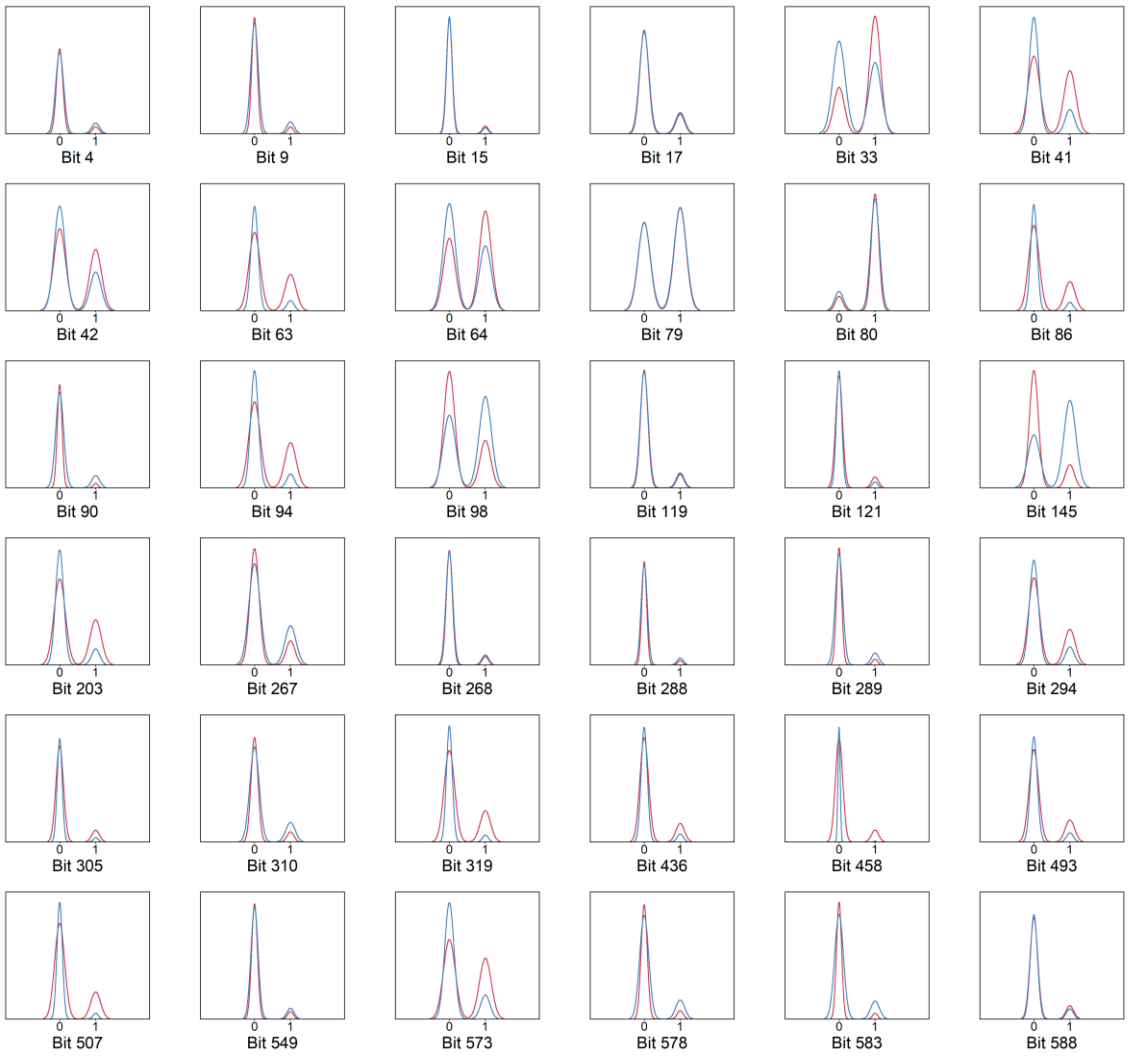
A 4.7 Fingerprint Descriptor Distribution

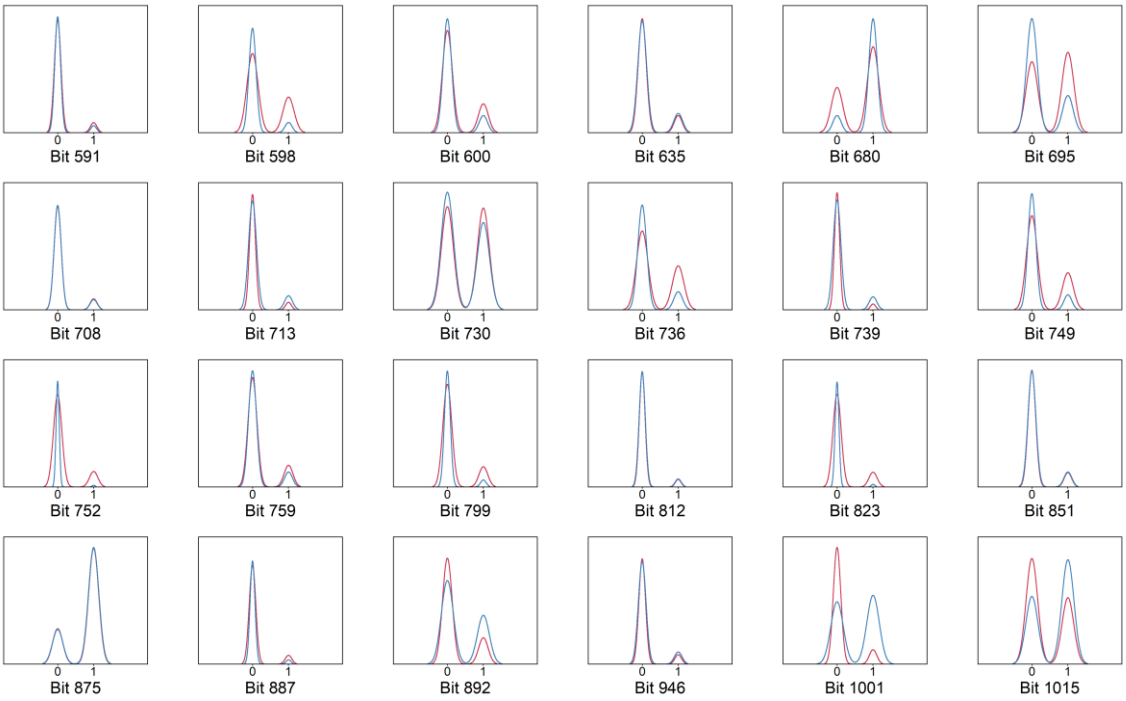
Plot of the 67 Morgan fingerprint descriptors and their distributions, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the “all solutions” QSPR model.



