UNCOVERING THE MUTATIONAL LANDSCAPE OF SARS-COV-2 USING MACHINE LEARNING METHODS

KIERAN DANIEL LAMB

Submitted in fulfilment of the requirements for the

Degree of Doctor of Philosophy

College of Medical, Veterinary and Life Sciences

University of Glasgow

TABLE OF CONTENTS

FIGURES LIST

TABLES LIST

ABBREVIATIONS

Abbreviation	Definition
ACE-2	Angiotensin Converting Enzyme 2
ADAR	Adenosine Deaminase Acting on RNA
AI	Artificial Intelligence
APOBEC	Apolipoprotein B mRNA Editing Catalytic Polypeptide-like
BERT	Bidirectional encoder representations from transformers
BLAST	Basic local alignment search tool
CBOW	Continuous bag of words
CLS	Classification
COSMIC	Catalogue of somatic mutations in cancer
CoV	Coronavirus
COVID	Coronavirus Disease
CTD	C-terminal domain
DF	Degrees of freedom
DL	Deep Learning
DMS	Deep Mutational Scan
DMV	Double-membrane vesicles
DNA	Deoxyribonucleic acid
EOS	End of sequence
ESM	Evolutionary scale model
ESSM	Enviroment specific substitution matrix
FCS	Furin cleavage site
FEL	Fixed Efects Likelihood (Program)
GISAID	Global Initiative on Sharing All Influenza Data
GPT	Generative pre-trained transformers
HDI	Human Development Index
HIV	Human immunodeficiency virus
HR1	Heptad repeat 1
HR2	Heptad repeat 2
ICTV	International committee on taxonomy of viruses
IFITM	Interferon-induced transmembrane
IFN	Interferons
ISG	Interferon stimulated gene
JAK	Janus kinase
KNN	K-nearest neighbours
LBA	Long branch attraction
LLM	Large language model
MASK	Mask token
MAVS	Mitochondrial antiviral-signalling proteins
MEME	Mixed Effects Model of Evolution (program)
MERS	Middle east respiratory syndrome
ML	Machine learning

MLM	Masked language model
MSA	Multiple sequence alignment
NCBI	National centre for biotechnology information
NLP	Natural language processing
NMF	Non-negative matrix factorisation
NNLS	Non-negative least squares
NSP	Non-structural protein
NTD	N-terminal domain
ORF	Open reading frame
OWID	Our world in data
OxCGRT	Oxford COVID-19 Government Response Tracker
PANGO	Phylogenetic assignment of named global lineages
PC	Principal component
PCA	Principal component analysis
PDB	Protein data bank
PLM	Protein language model
PPI	Protein-protein interaction
PSSM	Position Specific Scoring Matrices
QK	Query-Key
RBD	Receptor binding domain
RBF	Radial basis function
RIG	Retinoic-acid-inducible gene
RLR	RIG-like receptor
RNA	Ribonucleic acid
ROS	Reactive oxygen species
S.E	Standard Error
S1	Subunit 1
S2	Subunit 2
SARS	Severe acute respiratory syndrome
STAT	Signal transducers and activators of transcription
std	Standard
SVM	Support vector machine
SVR	Support vector regression
T-SNE	T-distributed stochastic neighbour embedding
TCGA	The cancer genome atlas
TLR	Toll-like receptors
TMPRSS2	Transmembrane serine protease 2
TOM70	Translocase of outer mitochondrial membrane protein 70
UMAP	Uniform manifold approximation and projection
UV	Ultra-violet
VOC	Variant of concern
VUI	Variant under investigation
WeSME	Weighted Sampling based Mutual Exclusivity
WHO	World health organisation
ZAP	Zinc finger antiviral protein

ACKNOWLEDGEMENTS

While the writing of this thesis was a task just for me, the years of work that form its content would not have been possible without help from the many wonderful people I met during my time as a PhD student.

First and foremost, I'd like to thank my supervisors Ke Yuan and David Robertson for guiding me through my formative years as a researcher. I've been incredibly lucky to work on such interesting questions and both of you have continued to push me to keep asking more. You have been exceptional role models for me, and I am grateful for the advice and the time you have both given me to grow.

I have been privileged to meet and get to know a number of fantastic researchers while at the CVR. To Joseph Hughes, Richard Orton, Joe Grove, Sarah Cole, Lea Meyer, Mazigh Fares, Ed Hutchinson, Orges Koci and many others, thanks for being such great colleagues and for all of the advice and assistance over my time here.

I'd like to thank all the members of the Robertson lab past and present: Francesca Young, Sejal Modha, Spyros Lytras, Dan Liu, Haiting Chai, Ewan Smith, Martha Luka, Rozeena Arif and Rubayet Alam. I was immediately made to feel at home within the lab despite my first few years being almost entirely on Zoom thanks to a certain pandemic. Now with many more days in the office, it's a privilege to come in and chat science with you all. Spyros, I could not have asked for a better friend and mentor through all of it. You taught me so much, from work relevant topics like phylogenetics to slightly

less work-related activities like how to properly enjoy a conference or how not to ride an electric scooter. I appreciate it all immensely. Fran, thanks for listening to me rant about everything at my desk and constantly distract you from your own work. I'm thankful to have such an excellent postdoc in the lab, and appreciate the many long chats figuring out everything from what transformers are to what meeting room we're meant to be in. Dan, it's been great having another PhD student to work with and its always nice to see a friendly face when we end up working late in the learning hub. Extra special thanks to at this point an honorary member of the lab, Alex Pancheva. From teaching me Java back in the second year of my undergraduate, to working together on lots of fun (and sometimes not so fun i.e. slurm) science, you've been a friend and mentor to me for years now and I appreciate it greatly.

I've been a part of the organising committee for the computational biology conference from the start of my PhD and I'd like to thank everyone who made that a fantastic experience including Alex, Fran, Jake Lever, Vinny Davies, Olympia Hardy, Ross McBride, Holly Hall, Lucas Farndale and John Cole.

To my favourite bioinformaticians who made my masters experience so special despite the pandemic, thank you so much to Olympia, Salvatore Esposito, Lee Oo, Paul Craig, Kiron Roy, Ioulia Tsatsani and Mania Kakavouli.

To my school friends Craig Campbell, Cameron McKay, Conor McKee, Ronan Diver, Chris Boland, Rachel O'Donnel, Conor Drummond, Meysoun Khan, and Juliet Adie, thanks for putting up with me for all these years.

To my wonderful PhD pals Anna Kirk, Jake MacLeod, Faye Watson, Hollie French, Matt Arnold, Harry Scott, Eilidh Rivers, Lois Mason, Kasim Waraich, Daniel Weir, Kelsey Davies, Holly Ireland, Andy Clarke, Innes Jarmson,

Stephen Devlin, Alex Wilson, Delia Cretu, Marianne Donald, Jack Frew, Cameron Best and so many others I'm forgetting. My time at the CVR would not be the same without you all. From loch swims to lunchtime debates on geese and everything in between, it's been a wild ride, and I couldn't have asked for better companions.

Thanks to Megan Saathoff and Stephanie Brown, my wonderful research assistants who are now going on to do PhD's themselves. Thanks for all of the hard work you did while working with me and I can't wait to see what you both do in your new labs.

To my Anna, getting to know you and work with you has been an absolute joy. You are the most incredible partner and friend, and this last year has been one of the best of my life. Whatever happens next, thank you so much for sticking through this experience with me. It means the world.

To Caitlin and Gemma, thank you both for being the greatest sisters a brother could have. You both have always been there for me, and I couldn't imagine having grown up without the pair of you.

To my Mum and Dad, Veronica and Steven. It's hard to put into words how thankful I am for everything you have given me. Without you both, I would not be where I am today. My love of research is a product of your parenting, and you are my biggest supporters in everything I do. Words can't express my gratitude for everything you have done for me before and during my PhD but thank you and I love you will have to do.

This thesis is dedicated to all of you.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation 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.

Kieran Daniel Lamb
June 2024

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative pathogen behind the Coronavirus disease 19 (COVID-19) pandemic. Following its emergence in Wuhan in the Hubei province of China, SARS-CoV-2 infected millions of people around the world and has since become one of the deadliest on record. As part of the pandemic response, an unprecedented number of viral genomes were sequenced to produce the worlds largest dataset of viral sequencing data. In this thesis, we used machine learning methods to discover more about the mutational landscape of the virus from this sequencing data. We use mutational signature analysis to discover the mutational processes providing the mutations SARS-CoV-2 uses to adapt and evolve over time. We show that these processes are dynamic, and shift in their activity throughout the pandemic. We show that different variants of concern (VOCs) show different levels of mutational process activity which may relate to differences between the intrinsic virology between these lineages. We next show how large language models (LLMs) that have traditionally been used in natural language processing (NLP) can be used to produce meaningful representations of viral proteins. These representations can distinguish between proteins from different virus VOCs, generate metrics that can evaluate every possible mutation in the protein, and even predict putative evolutionary trajectories that correlate with the real emergence dates. We also show that model logits identify epistatic interactions disturbed by mutations and identify positions of structural conservation. Much of this can be completed using a single sequence and can also be used in a surveillance scenario where new sequences can have their representations compared

against currently circulating or prior lineages. Finally, we show how identifying mutational patterns using co-occurrence highlights interesting pairs of mutations that may be selected for by the virus and its selective environment. Using mutational contexts, language models and the virus phylogeny, we can investigate how these mutations might benefit the virus and improve our understanding of how linked mutations appear in a circulating viruses. In summary, this thesis shows how techniques from machine learning can help us learn more about the evolutionary processes, dynamics and effects of changing viral proteins using genomic sequence data.

1 INTRODUCTION

The use of computational devices can be traced thousands of years into the recesses of human history. From the abacus and the Antikythera mechanism[1] (thought to be the world's oldest computer) to the conceptualisation and creation of Turing machines[1,2]; computational devices have long helped humanity solve its problems. Since the mid 20th century, research into machine learning (ML) has become a large part of modern computer science. The aim of machine learning is to develop methods that can use information to learn generalisable properties of some process or system. By using previous examples as input, machine learning methods aim to "learn" the function between inputs and outputs of the observed system. Assuming this function is generalisable, the method should be able to approximate the output of previously unseen inputs. Research into the action of neurons within the brain spurred on the creation of algorithms designed to mimic these functions. This formed the basis of neural networks, the artificial neuron[1]. Since then, the field has rapidly expanded to include hundreds of different methods and models designed to help understand complex systems.