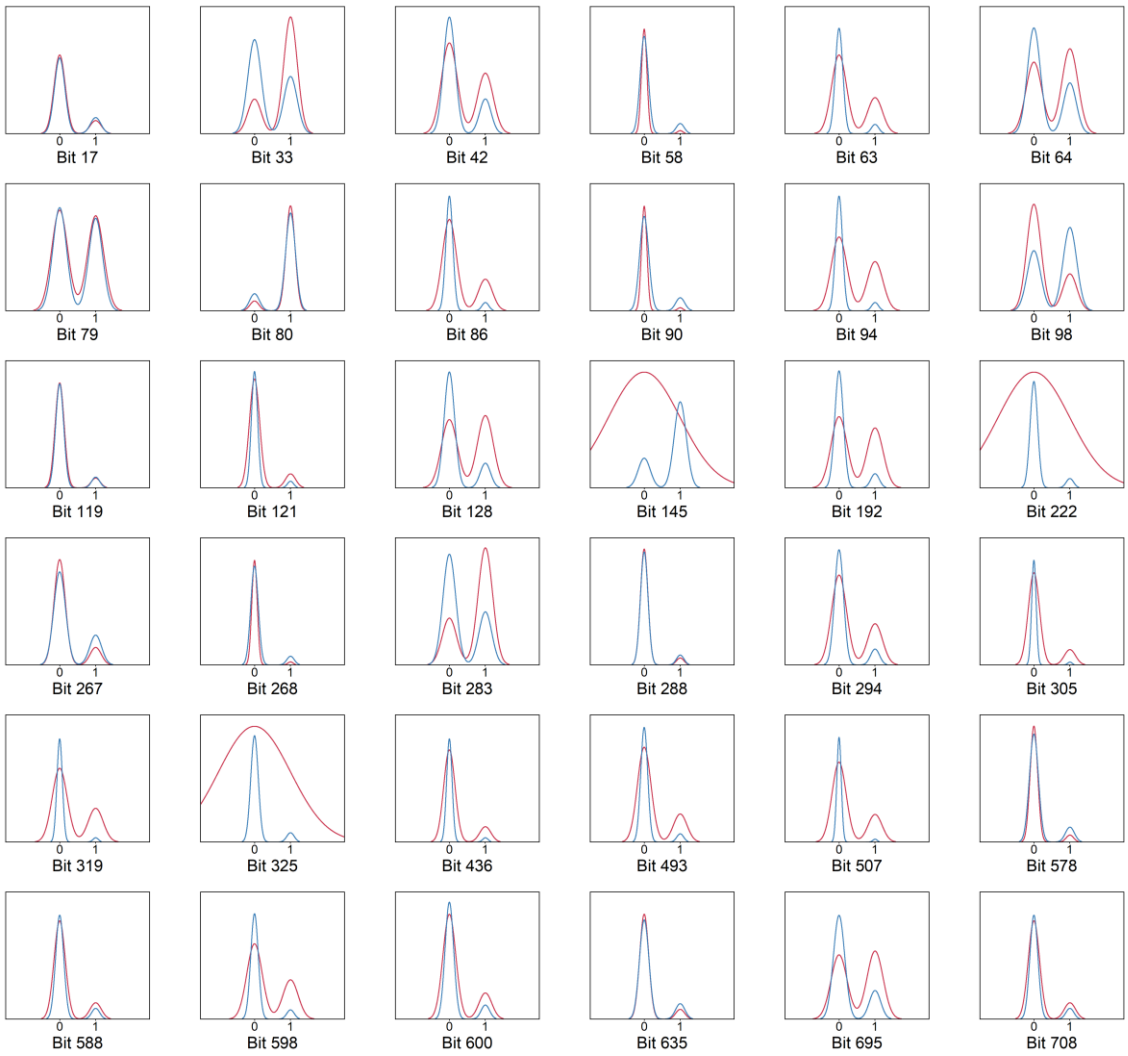

0 1
Bit 420
0 1
Bit 428
0 1
Bit 436
0 1
Bit 493
0 1
Bit 507
0 1
Bit 512
0 1
Bit 520
0 1
Bit 573
0 1
Bit 583
0 1
Bit 588
0 1
Bit 591
0 1
Bit 600
0 1
Bit 635
0 1
Bit 680
0 1
Bit 695
0 1
Bit 708
0 1
Bit 713
0 1
Bit 730
0 1
Bit 736
0 1
Bit 739
0 1
Bit 749
0 1
Bit 752
0 1
Bit 759
0 1
Bit 799
0 1
Bit 819
0 1
Bit 823
0 1
Bit 875
0 1
Bit 892
0 1
Bit 946
0 1
Bit 992
0 1
Bit 1001

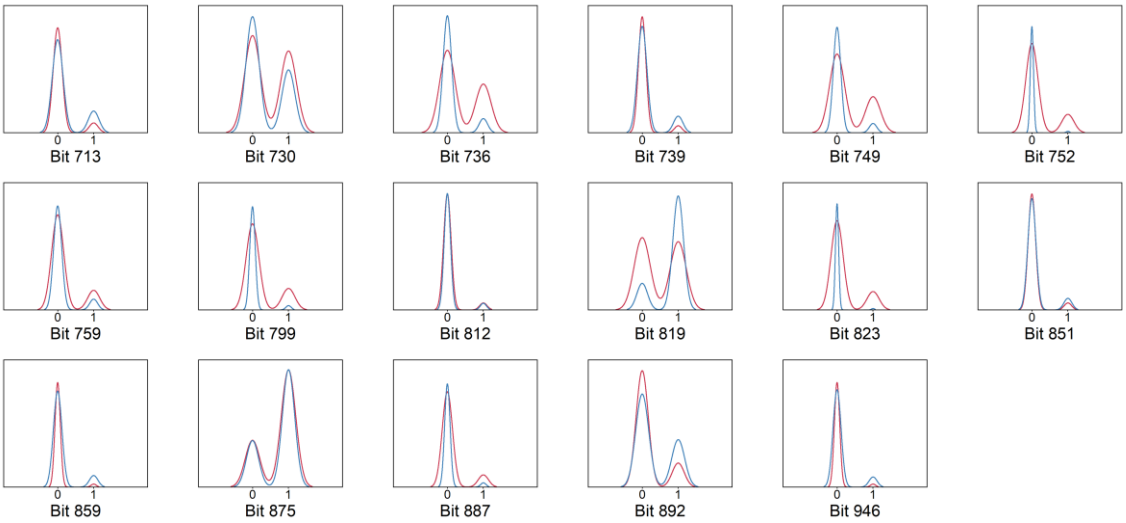
Plot of the 60 Morgan fingerprint descriptors and their distributions, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the "single component" QSPR model.




0 1
Bit 591
0 1
Bit 598
0 1
Bit 600
0 1
Bit 635
0 1
Bit 680
0 1
Bit 695
0 1
Bit 708
0 1
Bit 713
0 1
Bit 730
0 1
Bit 736
0 1
Bit 739
0 1
Bit 749
0 1
Bit 752
0 1
Bit 759
0 1
Bit 799
0 1
Bit 812
0 1
Bit 823
0 1
Bit 851
0 1
Bit 875
0 1
Bit 887
0 1
Bit 892
0 1
Bit 946
0 1
Bit 1001
0 1
Bit 1015

Plot of the 53 Morgan fingerprint descriptors and their distributions, separated by scattering quality classification (bad, red; good, blue) for the combined training and test set of the "aqueous" QSPR model.




0 1
Bit 713
0 1
Bit 730
0 1
Bit 736
0 1
Bit 739
0 1
Bit 749
0 1
Bit 752
0 1
Bit 759
0 1
Bit 799
0 1
Bit 812
0 1
Bit 819
0 1
Bit 823
0 1
Bit 851
0 1
Bit 859
0 1
Bit 875
0 1
Bit 887
0 1
Bit 892
0 1
Bit 946

A 4.8 Model Performance Metrics

Optimisation and performance statistics of the QSPR models developed for the training set.

		Resampling results of optimal model	Performance of optimal model on training set				
Model	Method	MCC	ROC AUC	Balanced Accuracy	MCC	K	Overall quality of model
All Solutions – Original Hyperparameters	RF	0.807	0.994	0.954	0.884	0.882	Good
	C4.5	0.807	0.998	0.978	0.961	0.961	Good
	kNN	0.529	0.907	0.855	0.663	0.649	Good
	SVM	0.450	0.800	0.808	0.676	0.664	Good
	XG	0.808	0.998	0.974	0.936	0.936	Good
	NB	-	0.820	0.736	0.447	0.441	Bad
	LR	0.715	0.923	0.871	0.712	0.707	Good
Single Component – Original Hyperparameters	RF	0.630	0.990	0.931	0.859	0.859	Good
	C4.5	0.681	0.974	0.887	0.774	0.774	Good
	kNN	0.421	0.766	0.706	0.412	0.406	Bad
	SVM	0.461	0.598	0.500	N/A	0.000	Bad
	XG	0.647	0.991	0.951	0.906	0.906	Good
	NB	-	0.811	0.707	0.414	0.407	Bad
	LR	0.596	0.921	0.829	0.655	0.654	Good
H_2O and D_2O – Original Hyperparameters	RF	0.807	1.000	1.000	1.000	1.000	Good
	C4.5	0.920	1.000	1.000	1.000	1.000	Good
	kNN	0.536	0.888	0.792	0.569	0.567	Good
	SVM	0.706	0.500	0.500	N/A	0.000	Bad
	XG	0.874	1.000	1.000	1.000	1.000	Good
	NB	-	0.880	0.867	0.683	0.636	Good
	LR	0.778	1.000	1.000	1.000	1.000	Good
Single Component – Updated Hyperparameters	RF	0.543	0.965	0.883	0.763	0.762	Good
	C4.5	0.677	0.980	0.898	0.793	0.792	Good
	kNN	0.382	0.764	0.739	0.484	0.467	Bad
	SVM	0.316	0.993	0.948	0.902	0.900	Good
	XG	0.640	0.995	0.975	0.950	0.950	Good
	NB	-	0.759	0.706	0.411	0.404	Bad
	LR	0.478	0.886	0.793	0.585	0.584	Good
H_2O – Updated Hyperparameters	RF	0.792	0.987	0.885	0.755	0.740	Good
	C4.5	0.814	0.995	0.963	0.915	0.911	Good
	kNN	0.518	0.873	0.790	0.579	0.579	Good
	SVM	0.570	1.000	1.000	1.000	1.000	Good
	XG	0.831	0.997	0.958	0.911	0.910	Good
	NB	-	0.886	0.768	0.565	0.484	Good
	LR	0.737	0.984	0.916	0.823	0.822	Good

Optimisation and performance statistics of the QSPR models developed for the in-domain external test set compounds using the maxdist applicability domain.

		Performance of optimal model on training set				
Model	Method	ROC AUC	Balanced Accuracy	MCC	K	Overall quality of model
All Solutions – Original Hyperparameters	RF	0.924	0.836	0.699	0.696	Good
	C4.5	0.920	0.848	0.718	0.716	Good
	kNN	0.808	0.780	0.540	0.537	Good
	SVM	0.577	0.628	0.341	0.304	Bad
	XG	0.894	0.803	0.615	0.615	Good
	NB	0.817	0.751	0.446	0.446	Bad
	LR	0.915	0.833	0.670	0.670	Good
Single Component – Original Hyperparameters	RF	0.990	0.851	0.702	0.694	Good
	C4.5	0.973	0.905	0.808	0.807	Good
	kNN	0.705	0.631	0.283	0.253	Bad
	SVM	0.625	0.500	N/A	0.000	Bad
	XG	0.951	0.887	0.772	0.769	Good
	NB	0.784	0.613	0.249	0.218	Bad
	LR	0.932	0.833	0.669	0.657	Good
H_2O and D_2O – Original Hyperparameters	RF	1.000	1.000	1.000	1.000	Good
	C4.5	1.000	0.929	0.894	0.889	Good
	kNN	0.786	0.750	0.553	0.538	Good
	SVM	0.500	0.500	N/A	0.000	Bad
	XG	1.000	1.000	1.000	1.000	Good
	NB	0.969	0.964	0.901	0.897	Good
	LR	1.000	1.000	1.000	1.000	Good
Single Component – Updated Hyperparameters	RF	0.957	0.778	0.604	0.535	Good
	C4.5	0.982	0.852	0.723	0.686	Good
	kNN	0.871	0.738	0.517	0.457	Good
	SVM	0.736	0.631	0.270	0.266	Bad
	XG	0.976	0.926	0.852	0.841	Good
	NB	0.926	0.738	0.517	0.457	Good
	LR	0.886	0.768	0.541	0.526	Good
H_2O – Updated Hyperparameters	RF	0.979	0.950	0.887	0.881	Good
	C4.5	0.986	0.950	0.887	0.881	Good
	kNN	0.871	0.900	0.789	0.767	Good
	SVM	0.800	0.643	0.436	0.320	Bad
	XG	1.000	0.900	0.789	0.767	Good
	NB	0.971	0.850	0.700	0.658	Good
	LR	0.971	0.950	0.887	0.881	Good

Optimisation and performance statistics of the QSPR models developed for the in-domain external test set compounds using the dsz applicability domain.