Viruses are some of the smallest replicating entities known to exist. Unable to replicate themselves, viruses must infect a host organisms' cells to proliferate and produce progeny virus. Viruses exist in large numbers and have incredible genetic diversity[1]. Their short lifecycle, higher mutation rates and ability to replicate many times in a single generation make them interesting to observe from an evolutionary perspective[1]. Observable evolution in species like humans may take many generations, while viruses can adapt on much shorter timescales since their generation times are significantly shorter. This also

allows viruses to be highly adaptable, since external pressures that threaten a virus survival can be potentially overcome through the quick accumulation of adaptive mutation. The SARS-CoV-2 pandemic has shown how quickly viruses (particularly RNA viruses) adapt to the ever-changing immune landscape and a new host species, further emphasising the importance of studying viruses. From its emergence into a mostly naïve immune landscape (save for prior exposure to other coronaviruses[1]) in Wuhan, China following a likely zoonotic spill over[1,2], to the present day where vaccination and infection have provided much of the population some immunity; SARS-CoV-2 has continually acquired new mutations enabling its adaptation to this dynamic landscape. How viruses evolve, interact with their hosts, replicate, and adapt represent incredibly complex systems that we need to continue to improve our understanding of. This thesis will attempt to show how using machine learning methods can help us understand the complexity involved in the evolution and adaptation of a virus during a global pandemic. It represents the intersection of computer science and virology during an unprecedented time in history. The pandemic had an enormous effect on the trajectory of this PhD, with much of the work occurring during 2020 and 2021 during the peak of the global lockdowns. Many of the questions we chose to tackle arose from the rapidly evolving situation we and the rest of the world found ourselves in.

We will begin with an introduction to virology, with a focus on RNA viruses and in particular SARS-CoV-2. We will then discuss machine learning methods and how they might be applied to biological sequence data, before outlining the work that contributes to this thesis.

1.1 VIROLOGY

A virus is a intracellular parasite comprised of a DNA or RNA genome and contained within a protective protein enclosure[3,4]. They are parasitic entities that require a host to be able to replicate themselves. Viruses do not have their own metabolism, and thus rely entirely on the host cells metabolic processes and organelles in order to replicate their genetic material and repackage their progeny into complete viral particles (virions)[3,4]. Since the discovery of the tobacco mosaic virus by M.W Beijerinck in the late 19th century[11], we have learned that viruses were the causal infectious agent behind many diseases that impacted humans and animals throughout the ages. To this day, viruses remain an ever-present threat. The SARS-CoV-2[1], Influenza[1], and HIV[1] have all resulted in worldwide pandemics over the 20th and early 21st century. Viruses also infect nearly all forms of life including plants, animals, bacteria and archaea. As our world becomes more connected, this creates its own problems. Increasing contact at interfaces between humans and wildlife can result in the increased chance of viruses from other animal species spilling over and transmitting between humans to cause new diseases[5,6]. They can also impact the food chain, with plant viruses affecting several different vital crop plants[7]. Viruses are thought to be the most diverse and prevalent entities on earth[6], and as such there are many distinct categories of viruses that we know to exist.

1.1.1 The Baltimore System

Viruses are classified in many ways. The classifications aim to provide useful demarcations that help to group similar viruses together. Factors such as genetic relatedness, genomic material type (DNA/RNA), strandedness (single/double) or even infection phenotype (respiratory, haemorrhagic, etc…) are all valid approaches depending on the circumstances.

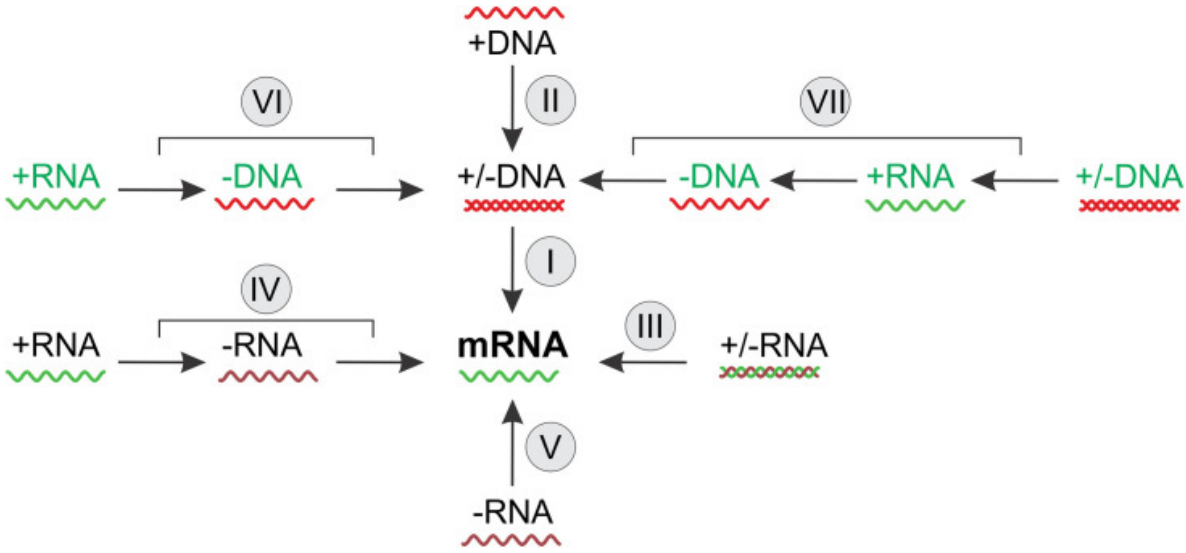


Figure 1.1: Diagram of the Baltimore classification system from Koonin et al[8]. Each roman numeral corresponds to the Baltimore classification.

One commonly used method for classification is the Baltimore system[8,9]. This divides viruses into 7 main categories based on the genomic material type and how the information from that material is transferred in cells via their replication cycle. Classes one, two and seven refer to DNA viruses while three-six refer to RNA viruses.

The DNA viruses are split into double-stranded DNA viruses (dsDNA Class I), single-stranded DNA viruses (ssDNA, Class II) and an extra seventh class that was discovered shortly after the classification was published[8]. dsDNA (Class I) viruses follow the traditional cycle of information within a cell. The DNA is replicated by the host DNA dependant DNA polymerase (DdDp), while viral

mRNAs needed to make viral proteins are transcribed by the host DNA dependant RNA polymerase (DdRp). ssDNA viruses (Class II) require the host DdDp to make the genome into dsDNA, but from there follow the same trajectory.

The RNA viruses are split by the strandedness as well as the strand sense, which is either positive (+) or negative (-). All RNA viruses (except for the retroviruses) require the translation of a virus encoded RNA dependant RNA polymerase (RdRp) that allows for replication of the viral RNA. For Class IV viruses (+ssRNA)**,** +RNA is the same sense as host mRNA and can immediately be used to for translation to make viral proteins, including the RdRp[1]. The RdRp can then transcribe the +RNA to make -RNA which forms a dsRNA intermediate with the +RNA[1]. This dsRNA can then be used to produce more +RNA strands which can either be packaged into new virions or be translated to create more viral proteins. This dsRNA intermediate is problematic, since it is quickly detected by host immune systems, so replication typically occurs within double membraned vesicles (DMVs) that help to evade from dsRNA sensing proteins from the host[10]. Class III (dsRNA) operate in much the same way, except without the need to produce the dsRNA intermediate since this is the starting point. To evade the immune system, the virus replicates within the virion rather than the DMVs of Class IV viruses. Class V viruses (-ssRNA) are packaged with the necessary RdRp required to transcribe +ssRNA before following the same processes as Class III viruses[1]. The retroviruses (Class VI and VII) operate by incorporating their genomes into the hosts DNA, which is then replicated using the hosts own proteins and mechanisms[1]. Class VI are +RNA while Class VII are dsDNA.

1.1.2 Virus Taxonomy and Evolution

While the Baltimore system is useful to identify broad virus categories, contemporary taxonomy now makes use of the genome sequences of virus to identify relatedness. The International Committee on Taxonomy of Viruses (ICTV) maintains the currently used viral taxonomy list[11]. Using SARS-CoV-2 as an example virus[12,13], taxonomy begins at its lowest level, and works up by joining related viruses together into increasingly diverse groupings.

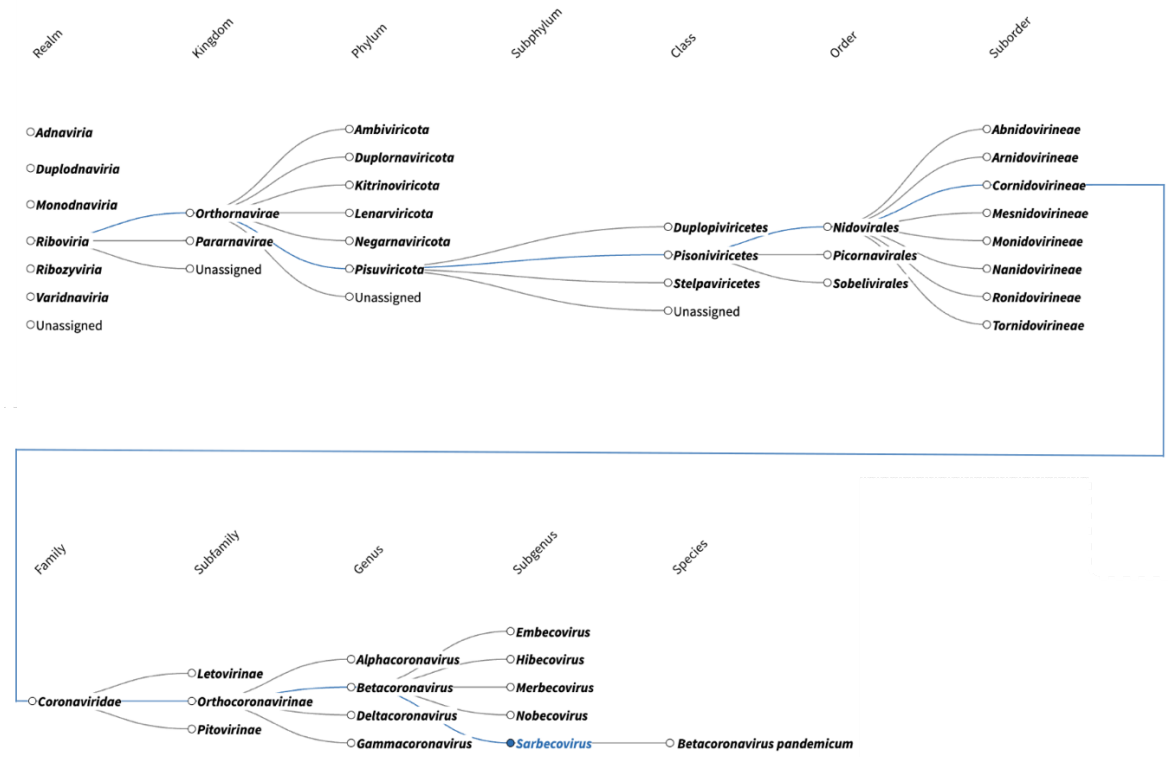


Figure 1.2: ICTV taxonomy tree levels of SARS-CoV-2. SARS-CoV-2 is a virus within the Betacoronavirus pandemicum species.