		Performance of optimal model on training set				
Model	Method	ROC AUC	Balanced Accuracy	MCC	K	Overall quality of model
All Solutions – Original Hyperparameters	RF	0.936	0.851	0.730	0.728	Good
	C4.5	0.931	0.851	0.730	0.728	Good
	kNN	0.783	0.750	0.489	0.488	Bad
	SVM	0.549	0.643	0.400	0.353	Bad
	XG	0.911	0.815	0.639	0.639	Good
	NB	0.812	0.720	0.446	0.446	Bad
	LR	0.925	0.857	0.714	0.714	Good
Single Component – Original Hyperparameters	RF	0.949	0.866	0.736	0.733	Good
	C4.5	0.976	0.911	0.822	0.822	Good
	kNN	0.664	0.595	0.214	0.192	Bad
	SVM	0.686	0.500	N/A	0.000	Bad
	XG	0.945	0.888	0.778	0.777	Good
	NB	0.765	0.572	0.168	0.146	Bad
	LR	0.925	0.820	0.653	0.643	Good
H_2O and D_2O – Original Hyperparameters	RF	1.000	1.000	1.000	1.000	Good
	C4.5	1.000	1.000	1.000	1.000	Good
	kNN	0.500	0.464	-0.116	-0.097	Bad
	SVM	0.500	0.500	N/A	0.000	Bad
	XG	1.000	1.000	1.000	1.000	Good
	NB	0.929	0.964	0.835	0.821	Good
	LR	1.000	1.000	1.000	1.000	Good
Single Component – Updated Hyperparameters	RF	0.957	0.778	0.604	0.535	Good
	C4.5	0.982	0.825	0.694	0.650	Good
	kNN	0.886	0.700	0.500	0.400	Good
	SVM	0.667	0.575	0.155	0.150	Bad
	XG	0.978	0.950	0.905	0.900	Good
	NB	0.943	0.700	0.500	0.400	Good
	LR	0.875	0.725	0.465	0.450	Bad
H_2O – Updated Hyperparameters	RF	0.979	0.950	0.887	0.881	Good
	C4.5	0.975	0.950	0.849	0.837	Good
	kNN	0.850	0.900	0.730	0.696	Good
	SVM	0.650	0.500	N/A	0.000	Bad
	XG	1.000	0.900	0.730	0.696	Good
	NB	0.950	0.850	0.632	0.571	Good
	LR	0.950	0.950	0.849	0.837	Good

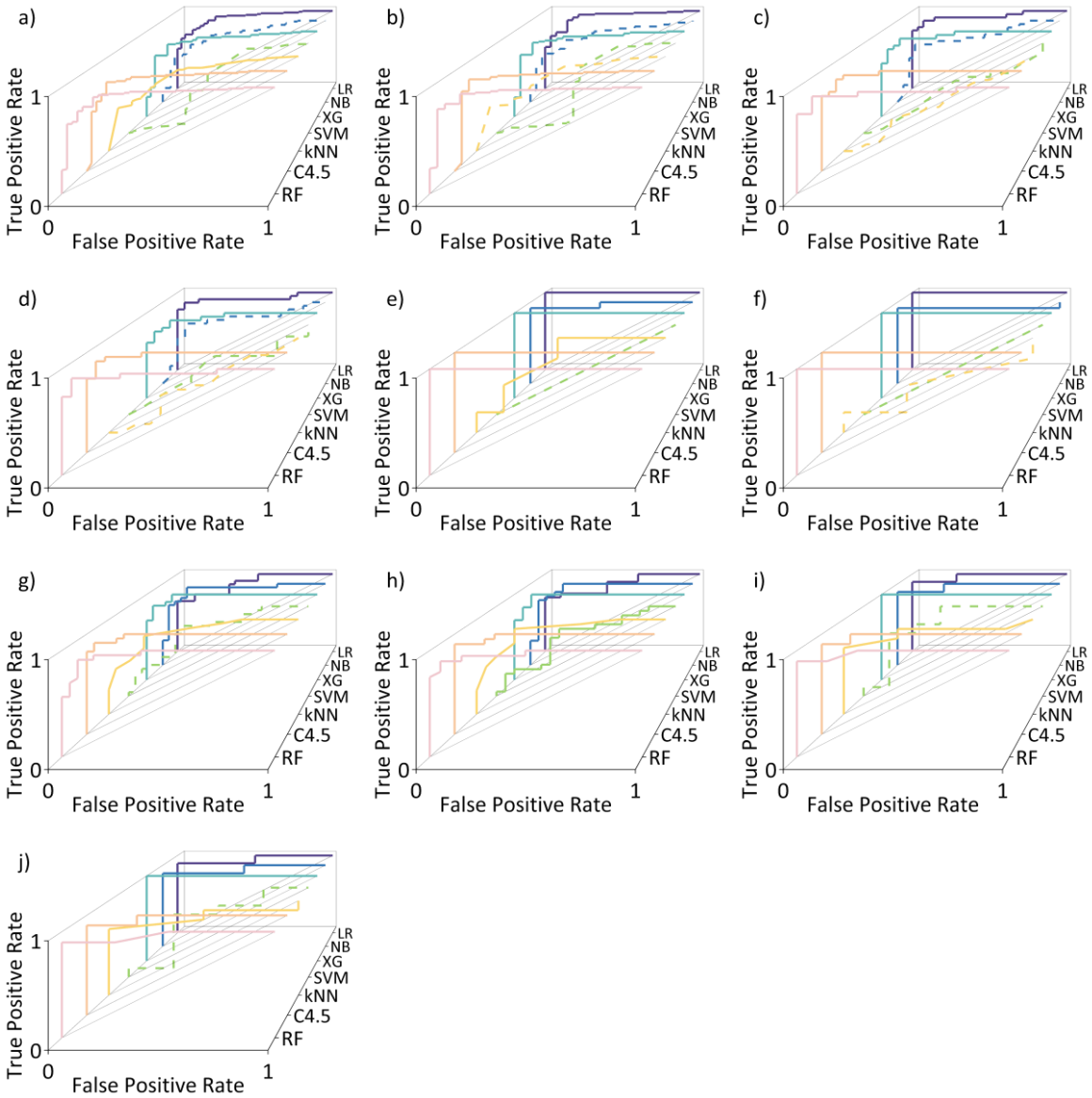
A 4.9 Partitioning Impact Assessment

Summary of the number of models which provided good predictive statistics for their respective training set and in-domain (maxdist and dsz) test set datapoints over 10 random stratified partitions. Models with poor training metrics but good test metrics are not counted.

Partition	Partition Number	Dataset				
		All Solutions	Single Component	H_2O and D_2O	Single Component	H_2O
Hyperparameters		Original	Original	Original	Updated	Updated
Training	1	6	5	6	6	7
	2	6	6	6	5	6
	3	5	4	6	4	7
	4	5	4	6	6	7
	5	6	4	6	6	7
	6	6	4	6	5	7
	7	6	5	6	5	7
	8	6	5	6	5	7
	9	6	4	6	5	7
	10	6	4	5	5	7
Maxdist	1	4	1	5	3	3
	2	4	2	5	3	6
	3	4	4	6	0	6
	4	4	4	1	3	4
	5	4	4	3	4	6
	6	4	3	5	4	5
	7	5	2	5	3	6
	8	5	4	6	4	0
	9	4	4	6	4	6
	10	4	3	5	4	4
Dsz	1	4	1	4	2	4
	2	3	0	5	3	6
	3	4	4	5	0	6
	4	3	4	1	1	4
	5	4	4	2	4	6
	6	5	3	4	5	5
	7	4	1	5	4	6
	8	4	4	5	4	0
	9	5	4	5	4	6
	10	3	2	5	3	3

A 4.10 Receiver Operator Characteristic Curves

Receiver operator characteristic (ROC) curves for all reported optimised models. Curves with a solid line represent models with good predictive statistics, while curves with a dashed line represents models with bad predictive statistics, as described in Appendix 4.7. Curves have been separated on the basis of model and applicability domain as follows: All solutions a) maxdist, b) dsz; Single component c) maxdist, d) dsz; H_2O and D_2O e) maxdist, f) dsz; Single component (updated hyperparameters) g) maxdist, h) dsz; H_2O (updated hyperparameters) i) maxdist, j) dsz.



A 4.11 Validation Predictions

Summary of the validation predictions deemed within the maxdist applicability domain of the respective model.

Validation predictions of the “All solutions Model”

Solution Constitution				Prediction	Combined Confidence
Solute	Solvent	Concentration	Scattering Quality		
PBI-L 1Py	Acetone	5	Good	Correct	0.24 ± 0.19
PBI-M	H2O	5	Good	Correct	0.69 ± 0.11
PBI-E	H2O	5	Good	Correct	0.64 ± 0.12
PBI-V	H2O	5	Good	Correct	0.62 ± 0.13
C6-NDI-F + 2Nap-FF	H2O	5	Good	Correct	0.74 ± 0.11
NDI-F + 2Nap-FF	H2O	5	Good	Correct	0.8 ± 0.14
PBI-A 1Py + 2Nap-FF	H2O	5	Good	Correct	0.98 ± 0.02
PBI-E + 2Nap-FF	H2O	5	Good	Correct	0.99 ± 0.01
PBI-G + 2Nap-FF	H2O	5	Good	Correct	0.91 ± 0.07
PBI-D	H2O	10	Good	Correct	0.8 ± 0.13
PBI-H	H2O	10	Good	Correct	0.66 ± 0.11
PBI-I	H2O	10	Good	Correct	0.65 ± 0.2
PBI-L + 2Nap-FI	H2O	5	Good	Correct	0.89 ± 0.12
NDI-GF + 6MeO2Nap-FI	H2O	5	Good	Correct	0.56 ± 0.31
NDI-GF + C6-NDI-F	H2O	5	Good	Correct	0.48 ± 0.34
PBI-V + PBI-H	H2O	5	Good	Correct	0.91 ± 0.06
PBI-V + PBI-E	H2O	5	Good	Correct	0.96 ± 0.05
PBI-E + 2Nap-IF	H2O	5	Good	Correct	0.98 ± 0.01
PBI-E + NDI-W	H2O	5	Good	Correct	0.98 ± 0.01
PBI-H + C6-NDI-L	H2O	5	Good	Correct	0.9 ± 0.07
ThNap-FF + PBI-W	H2O	5	Good	Correct	0.99 ± 0.01
PBI-W + 6MeO2Nap-FI	H2O	5	Good	Correct	0.95 ± 0.05
PBI-W + NDI-G	H2O	5	Good	Correct	0.99 ± 0.02
PBI-DOPA + C6-NDI-L	H2O	5	Good	Correct	0.89 ± 0.07
PBI-DOPA + NDI-YI	H2O	5	Good	Correct	0.79 ± 0.08
PBI-H + NDI-YI	H2O	5	Good	Correct	0.7 ± 0.29
PBI-D + NDI-V	H2O	5	Good	Correct	0.96 ± 0.05
PBI-D + NDI-A	H2O	5	Good	Correct	0.78 ± 0.05
PBI-I + NDI-I	H2O	5	Good	Correct	0.84 ± 0.17
PBI-L + NDI-H	H2O	5	Good	Correct	0.25 ± 0.26
PBI-H + NDI-V	H2O	5	Good	Correct	0.92 ± 0.07
PBI-G + NDI-A	H2O	5	Good	Correct	0.54 ± 0.19
PBI-M + NDI-A	H2O	5	Good	Correct	0.67 ± 0.09
PBI-M + 6MeO2Nap-FI	H2O	5	Good	Correct	0.77 ± 0.16
PBI-Y + NDI-G	H2O	5	Good	Correct	0.56 ± 0.09
PBI-W	D2O	10	Good	Correct	0.93 ± 0.06
PBI-M	D2O	10	Good	Correct	0.87 ± 0.09