At the highest levels of the taxonomy tree, viruses are categorised by common features such as a shared gene (RdRp for *Riboviria*) or a nucleic acid type (*Monodnaviria* are all ssDNA viruses) in a structure similar to the Baltimore system. At these high levels (*Realm, Order* etc) the virus groups are often so divergent from one another that it is exceedingly difficult or impossible to join

based on a shared evolutionary history. Lower levels in the taxonomic tree are joined together using increasingly larger sets of shared features from individual nucleotides, genes and even sets of genes.

1.1.3 PHYLOGENETICS

To determine the evolutionary history between sequences, we can use phylogenetics. Since the famous early sketches by Charles Darwin, tree structures have been used to represent the relationships between different species[14,15]. Due to their hierarchical nature, trees are intuitive representations for describing patterns of evolutionary relatedness. In a phylogenetic tree (or phylogeny) species are represented as terminal (or leaf) nodes in the tree structure. Each terminal node is joined to an internal node by a branch. Internal nodes represent an unobserved common ancestor between either two leaf nodes, or a leaf node and another internal node. In this way, the phylogeny enforces a hierarchy with the topology of the tree describing how each species in the tree relates to the others. With traditional species phylogenies, each internal node represents a speciation event where a species has split into in two new species[1]. With genomic sequencing, these internal nodes can also represent where a new mutation occurs making the internal node the nearest common ancestor. The simplest phylogeny (Cladogram) therefore simply describes the branching pattern or speciation order[1]. Phylograms can extend this by scaling branch lengths to reflect the divergence between species[1]. The sum of the branch lengths then indicates the level of divergence between the two species.

For many years, phylogenies described groupings of species often by morphological or phenotypical traits (such as Darwin's finches[16]). With the

advent of modern sequencing technologies, it is more common to make phylogenies using genome sequences of the species in the tree. However, this is not as simple as aligning the full genomes of different species, since genes often have independent evolutionary histories. Phylogeography can be used to supplement sequencing by using the geographic information of the sampled sequences[17]. This allows for models to be constructed that consider how geography can affect the evolutionary history of a sequence, which can be very useful when tracking the spread of viruses, particularly when geographical features like altitude, bodies of water, or transport links can be driving factors[18,19].Molecular phylogenetics and phylogeography can give a much better estimation of the relatedness between species than a shared phenotype, since this can arise independently through convergent evolution. Sequencing data also allows for an estimation of divergence since the difference between 2 sequences can be calculated using the difference between their nucleotide or amino acid compositions. If substitution rates for the sequences are known or inferred, then sequences can also be dated, and speciation events estimated.

1.1.4 Sequence Alignment

Phylogenetic trees inferred from sequencing data are built using multiple sequence alignments (MSA)[1]. Sequences that are related share homologous sites i.e. characters that are conserved between sets of sequences. These sites are important since they allow sequence fragments that have different lengths and characters to be lined up such that their similarities and differences can be identified. The more distantly related 2 sequences are, the more difficult they are to align since there are typically fewer homologous sites. Alignments of distantly related sequences often improve with more sampling, since

sequences that are more closely related to other sequences can help to "fill in" missing evolutionary events such as substitutions, insertions, deletions, or duplications. This means that while sequence data is more informative than observable traits, alignments still represent an evolutionary hypothesis of how sequences of species are related. New sequences can change this hypothesis, and ultimately also change the phylogeny that is produced by it.

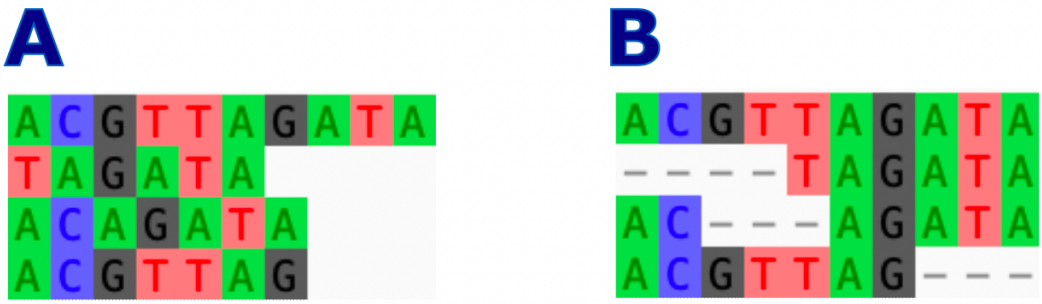


Figure 1.3: (A) Unaligned nucleotide sequences. (B) Aligned version of the same nucleotide sequences.

1.1.5 Phylogenetic Methods

Once an MSA has been created, there are several methods that can be used to create a phylogenetic tree. These can be broadly characterised as either distance methods of as character methods[15]. Distance based approaches first assume a substitution model for the nucleotides or amino acids. This is necessary as there are observable differences in the substitution rates between different nucleotides and amino acids that are determined by several properties including their biochemistry and their position within the sequence. One of the earliest known examples of this phenomenon was the discovery of transition biases in DNA. The four DNA nucleotides (adenine, guanine, cytosine, and thymine) can be broken into 2 classes of nucleotide called purines and pyrimidines that are based on their structures (Figure 1.4). It was

discovered that substitutions that involve a change within a structural class (transitions) were more frequently observed than substitutions between structural classes (transversions)[20,21]. Further evidence of non-uniform substitution biases can be found between nucleotide contexts, genes, and even whole species.

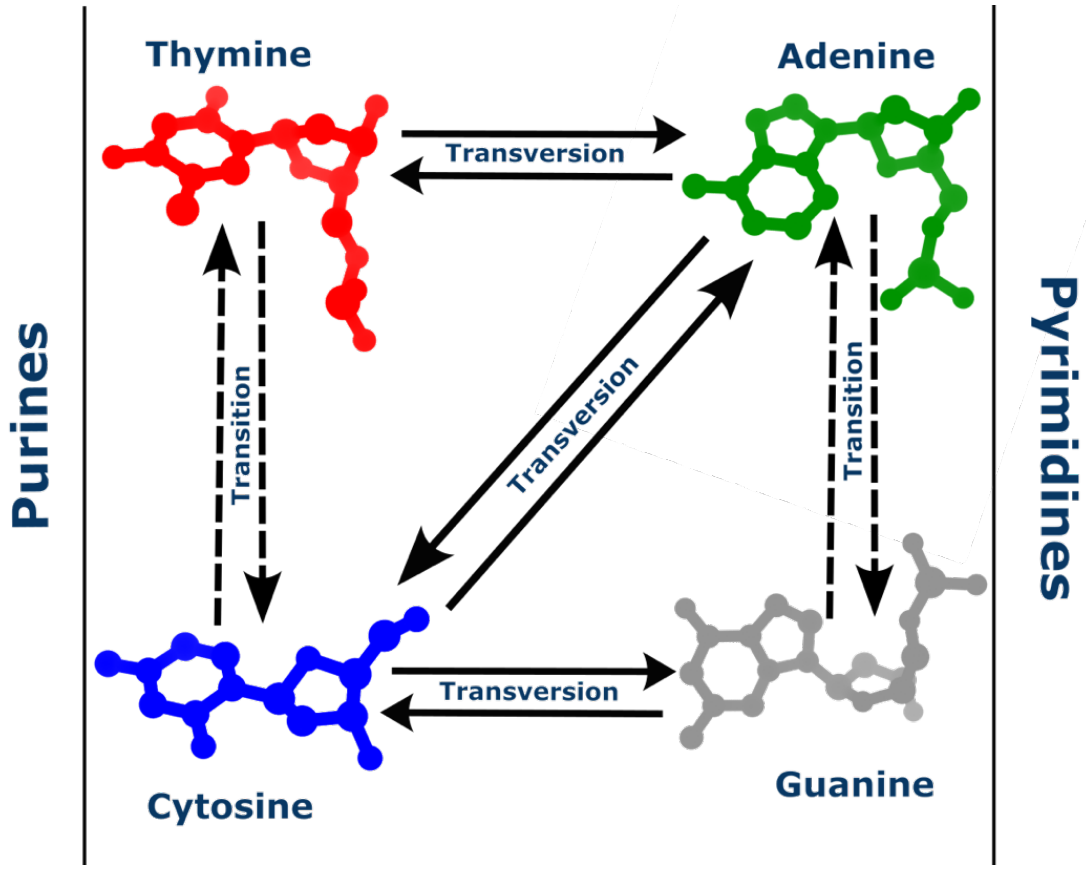


Figure 1.4: Nucleotide structures with arrows showing the substitution types. The left side are the purine nucleotides which contain a single ring structure attached to a sugar phosphate group. The right side are the pyrimidines which contain two ring structures attached to the sugar phosphate backbone.

Several different substitution models exist to accommodate these different biases. Once a model has been selected, the alignment can be used to calculate distances using the substitutions and the rates for each substitution category.