PBI-V	D2O	10	Good	Correct	0.79 ± 0.06
PBI-Y + PBI-G	H2O	5	Good	Correct	0.6 ± 0.11
NDI-M	Acetone	5	Bad	Correct	0.45 ± 0.18
C6-NDI-A	H2O	5	Bad	Correct	0.4 ± 0.16
2Nap-FI	H2O	5	Bad	Correct	0.62 ± 0.15
NDI-L	H2O	5	Bad	Correct	0.89 ± 0.17
CF3Ph-FV	H2O	5	Bad	Correct	0.95 ± 0.05
2Nap-VI	H2O	5	Bad	Correct	0.83 ± 0.16
C6-NDI-F	DMSO	5	Bad	Correct	0.86 ± 0.07
2Nap-VL	DMSO	5	Bad	Correct	0.97 ± 0.01
NDI-Y	DMSO	5	Bad	Correct	0.95 ± 0.02
NDI-H	DMSO	5	Bad	Correct	0.94 ± 0.03
NDI-A	DMSO	5	Bad	Correct	0.98 ± 0.01
2Nap-IV	DMF	5	Bad	Correct	0.95 ± 0.05
6MeO2Nap-FI	DMF	5	Bad	Correct	0.92 ± 0.04
2Nap-IF	Ethanol	5	Bad	Correct	0.59 ± 0.12
NDI-Y	Ethanol	5	Bad	Correct	0.43 ± 0.33
NDI-L	THF	5	Bad	Correct	0.9 ± 0.07
2Nap-FI	THF	5	Bad	Correct	0.85 ± 0.05
NDI-YI + 2Nap-V	H2O	5	Bad	Correct	0.5 ± 0.25
NDI-YI + C6-NDI-A	H2O	5	Bad	Correct	0.23 ± 0.13
PBI-V	d6DMSO	5	Bad	Correct	0.44 ± 0.26
NDI-I + 2Nap-V	H2O	5	Bad	Correct	0.76 ± 0.03
NDI-I + NDI-F	H2O	5	Bad	Correct	0.55 ± 0.16
NDI-L	DMF	5	Good	Incorrect	0.93 ± 0.05
NDI-M	DMF	5	Good	Incorrect	0.61 ± 0.28
6Br2Nap-VF	THF	5	Good	Incorrect	0.42 ± 0.18
NDI-L + C6-NDI-L	H2O	5	Good	Incorrect	0.67 ± 0.2
NDI-GF + NDI-A	H2O	5	Bad	Incorrect	0.65 ± 0.14
ThNap-FF + NDI-YI	H2O	5	Bad	Incorrect	0.46 ± 0.25
PBI-H + 6MeO2Nap-FI	H2O	5	Bad	Incorrect	0.81 ± 0.11
PBI-M + C6-NDI-A	H2O	5	Bad	Incorrect	0.66 ± 0.13
PBI-M + NDI-H	H2O	5	Bad	Incorrect	0.85 ± 0.13

Validation prediction of the “Single component” model

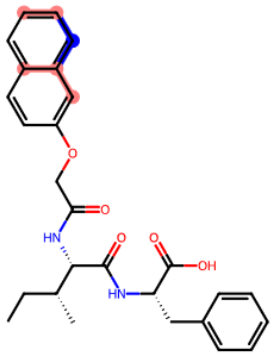
Solution Constitution				Prediction	Combined Confidence
Solute	Solvent	Concentration	Scattering Quality		
PBI-V	H2O	5	Good	Correct	0.74 ± 0.15
PBI-F	H2O	5	Good	Correct	0.89 ± 0.13
PBI-M	H2O	10	Good	Correct	0.89 ± 0.03
PBI-DOPA	H2O	10	Good	Correct	0.92 ± 0.06
PBI-DOPA	D2O	10	Good	Correct	0.88 ± 0.1
PBI-D	D2O	5	Good	Correct	0.74 ± 0.17
PBI-Y	D2O	10	Good	Correct	0.91 ± 0.05
PBI-F	D2O	10	Good	Correct	0.92 ± 0.08
2Nap-IF	Acetone	5	Bad	Correct	0.3 ± 0.25
2Nap-VI	Acetone	5	Bad	Correct	0.64 ± 0.18
NDI-V	H2O	5	Bad	Correct	0.72 ± 0.17
2Nap-VL	DMSO	5	Bad	Correct	0.96 ± 0.02
6Br2Nap-VF	DMSO	5	Bad	Correct	0.64 ± 0.22
1Nap-FL	DMSO	5	Bad	Correct	0.96 ± 0.02
2Nap-IV	DMSO	5	Bad	Correct	0.96 ± 0.01
6Br2Nap-VF	DMF	5	Bad	Correct	0.54 ± 0.37
2Nap-FI	DMF	5	Bad	Correct	0.93 ± 0.07
7MeO2Nap-FV	DMF	5	Bad	Correct	0.95 ± 0.06
C6-NDI-F	H2O	10	Bad	Correct	0.5 ± 0.34
NDI-L	Ethanol	5	Bad	Correct	0.67 ± 0.2
NDI-I	THF	5	Bad	Correct	0.49 ± 0.22
C6-NDI-F	H2O	5	Good	Incorrect	0.75 ± 0.17
C6-NDI-L	H2O	5	Good	Incorrect	0.47 ± 0.29
NDI-I	H2O	5	Good	Incorrect	0.56 ± 0.29
C6-NDI-F	THF	5	Good	Incorrect	0.68 ± 0.07
6MeO2Nap-FI	THF	5	Good	Incorrect	0.89 ± 0.08
PBI-L 1Py	DMF	5	Bad	Incorrect	0.47 ± 0.1
NDI-Y	H2O	10	Bad	Incorrect	0.26 ± 0.08

Validation prediction of the H_2O model

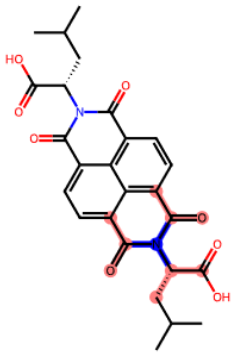
Solution Constitution				Prediction	Combined Confidence
Solute	Solvent	Concentration	Scattering Quality		
PBI-V	H2O	5	Good	Correct	0.71 ± 0.21
PBI-F	H2O	5	Good	Correct	0.92 ± 0.04
NDI-GF	H2O	10	Good	Correct	0.75 ± 0.2
NDI-V	H2O	5	Bad	Correct	0.9 ± 0.07
NDI-M	H2O	10	Bad	Correct	0.61 ± 0.31
C6-NDI-F	H2O	5	Good	Incorrect	0.73 ± 0.08
C6-NDI-L	H2O	5	Good	Incorrect	0.55 ± 0.1
NDI-I	H2O	5	Good	Incorrect	0.76 ± 0.09
2Nap-FI	H2O	10	Good	Incorrect	0.64 ± 0.15

A 4.12 Feature Importance

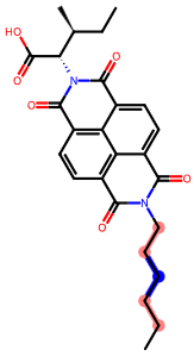
Beeswarm plots of the Shapley values of the 12 most important descriptors of the XGBoost model of the final QSPR models. A datapoint with a positive Shapley value indicates that the descriptor contributes to a “Good Scattering” classification, and *vice versa*. Datapoints are coloured based on their normalised value with purple representing a descriptor value of 0, red representing a descriptor value of 1, and intermediate values assigned a colour on the spectrum between these extremes. Representative molecules from the dataset with relevant structural feature highlighted are provided for the fingerprint bits described. The core atom represented by the fingerprint bit of interest is highlighted in blue, with the extended connectivity captured by the structural descriptor highlighted in red.