A common distance-based method is the neighbour joining tree[1], which constructs the tree using agglomerative clustering. This groups sequences into "neighbourhoods" using the distances, before joining neighbourhoods together into increasingly larger sets until all sequences are joined together. The agglomerative method inherently creates a hierarchy which intuitively forms a tree structure. The benefit of distance methods is that they are very fast, which becomes important as the number of sequences in the tree increases. More complex methods often use methods like neighbour joining to build initial trees to iterate upon and reduce the number of possible trees to check. The most simplistic character-based method of tree construction is maximum parsimony. This method takes an Occam's razor like view of evolution and assumes that the tree topology explained by the fewest substitutions is the most likely tree[15]. It does not usually make use of a substitution model and as such has no underlying assumptions about substitution rates. As such, maximum parsimony is simple to implement and understand but is potentially overly simplistic in describing more complex trees. If there are biases that are well understood and could be useful in building the correct tree, there is no way to incorporate these using maximum parsimony since all substitutions are considered equal. As such, other character-based methods like maximum likelihood or Bayesian approaches are typically used over maximum parsimony. A weighted parsimony method was introduced to tackle some of these issues, however both maximum parsimony and weighted parsimony suffered from a phenomenon called long branch attraction (LBA)[22,23] (although other approaches can also suffer from this problem as well). LBA occurs when samples in the tree that are very divergent are grouped with a sample despite

not being related to each other. This is because by chance a divergent lineage may contain substitutions that are contained in other parts of the tree yet were acquired independently such that the tree topology is then inferred incorrectly[1]. Long branches are often located near more basal nodes in the tree, since often trees are routed on outgroup sequences or samples. Maximum likelihood and Bayesian phylogenetics approaches are the current state of the art for phylogeny construction. They expand upon maximum parsimony by use a more statistical model allowing for parameters that describe properties of the tree rather than simply minimising substitutions. This gives both approaches a significant advantage over maximum parsimony, however both methods are significantly more time consuming, and as such are difficult to apply to larger datasets. This is becoming increasingly problematic as sequence datasets are growing increasingly large, especially in the wake of the SARS-CoV-2 pandemic.

Since the first full virus genome (called bacteriophage MS2) was sequenced[24] in 1976 by Walter Fiers using Sanger sequencing[1], the rapid development of sequencing technologies has enabled the first large-scale use of virus sequencing that occurred during the SARS-CoV-2 pandemic. Between December 2019 and January 2024 more than 16 million SARS-CoV-2 were deposited in the Global Initiative on Sharing All Influenza Data[25] (GISAID) database, dwarfing the size of the previously most sequenced virus Influenza with ~1.2 million sequences collected over decades in the NCBI virus database. Metagenomic sequencing is now a possibility as well, with many studies expanding our current view of viral diversity[26,27] on a regular basis.

1.1.6 SARS-CoV-2

SARS-CoV-2 is an enveloped single-stranded positive-sense RNA virus and causative agent of the COVID-19 pandemic[1,2,28]. It emerged in Wuhan, China following most likely zoonotic spillover event[29,30], before quickly spreading within the city and subsequently the rest of the world. The virus is a member of the *Coronaviridae*, a family of viruses first characterised by their distinctive protein projections resembling a solar corona under a microscope[31,32]. Coronaviruses are known to infect humans and include a number of common cold viruses as well as the epidemic viruses SARS-CoV and MERS-CoV. SARS-CoV-2 is most closely related to SARS-like coronaviruses that had previously been found in bats, although the animal reservoir of the virus has yet to be identified[28,33]. The virus has a genome of ~30k bases which encodes 4 structural proteins (membrane (M), envelope (E), spike (S) and nucleocapsid(N)) and several non-structural proteins (Figure 1.5A). Key among these is the S protein, which is the virus's glycoprotein that mediates cell receptor binding and entry.

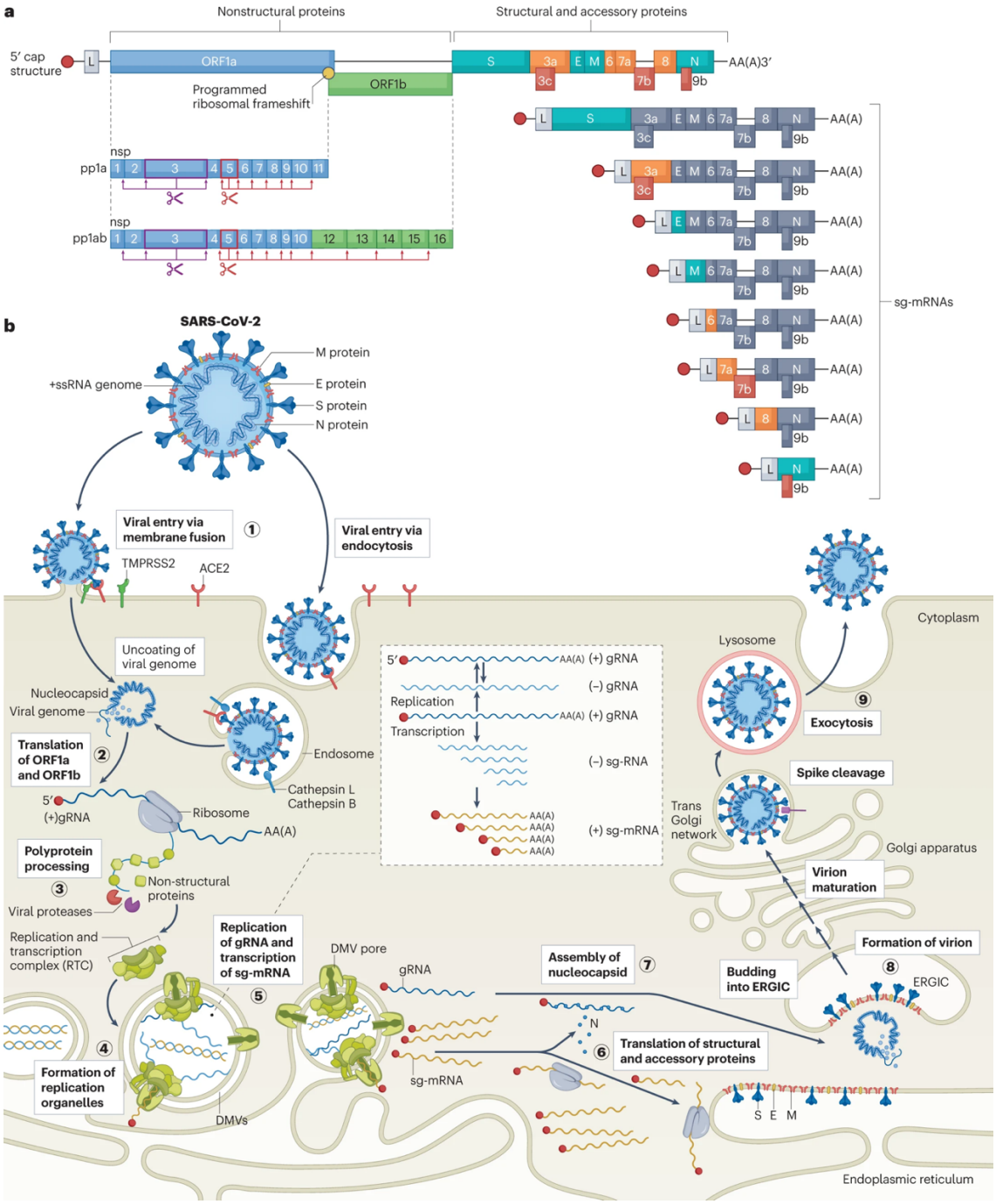


Figure 1.5: Diagram from Steiner et al.[34] showing the SARS-CoV-2 genomic structure, virion structure and cellular lifecycle. (A) shows how the SARS-CoV-2 genome is arranged into its different ORFs, proteins and sub-genomic mRNAs. (B) shows how the virion is constructed from the different structural proteins. It then shows how the virus enters the cell, and is unpackaged, translated, transcribed, replicated, re-packaged and released from the cell.

1.1.6.1 Cell entry

The SARS-CoV-2 spike is a trimeric protein with 2 major subunits S1 and S2. S1 contains the receptor binding region (RBD) while S2 contains the fusogenic region that allows for membrane fusion. The virus can enter cells via 2 different pathways: the cell surface entry or endosomal entry[34,35] (Figure 1.5B). The cell surface entry first involves the S protein RBD binding to the human angiotensin-converting enzyme 2 (ACE-2) receptor on the cell surface. The binding of the RBD to the cell ACE-2 receptor triggers a conformational change that exposes a cleavage site S2′. Cleavage sites are regions of a protein that promote the binding of proteases, cellular enzymes that cut protein chains. Cleavage of S2′ by transmembrane serine protease 2 (TMPRSS2)[34–36] initiates a further conformational change that permits membrane fusion using the S2 subunit and injection of the viral genome into the cell cytoplasm[34,35]. The endosomal entry route also involves the spike binding to ACE-2, except instead of cleavage at the membrane, the ACE-2 bound virus is taken into the cell via endocytosis[37]. The cell membrane envelopes the virion into an endosome, effectively swallowing it and bringing the endosome into the cell cytoplasm. Next, another family of proteases called cathepsins are used to cleave the S2′ site and trigger endosomal fusion with the virion and release of the viral genome into the cytoplasm.

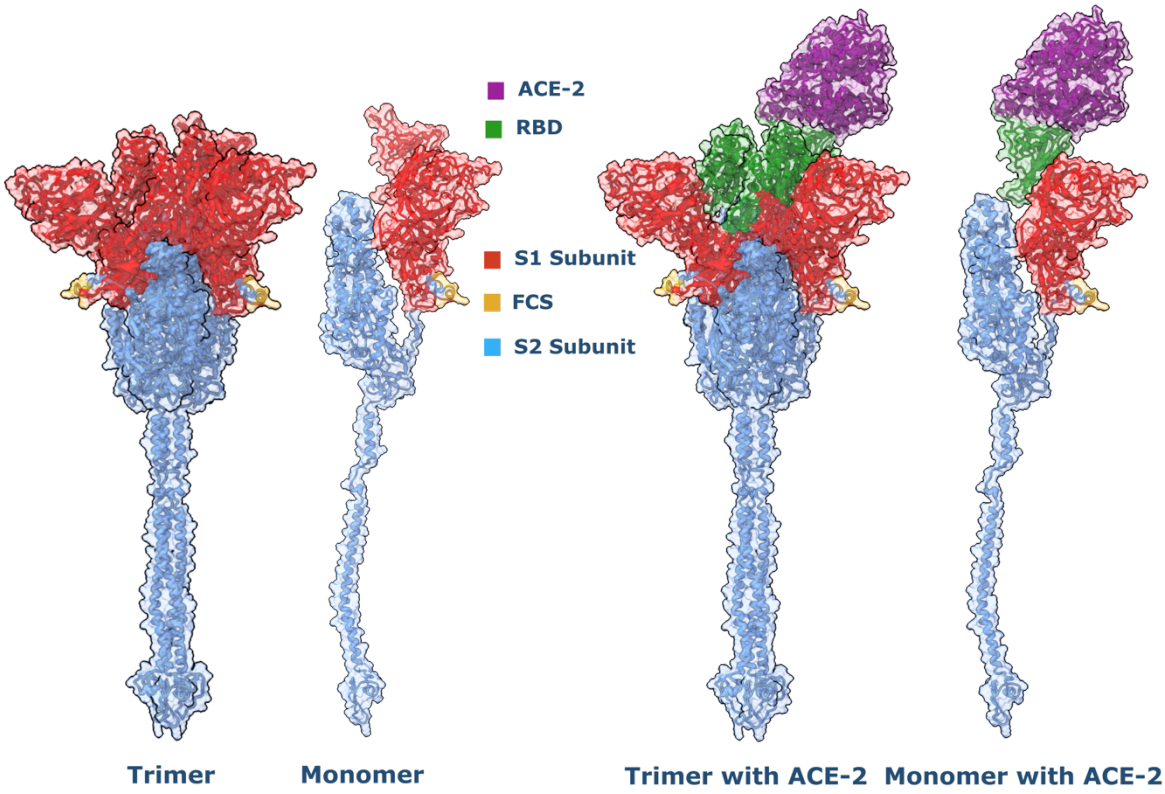


Figure 1.6: Spike protein of SARS-CoV-2. S1 subunit is coloured in red, the furin cleavage site (FCS) in orange, the S2 subunit in blue, receptor binding domain (RBD) in green and ACE-2 in purple. Trimeric spike and monomeric forms of the protein are both shown. Figure made in Chimera[38], with structures from Woo et al[39].

The spike protein was the target of much of the controversy around the viruses' origins due to the presence of a polybasic furin cleavage site (FCS) at the S1/S2 boundary of the protein. While other human coronaviruses were found to contain these sites, they have not been observed in betacoronaviruses like SARS-CoV-2[40]. Furin proteases cleave the covalent bonding between the S1 and S2 subunits of the S protein before newly formed virions leave the cell. Prior experimental evidence has shown that insertion of an FCS into SARS-CoV can enhance cell-cell fusion[40,41] and is in fact required for SARS-CoV-2 to enter human lung cells[42]. Similarly, an FCS was shown to be necessary for

SARS-CoV-2 to become transmissible in ferrets, which are useful animal models of diseases in human[36]. The FCS appears to allow the spike protein to more easily enter an "up" conformation which permits receptor binding of the RBD to ACE-2[43].

1.1.6.2 Virus replication and translation

Once the viral genome has entered the cytoplasm of the cell, it is uncoated, and the ORF1a/b polyprotein is translated and proteolytically cleaved into the viruses' non-structural proteins (NSPs). ORF1a/b has two translated forms due to a -1 slippage point being present between what should be the end of ORF1a and the beginning of ORF1b[1]. The presence of a 7 nucleotide "slippery" sequence followed by a RNA pseudoknot results in the ribosomes involved in translation of the ORF slowing down, and either backtracking by 1 nucleotide which allows continued translation over to ORF1b, or continuing as normal and terminating at the stop codon of ORF1a[1,2]. This frameshift into ORF1b allows for translation of NSPs 12-16. The frameshift occurs between approximately 57% +/-12% of the time[1], allowing modulation of 2 ORF1a/b forms which has been shown to be necessary for viral fitness[1–3].

Since SARS-CoV-2 is a +ssRNA virus, the viral RNA can immediately be translated. Among these NSPs are the viral replication proteins. The virus then forms double-membraned vesicles (DMVs) to hide the viral replication from the host innate immune responses. Within the DMV, the viral replication and transcription complex (made from NSPs 7,8 and 12) is used to produce negative sense RNA for further positive sense RNA genome production. This replication stage may produce dsRNA complexes in the DMVs, although it is unknown whether this occurs during the replication phase or subsequently[44].

Regardless of how they are formed these dsRNAs are a distinctive sign of viral replication and are a useful signal for host cells.

1.1.6.3 Host immune signalling and responses

Viruses compete for host cell resources and as such are parasitic entities. In response, cellular hosts have evolved ways of detecting viral infection and producing an immune response to clear the host of the virus. There are two categories of host immunity, the first is known as innate immunity while the second is known as adaptive immunity. Innate immunity is the first line of defence for a host. It is made up of a mixture of physical barriers and non-specific antiviral molecules that react to viral infection. For respiratory viruses like SARS-CoV-2, one of the initial innate immune barriers is the mucus that lines the surface of the airways[45]. Mucus is a physical barrier that prevents particles from reaching the cell surfaces and contains a mixture of proteins that can bind to viral glycoproteins to prevent receptor binding. When the virus does enter the cell, a host of sensing molecules known as pattern recognition receptors (PRRs) are used to recognise classical signs of viral infection. For SARS-CoV-2, these are thought to be RNA intermediates forming during replication (dsRNA,-ssRNA, stress granules) and that are recognised by RIG-like (RLR) and Toll-like (TLR) receptors[46]. The DMV formation during the replication cycle of SARS-CoV-2 is one method the virus uses to avoid these receptors, however many of the NSPs also appear to prevent sensing. This can be via direct antagonization of the PRRs such that they are unable to bind to the viral RNAs, or by preventing overaccumulation of replication intermediates[46]. Detection by PRRs triggers a cascade of activity that results in the production of interferons (IFNs). IFNs are a family of cytokine signalling molecules that trigger the activation of interferon-

stimulated genes (ISGs) and are released by infected cells. ISGs have many functions but they primarily prepare cells for viral infection by producing antiviral proteins and changing the cells state (i.e. altering cellular processes such as protein synthesis and growth) to best respond to infection[47]. Many ISGs produce proteins that directly interact with the viral genomic material and degrade, edit, or restrict it. These include proteins such the zinc-finger antiviral protein (ZAP), which restricts viral RNA by binding to CpG rich regions and preventing translation[48] or members of the Apolipoprotein B mRNA Editing Catalytic Polypeptide-like (APOBEC) family of editing enzymes that induce cytidine deamination of the viral genome[49]. Other ISGs such as the IFN-induced transmembrane proteins (IFITM) prevent viral entry particularly via endocytosis[49]. SARS-CoV-2 like many other viruses is targeted or interacts many of these different ISGs.

Cells that have already succumbed to infection by the virus can also alert the body using the major histocompatibility complex (MHC) class 1 pathway[1–3]. This pathway is used by cells to present antigens that come from the infecting virus on the cell surface. Viral antigens (like the SARS-CoV-2 spike protein) are broken down in the cell cytosol into small peptide chains that are carried to the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP) protein. The MHC class 1 proteins (a dimeric protein made up of a heavy chain polypeptide and a β_2M fold protein) are bound together in the ER, before associating with TAP which joins the MHC class 1 complex to the viral antigen peptide. This is then transported to the cell surface where it is presented on the cell surface, and can be bound by the T cell receptors (TCRs) of cytotoxic T cells[1,2]. These T cells on recognition of the MHC class 1 complex subsequently destroy the infected cell[1].

1.1.6.4 Adaptive immune response

Upon host identification of viral infection, the adaptive immune response kicks in. This is a complex mechanism of systems, cells and signalling molecules that help to prevent virions from infecting cells, as well as dealing with the infected cells that have already been overwhelmed. The neutralisation of viral particles before they infect cells is primarily of interest here.

Virions are targeted by the adaptive immune response upon recognition of their surface receptor molecules. In the case of SARS-CoV-2 this is primarily the spike protein. The spike protein protrudes from the virion surface and can be recognised by host immune cells such as B-cells. B-cells recognise antigenic markers (antigens) on the surfaces of invading entities and can generate proteins (antibodies) that bind to them. These antibody proteins can be used to neutralise the virus by preventing it from binding to cellular receptors or to signal other immune cells that the virus needs to be disposed of.

Since the spike is the primary antigenic protein, it is under intense selective pressure to change. Antibodies typically bind to small epitopes on the protein surface and are specific to this region. As such, mutations in these regions can result in the antibodies failing to bind and prevent neutralisation of the virus. This makes these regions of the protein subject to substantial selective pressures, and as such they typically change faster than other regions of the protein. The SARS-CoV-2 vaccines currently available all target the spike protein, so understanding how the protein changes and adapts overtime is important for understanding our immunity to the virus.

1.1.6.5 SARS-CoV-2 Evolution

SARS-CoV-2 has continued to evolve following its emergence into humans[1–4]. Due to short generation times and quick (yet erroneous) replication at each generation, viruses evolve at timescales that are perceivable to us, and thus can be studied and observed in near real time[1]. RNA viruses in particular change rapidly due to their error prone polymerase and (often) lack of any proofreading mechanism or error correction[50]. This inevitable introduction of error seems undesirable, however it in fact allows the virus to increase its diversity overtime which can help it adapt to a changing environment. SARS-CoV-2 has a large genome for an RNA virus, possible due to the presence of a proofreading enzyme[51–53]. Mutation rates that are too high can render the virus inactive by introducing too many deleterious mutations per replication cycle, and longer genomes mean an increased likelihood of a new mutation per genome[1]. While most mutation observed is thought to be neutral or approximately so[54], SARS-CoV-2 has seen a striking amount of adaptive change, particularly within the spike protein. While mutations may appear random, selective pressures mean that truly deleterious mutations are often removed by purifying selection and are never propagated. Alternatively, they are positively selected for due to the mutation providing a selective advantage. An initial introduction into a new host followed by an increasingly less naïve immune landscape has meant the virus has been subjected to a series of powerful environmental selective pressures. Selective pressures can be observed particularly through the phenomenon of convergent evolution, where the repeated occurrence of the same substitution or phenotype is generated from different evolutionary trajectories. In its most simple form, this is the same nucleotide, however convergence can occur across the different levels of

life. Multiple nucleotide substitutions can result in the same amino acid being translated, different combinations of amino acids can produce the same secondary structure and folds, and entirely different proteins can evolve to perform similar or identical functions[1–4]. Throughout the SARS-CoV-2 pandemic, there have been several mutations (particularly within the Spike protein) that have convergently appeared both in circulating lineages[56–58] also in wastewater sampling[59]. Wastewater samples are of particular interest due to the presence of so-called "cryptic lineages" which are often hypermutated and contain mutations not often observed in the circulating lineages. This has led to hypothesise on their origin ranging from chronically infected patients to animal reservoirs to replication and lack of sampling of SARS-CoV-2 in the gastrointestinal tract[59]. Following on from the Omicron variant of concern's emergence, many of the cryptic lineages shared mutations with the emergent variant of concern suggesting that despite having a different ancestor, they may be subjected to similar selective pressures resulting in this convergent evolution.