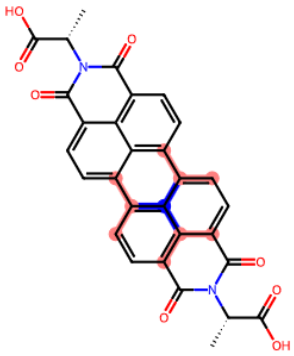
Bit 63



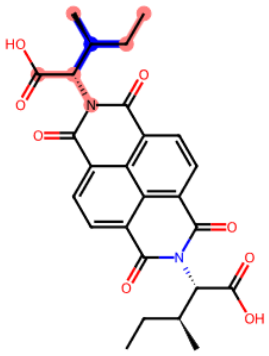
Bit 98



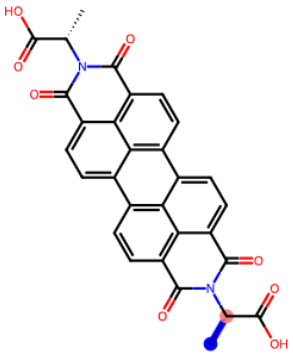
Bit 119



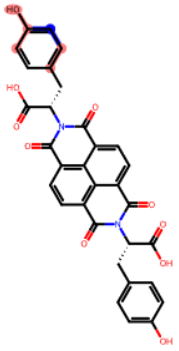
Bit 145



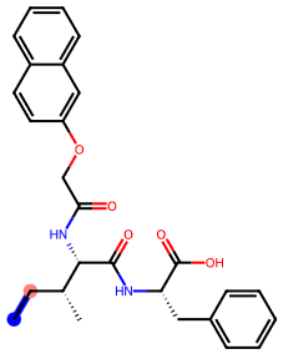
Bit 268



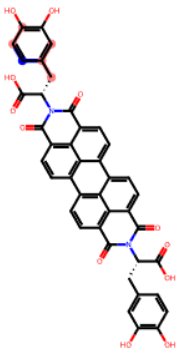
Bit 283



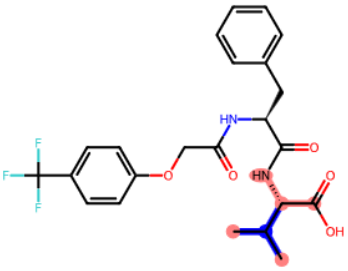
Bit 289



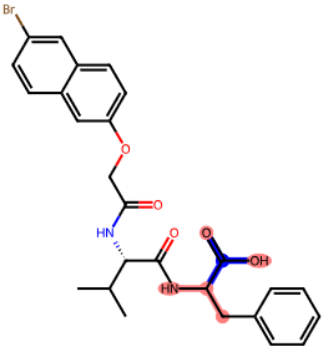
Bit 294



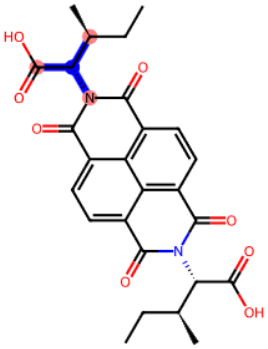
Bit 310



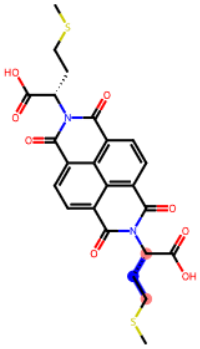