1.1.6.6 SARS-CoV-2 lineage naming schemes and nomenclatures

During the pandemic, monitoring of SARS-CoV-2 sequences was undertaken in order to identify potential outbreaks of the virus[1,2]. It quickly became apparent that a naming scheme was necessary in order to identify specific genotypes of the virus without listing each of the mutations. The Pango lineage nomenclature[1,2] was created to label SARS-CoV-2 genotypes that were at least one mutation apart from a parental SARS-CoV-2 lineage, and that demonstrated onward transmission within a new location. This aimed to label lineages that were producing infection outbreaks in new locations, however

future lineage allocation became less focused on the epidemiological evidence due to the quantity of sequencing later in the pandemic. The hierarchical nomenclature used a combination of letters, numbers (i.e. A.1, B.1.1.1) as labels, and would dynamically reset to a new letter combination after 3 levels of numbers (e.g. B.1.1.7.1 became Q.1). While useful to researchers, the nomenclature was considered too complicated for use by public health officials for communicating information about the virus[1,2]. This prompted the creation of the Greek alphabet labelling of important variants, including the major variants of concern (Alpha, Beta, Gamma, Delta and Omicron) as well as some other variants of interest (VOIs). These variants differ from the Pango lineages, and often several Pango lineages are classified as the same WHO designation (i.e. BA.1 and BA.2 are both Omicron variants).

1.2 Machine Learning

Machine learning (ML) is a domain combining computer science and statistics that focuses on how prior data can be used to help analyse data. Machine learning can help explain relationships between variables, estimate values where they are missing and predict future or previous outcomes based on our data. Biological data is now at the forefront ML research since it can help with tasks like drug discovery and diagnosis of disease. With plentiful genomic data and healthcare records becoming increasingly digitised, ML could help usher in a future of precision medicine.

The application of ML in life sciences has never been more apparent than with the revolution in protein structure prediction ushered in by AlphaFold[60] and ESMFold[61]. However, these breakthroughs only represent the beginning of the challenges ML may be used to solve in future. Biological data happens to be inherently complex, high dimensional, and noisy which makes it a perfect candidate for ML techniques. Here, we will outline how machine learning can be applied in particular to biological sequence data. We will describe how we can use sequences to derive features for ML tasks and how these features encode meaningful signals of relevant biology. We will discuss some of the basic ML approaches, as well as how they can be applied to sequence data and what they can be used for.

1.2.1 Signals in the sequences

DNA, RNA, and amino acid sequences are often represented as strings of arbitrary characters despite these molecules each having unique biochemical and structural features. This representation omits much of what makes these molecules unique, yet these sequence representations are incredibly powerful

for many different tasks. As previously mentioned, sequence data is the foundation of molecular phylogenetics and allows us to infer the evolutionary relationships between biological entities. This signal comes from the homologous fragments of sequences found between related organisms, but there are many other signals that can emerge from sequence composition. Viruses contain many such informative signals. Much of viral evolution over time is driven by adaptation to the hosts they infect. Viruses with human hosts typically have suppressed CpG content in part because host anti-viral molecules like ZAP target their CpG rich regions in the genome[49,62]. Humans use DNA methylation of CpG sites as a mechanism to regulate gene expression, however subsequent deamination of cytosines can cause the cytosine to become thymine further suppressing CpGs[63,64]. Viruses that expose their genomes within the host cytoplasm have also been shown to have reduced CpG content vs viruses that do not, again showing the usefulness of sequence composition on understanding viral properties[65]. CpGs are one di-nucleotide category, however there are several other of di-nucleotide categories as well as other k-length categories. These k-length sequence features are called a k-mers, and they have been shown to be predictive for tasks such as virus host prediction[66,67], bacterial phenotypes[68] and the identification of viral sequences in metagenomic samples[69]. K-mers are useful since they are often biologically meaningful, informative and can be easily used in ML tasks. Domains are another useful sequence-based feature. Sequences typically contain functionally relevant sections that are known as domains. Presence or absence of notable domains can typically be used to make predictions on things like protein function or structure. Conservation of domains with uncertain

function can also be predictive of useful functionality. Again, these signals are all contained within the sequence representation of the organism. New sequence representations continue to expand the usefulness of sequence data including innovations like sequence embeddings and new sequence alphabets such as the 3Di alphabet used by FoldSeek, a tool for searching for sequences with similar folds[70]. The 3Di alphabet is created using an autoencoder to extract features of an amino acids local structural context and encode this as a 20-character alphabet. This makes it compatible with existing alignment tools since it still uses traditional characters, yet these characters now more directly encode information about the protein structure. There are many other sequence representations, but the main takeaway is that all of these representations contain useful signals which can be exploited by machine learning methods to accomplish various tasks.

1.2.2 Supervised Machine Learning

Supervised machine learning relies on labelled data being used to train models. The "supervisor" is essentially the data labeller, with the assumption being that they are a domain expert and thus know the correct label for each data point. With well annotated datasets, supervised ML often achieves excellent performance however labelling is time consuming and typically restricts dataset size as a result. While most applications strive to use gold standard datasets (i.e. in medical imaging a dataset labelled by a radiologist or clinician), this is often not achievable in an era of exponentially growing datasets. Despite these restrictions, where well labelled data is available a supervised approach is often effective.

1.2.2.1 Regression

If there is a need to predict continuous values from data, a regression is what will typically be used. A regression model is one of the simplest forms of supervised ML and is still used often due to its explainability and broad applicability to many tasks. The linear regression is the simplest of these models.

$$y = mx + c$$

This equation describes a line (if the data is in 2 Dimensional, this becomes a hyperplane in higher dimensions), with m representing the gradient, c representing the y-intercept and x and y being 2 sets of observed variables. By iterating through possible values of c and m and checking if those values minimise the error between points and the line (known as the least squares method), linear regression can estimate the relationship between these two sets of values. This is useful, since assuming the data used to generate the line is generalisable, the line can be used to predict values of y given any value of x. It is also easily interpretable, since there is a clear and obvious intuition i.e. given a change in the value of x, we would observe the value of y to be this. Extensions can be added in the form of multiple regression, where an additional gradient term is added to account for each new dimension, although fundamentally the equation is the same. There are many extensions that can be made to linear regression such as non-negativity constraints or using polynomials that allow them to work on different datasets.

1.2.2.2 Classification

Classification is another common ML task that involves placing an observation into a set of classes, rather than a prediction of a continuous value. At its most basic, this would be binary classification where there are only two classes, but

this can continue to multiple classification as well. There are several different methods that allow for classification including k-nearest neighbours (KNN) and support vector machine (SVM).

KNN is a simple classification method that measures the distance of a sample to all known labelled samples. After measurement, the sample is assigned a class by taking the k-nearest neighbour samples, checking their classes, and determining which class has the largest number of neighbours the sample.

SVM works by defining what are known as support vectors that controls the margin that separate the 2 classes with a threshold for misclassifications allowed. The useful thing about SVM is that it can be used to find class boundaries by projecting data points into higher dimensions. Using what is known as the kernel trick, SVM can project the sample values into a higher dimension and work out if that new dimension separates the data better than prior dimensions. If it does, that dimension is used as the classifier dimension.

1.2.3 UNSUPERVISED MACHINE LEARNING

Unsupervised learning is typically used when there is a lack of labelled data or for exploratory data analysis. The lack of labels means that techniques like SVM will not work, since the labels are needed to work out the decision boundary between the classes. One of the most popular methods of unsupervised learning is clustering. Clustering is a technique where data points are grouped together, typically using a distance metric and an estimated number of possible clusters to identify. There are many varieties of clustering that should be used based on expectations of what clusters in the data might look like. This can mean the shape of clusters i.e. circular vs elliptical or features like cluster density.

K-means[1] clustering initialises *K* cluster centres at random positions and assigns datapoints to their closest cluster centre. After assignment the cluster centre is moved to the average of the assigned datapoints. This process is repeated multiple times until the cluster centres stop moving i.e., they converge to either a global or local minima. The initialisation of the *K* cluster centres can skew the algorithm to local minima, so K-Means is often repeated multiple times to determine an optimal clustering. Because of K-means use of a mean as the cluster centroid, clusters are expected to be spherical in nature, so non-spherical clusters will be poorly identified by K-means.

Agglomerative clustering[1] is a method we previously mentioned when discussing phylogenetic methods. This clustering works in a hierarchical manner, with datapoints iteratively clustered into bigger and bigger clusters, before a cut-off is assigned to determine how many clusters is required. This type of clustering is particularly useful for datapoints where there is a shared common relationship that is hierarchical, which is why it is useful for phylogenetic methods.

1.2.4 Self-supervised Machine Learning

Self-supervised machine learning is an increasingly popular approach to model training. Unlike supervised approaches, self-supervised learning uses the data itself as the labels rather than requiring prior labelling of samples. Current large language models are a good example, where the learning objective is predicting elements of the training text such as the next sentence, next token or a masked token. These tasks are derived directly from the data, and as such require no prior label assignment. We will discuss these tasks further when discussing language models and deep learning.

1.2.5 DEEP LEARNING

Deep learning (DL) is a subset of machine learning that expands upon earlier research into artificial neurons. These ideas could be traced back to the research in psychology and neuroscience by McCullough and Pitts[71] in the 1940s that described the action of neurons within the brain using logical operations. The artificial neuron could take a series of inputs which were weighted and would return an output that was dependant on whether the inputs had managed to activate the neuron or not. This is analogous to the action of real neurons, which do not fire unless the activation threshold has been met. The term "deep" learning arises from how in later implementations neurons were stacked into many multiple layers resulting in increasingly "deep" interconnected networks of neurons[1]. The power of deep learning arises from these deep interconnected layers since the activation of each neuron is dependent on the neuron activations from the prior layers and the initial inputs[1]. Each input produces different combinations of neuron activations which ultimately change the values produced at the output layer. The different activations of neurons within these intermediate or "hidden" layers are related to differing features between the inputs[1]. These features typically get more complex the more layers that are added. For example, a network describing protein sequence data may have early layers describing the amino acid properties, while later layers might inform the network about the amino acid positioning in 3-dimensional space. Networks can be trained to learn these properties by changing neuron weights to minimise the error for the prediction task of interest. On a forward pass through of training inputs, the model produces an output that predicts the desired output. Depending on how close to the correct output the model gets (i.e. how minimal is the model error), the

model performs backpropagation which adjusts the model weights to improve future model pass throughs[1]. Weights are often randomly initialised at the beginning of model training, and due to having potentially billions of parameters it often takes many millions of training examples for the model to learn appropriate weights to accomplish the training task. This is one of the reasons deep learning has grown in popularity in recent times since there is now access to datasets of the required magnitude and hardware advanced enough to allow these networks to be trained in reasonable timescales.

1.2.5.1 Deep learning for sequential data

Natural language processing (**NLP**) is a field that has recently taken advantage of deep learning methods. NLP focuses on how computers can be used to process and interpret language. This ranges from problems such as teaching computers to understanding the rules (i.e. the grammar) and structure of a language, to subsequently understanding the meaning of units of language such as words and sentences (i.e. the semantics).

Grammar is defined in the Oxford English Dictionary[1] as:

"The area of study concerned with the structure of a language or of languages in general; esp. the study of the structure of sentences and words, that is, syntax and morphology (sometimes specifically inflectional morphology)."

Syntax governs the structure of words in a sentence, while morphology looks at the structure or the words. Grammar contains both syntax and morphology, both of which contribute to semantics. This relationship between grammar and semantics was explored through the distributional hypothesis[69], an idea from linguistics that can be implemented using deep learning methodologies. It posits the idea that words with similar meanings are often found in similar

contexts and was popularised by John Firth with the statement “you shall know a word by the company it keeps”. Since then, there has been extensive research into understanding the distribution and context of words in natural language and how computers can be used to help with this. First, text must be converted into a numerical representation to be compared with other pieces of text.

1.2.5.2 Word Embeddings

The one hot encoding/vector is one such method of numeric representation. Text is represented as a vector of 1’s and 0’s depending on the presence or absence of words. If a word is present, then the value of the word in the vector becomes 1, else it remains as 0. This implicitly encodes context since vectors differ by the words in the encoded text. This can be useful since word presence can provide a lot of information about the meaning or relevance of a piece of text. For example, the inclusion of the word “Saturn” in a piece of text makes it likely that the text is discussing astronomy in some way. However, in many instances, word presence alone can prove misleading. A one hot encoding does not contain any information about the order of the words in the text, nor how often they occur. Order (i.e. syntax) is important since the meaning of text can be changed by its order. The sentences “The green house was near the pink tree” is identical to “The pink house was near the green tree” as a one hot encoded vector, but clearly the objects have different colours depending on the order of the words. This is even more critical for words like “it” where their placement in a sentence changes the meaning entirely. An example could be the passage “I went to the park with my dog and saw a cat. It was chasing a bird”. If the words “dog” and “cat” were swapped, the meaning of it would

change to refer to the dog rather than the cat, despite "it" remaining the same word in the same position.

Embeddings are trained representations that aim to encode information about an object in a numerical fashion. A one hot encoding can be seen as a very basic untrained embedding, however trained representations can often encode more interesting properties than simply the content of the text. Early word embeddings such as word2vec[72] could learn the semantic properties of words impressively allowing for basic word arithmetic such as the now famous $King - Man + Woman = Queen$ example. This emergent property of the word2vec representation captures why trained embedding representations have become popular: they learn properties of the input that extend beyond the literal characters that the word is comprised of. Word2vec was comprised of 2 distinct architectures, one which used the surrounding context of a word to predict the current word (known as the continuous bag of words or CBOW) and another which used a word to predict the other surrounding words (known as the skip-gram). Due to the different training objectives, the CBOW embeddings performed better on tasks relating to grammar, while the skip-gram embeddings performed better on semantics. Variations and improvements on these architectures have developed more recently, most notably the transformer architecture.

1.2.5.3 The Transformer

The creation of the transformer architecture by Vashwani et al.[73] has ushered in somewhat of an "AI revolution" with many different fields adopting the architecture and achieving impressive results. Much like the aforementioned CBOW architecture, the transformer aims to capture the contextual

information of words in order to learn semantic and grammatical properties about natural language. It is comprised of two main sections called the encoder and the decoder. While both are useful and were initially designed to be used together, many current models use one or the other or both depending on their use case. The full transformer was designed as a sequence-to-sequence (seq2seq) model for language translation, with the encoder section used to learn the context of the initial language sequence, and the decoder used to predict the words from the other language using the encoder context[1].

First, the input sequences are embedded using any fixed length embedding approach (such as a one-hot encoding). Next, this initial embedding is combined with a positional embedding to make a new embedding that is aware of the sequence context and its content. This is a pre-requisite to using either the encoder or decoder blocks of the transformer. Both the encoder and the decoder next utilise what is known as self-attention (Figure 1.7 C), a key mechanism that was introduced as part of the transformer paper.

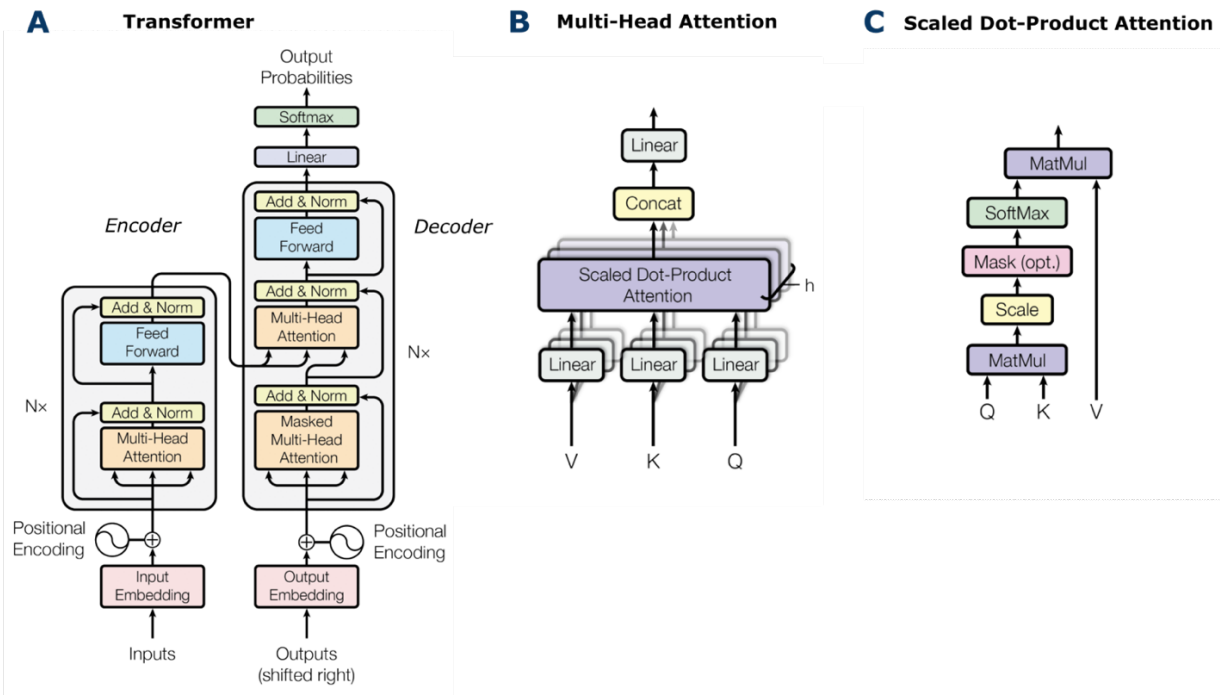


Figure 1.7: From Vashwani et al.[73] (A) The full transformer model architecture. The model is comprised of an encoder (the left block) and a decoder (the right block). (B) The multi-head attention block, which is present in the encoder and the decoder. (C) The attention block, the key element of the transformer mechanism.

A single self-attention block contains 3 matrices called the Query (Q), Key (K) and Value (V) which all learnable parameters of the model. The Q and K matrices are multiplied together to identify how similar the input is to itself, which is why it is called a self-attention block. The QK matrix is then passed through a softmax function which bounds the values between 0 and 1, and effectively make the QK matrix a filter. When this is multiplied by V, the final self-attention matrix is produced which tells the model which elements in the sequence should be attended to.

With the addition of a scaling parameter equivalent to the number of dimensions of the K matrix, the final self-attention formula equates to:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})$$

This mechanism is how the model can use the context of other words in the sequence to understand the meaning of words such as "it". Each of these attention heads can learn different properties, which is why a typical model includes several heads per encoder/decoder.

The main difference between the encoder and the decoder is that the decoder is generative and "auto-regressive" meaning that it generates an output based upon the previous values. The decoder generates a new token based upon the context provided by the encoder, and the prior tokens produced from the decoder. In the case of ChatGPT, which is an example of a decoder only model called Generative Pretrained Transformer (GPT)[74], the prompt given by the user is used to predict the likely next tokens which happen to be the response. Each new word of the decoder output directs the model on what the next token should be until the <EOS> token is produced which stands for end of sequence. This is why the decoder uses a masked attention head (Figure 1.7A) during training time, the model is only allowed to use attention on the tokens prior to the current token in the sentence. A full self-attention head would allow attention being calculated for all words in the sentence, including those after the current word, allowing the model to cheat in predicting its next token, hence masking is needed to hide the future words. The masked attention is then combined with the encoders own attention output using a cross attention mechanism. As the models we will discuss are predominantly encoders or

decoder only variants, we won't discuss this in detail and instead will focus on the details as well as the pros and cons for decoders and encoders.

1.2.5.4 To encode or to decode, that is the question.

Much like the CBOW and the skip-gram, encoders and decoders learn different properties because of their different architectures and training objectives. Encoders typically learn the properties of sequence via the Masked Language Model (MLM) objective, or the Next Sentence Prediction (NSP) objective. The popular BERT (Bi-directional Encoder Representations from Transformers) model[75] used both of these objectives as part of its training. MLM essentially randomly masks a proportion of the input sentence with <MASK> tokens and asks the model to predict the masked tokens using the rest of the sequence. In doing so, the encoder learns about token distributions in language and how sentences are constructed. The NSP objective uses a classification (<CLS>) token that is placed at the start of every sequence input during training and uses this to predict if a selected sequence is indeed the true next sequence. This token is not a traditional character, but its inclusion in the sequence allows it to learn about sequence context and thus its value can be used to classify sequences. In the BERT paper, 50% of the time the sequence is the next sequence, while 50% of the time it is a randomly selected sequence. Decoder models learn using the next token prediction task. As such, encoder models like BERT are typically used when representations of whole sentences are required and can be used for prediction. Decoder models are often used for their generative properties and have become increasingly used as chatbots since they can use questions as prompts to autoregressively generate text as an answer.