Bit 507



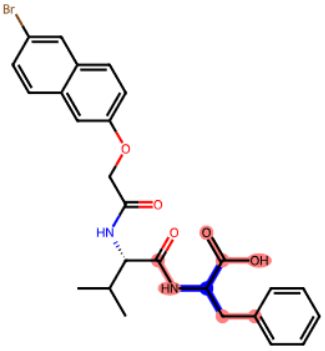
Bit 600



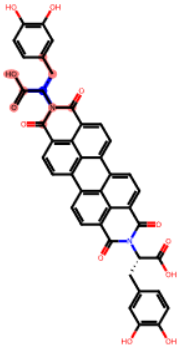
Bit 635



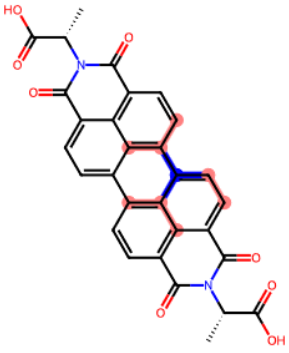
Bit 739



Bit 851

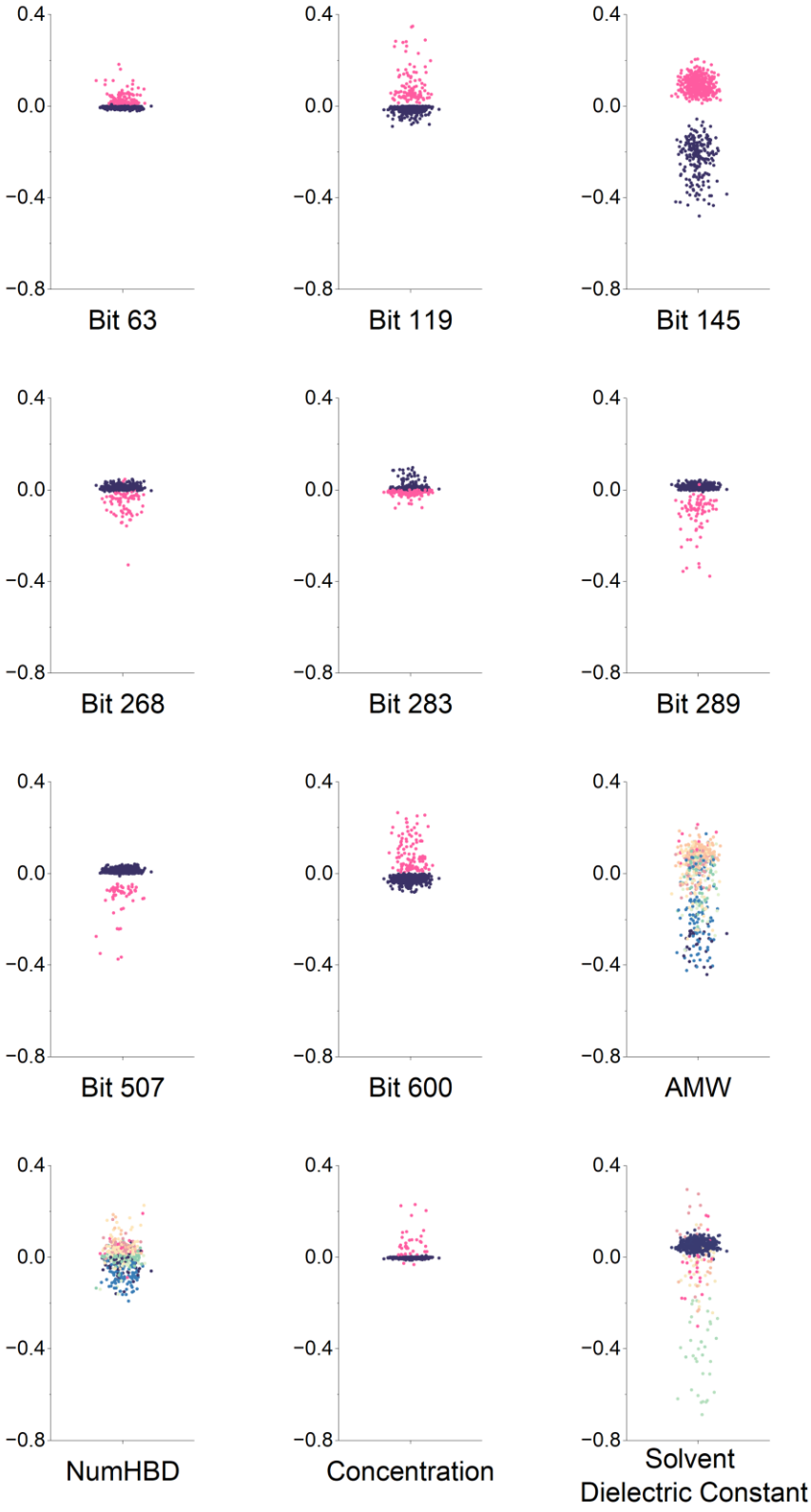


Bit 892

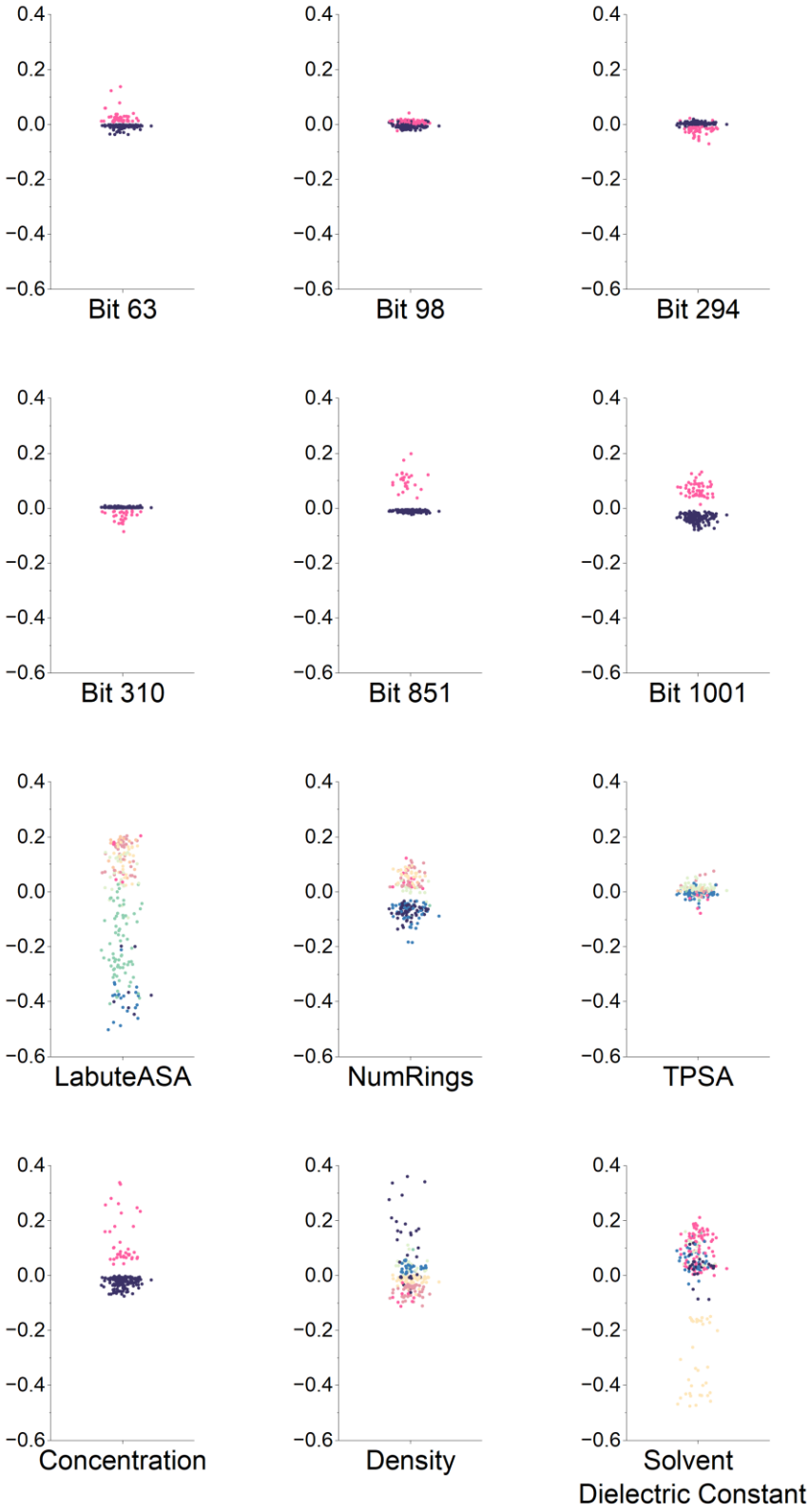


Bit 1001

Important features of the XGBoost model for the “All solutions” dataset.



Important features of the XGBoost model for the “Single component” dataset.



Important features of the XGBoost model for the “Aqueous” dataset.