1.2.6 BIOLOGICAL SEQUENCE EMBEDDINGS

Biological sequences share several similarities with natural language. Both nucleotide and protein sequences use the alphabet to represent their constituent biological tokens i.e. nucleotides and amino acids. They also both carry meaning that extends past their literal token composition. As such, learning meaningful representations of both nucleotide and protein sequences may provide insight into similarities between sequence that go past basic composition. The evolutionary scale model (ESM) by Rives et al.[76] was one of the first large protein language models based on a BERT style encoder architecture. It was trained using the UniRef database of 250 million sequences across all of life and aimed to learn generalisable properties about protein structure, function and composition. It was quickly apparent that the model had learned interesting properties since low dimensional representations (t-distributed Stochastic Neighbour Embedding or t-SNE) of individual amino acid embeddings appear to group by attributes such as charge and molecular weight. Earlier versions of the model managed to perform successful variant effect prediction of sequences, predict secondary structure and even produce contact maps between residues. Later versions improved upon all of these capabilities and with the use of the folding trunk and structure module (modules from the AlphaFold pipeline that take alignment and embedding information and produces atomic structure coordinates) from AlphaFold even managed to produce tertiary protein structures[60,61]. The wealth of information within these protein language model embeddings has meant that they are now being used for all sorts of tasks including identifying evolutionary trajectories[77], predicting protein-protein interactions(PPIs)[78], and even generating new never before observed

proteins[79,80]. Nucleotide transformers are also starting to emerge, although the context window of current transformers is often too small for full sequences, although a number of methods are beginning to tackle these issues.[81]

1.2.7 Dimensionality Reduction and Source Separation

The curse of dimensionality refers to the inherent difficulties associated with the visualisation and computation of data containing more than 3 dimensions. Several properties about high dimensional data make it difficult to use. These include visualisations only being capable of showing a low number (typically only two) of dimensions at a time, distances measures becoming less meaningful as dimensions increase and the increasing sparsity of data points as within high dimensional representations[82]. In this era of rapidly expanding biological feature sets, issues with interpreting high dimensional data will become a necessary obstacle for most researchers. Embeddings are an excellent example of this, given they are high dimensional representations, yet even experiments with more than 3 columns of results invoke the curse and its trials.

To tackle these problems, techniques have been developed over the years to create low dimensional representations of high dimensional data. These representations allow for high dimensional feature sets to be analysed or visualised in a classic low dimensional space, with some important caveats. These caveats depend on the assumptions made by the dimensionality reduction technique on such as the linearity, the type of distributions and even the type of values i.e. large, small, positive, negative, or any combination of these and more.

1.2.7.1 Principal Component Analysis (PCA)

PCA[1] is one of the most popular and well-known dimensionality reduction methods. It is a linear method that calculates principal components i.e. vectors that maximise the variation between features. By zero-centering the features, creating a line that goes through the origin, mapping samples onto the line, and then maximising the summed square of distances between the points for each sample and the origin, we make a fitted line that equates to a principal component (PC). A new PC is added per combination of features, with each PC being perpendicular to the previously calculated PCs. The variation (also known as the eigenvalue) is equivalent to the average of the summed square of distances. While PCs are calculated for every feature pair in the data, only a subset (ideally 2 for visualisation) are selected to represent the data. If the variation of the first 2 PCs represent enough of the total variation in the data, then downstream tasks like classification or visualisation of the first 2 PCs can be produced and are likely to be representative of the data. The less each individual PC can capture variance, the less likely the dimensionality reduction is to adequately represent the data. PCA is popular due to its simplicity and interpretability. However, it is likely to struggle in the presence of non-linear relationships, where variation is unlikely to be well described through linear projections.

1.2.7.2 Uniform manifold approximation and projection for dimension reduction (UMAP)

Unlike PCA, the UMAP[83] is a dimensionality reduction technique that should only be used for visualisation. Rather than performing a linear operation on the features, UMAP tries to create an optimal grouping of objects by

prioritising local distances over global distances. This makes it very useful for grouping together similar samples in a dataset, but importantly means that the UMAP dimensions have no explicit meaning unlike with PCA where dimensions are linearly derived from the features. As such, UMAPs offer a nice visualisation of the approximate arrangement of a dataset, but generating hypotheses by looking at UMAP distances should be avoided. For exploratory data analysis, it can be an excellent tool for understanding local structures within a high dimensional dataset.

1.2.7.3 Non-Negative Matrix Factorisation (NMF)

Non-negative matrix factorisation[1] creates low dimensional representations by performing matrix factorisation, a method which breaks a matrix of values into 2 sub-matrices (Figure 1.8). One matrix contains the basis vectors, similar to the principal components of PCA. The other matrix contains the weights for each of the basis vectors, which indicate how much each vector contributes to the original data matrix. NMF can do this because there is a non-negativity constraint for the initial dataset. This requirement means that each basis vector can be thought of as an additive component of the data i.e. the original dataset can be rebuilt by adding together the basis vectors together at different quantities.

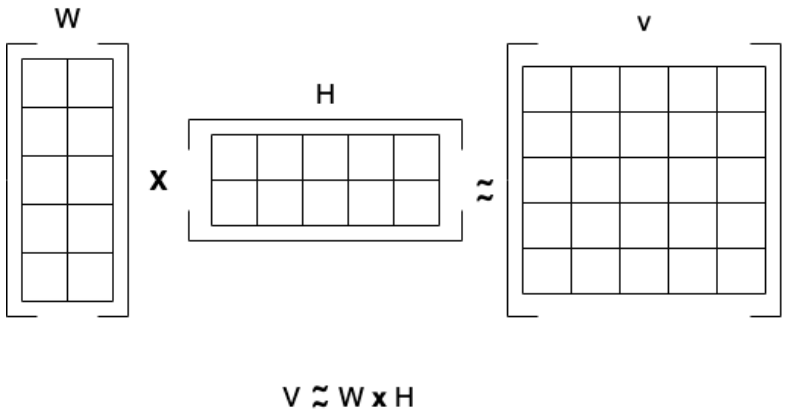


Figure 1.8: Equation and visual representation of NMF. The goal of NMF is to factorise the matrix V into the W and H matrices. W represents the weights matrix while H represents the feature matrix. The multiplication of W and H creates an approximation of matrix V.

For datasets where there is a non-negative constraint such as signal processing or even cancer evolution, the basis vectors of NMF can be related to meaningful signal sources. Lee and Sung[84] demonstrate that NMF applied to images of human faces produces feature vectors that describe components of the faces such as noses, eyebrows and mouths, with weight matrices describing how much of these are represented in the image i.e. how prominent are the eyebrows. This additive property of NMF features make them very interpretable since they directly and intuitively contribute to the data.

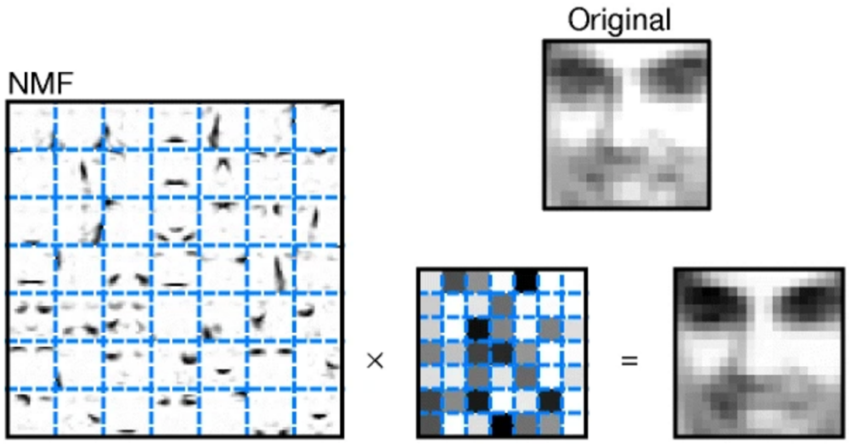


Figure 1.9: Figure from Lee and Sung[84] showing NMF decomposition on a human faces database. The feature matrix shows parts of faces such as noses or mouths, while the weight matrix shows the how much of these attributes are represented in the real face when reconstructed.

Of particular interest is how NMF can be applied to genomic data. This topic has an extensive history within the field of cancer genomics, with NMF being used to extract what are known as mutational signatures across cancer types[85–87]. By modelling evolution as a set of purely additive process (i.e. mutations can only be added), NMF can be used to extract the signals that are left in the sequencing data by the processes adding mutations to the genome. Mutational processes such as the APOBEC family of cytidine deaminases are known to have specific nucleotide contexts and substitution types. By starting with a count matrix of substitution-context pairs for each mutation in a cancer sample, NMF extracts signature vectors with mutations that occur at similar frequencies and at similar contexts. These can then be compared with the known mutational contexts for known biological mutagens and subsequently validated. Furthermore, the weight matrix from NMF allows can then be interpreted as the amount of exposure this sample has had to the mutational

processes that were extracted from the samples. NMF therefore allows for identification of active mutational processes and an estimation of their importance within the cancer type. This has allowed for validation of key mutagenic processes behind skin and lung cancers (UV light exposure and smoking) as well as the identification of processes present across the spectrum of cancers[85].

1.3 THESIS OUTLINE

1.3.1 MUTATIONAL SIGNATURE DYNAMICS INDICATE SARS-COV-2'S EVOLUTIONARY CAPACITY IS DRIVEN BY HOST ANTIVIRAL MOLECULES

Kieran D. Lamb (1,2), Martha M. Luka (1,2), Megan Saathoff (1), Richard J. Orton (1), My V.T. Phan (3,4), Matthew Cotton (1,3,4,5), Ke Yuan (2,6,7), David L. Robertson (1)

1. Medical Research Council - University of Glasgow Centre for Virus Research, University of Glasgow;
2. School of Computing Science, University of Glasgow, Glasgow, Scotland, United Kingdom;
3. Medical Research Council/Uganda Virus Research Institute and London School of Hygiene & Tropical Medicine Uganda Research Unit, Entebbe, Uganda;
4. College of Health Solutions, Arizona State University, Pheonix, Arizona, United States of America;
5. Complex Adaptive Systems Initiative, Arizona State University, Scottsdale, Arizona, United States of America;
6. School of Cancer Sciences, University of Glasgow, Glasgow, Scotland, United Kingdom;
7. Cancer Research UK Scotland Institute, Glasgow, Scotland, United Kingdom;

This chapter is taken from the paper "Mutational signature dynamics indicate SARS-CoV-2's evolutionary capacity is driven by host antiviral molecules" published in PLOS Computational Biology. This chapter uses regression models and mutational signature analysis to investigate SARS-CoV-2 waves of infection, uncover the putative mutational processes behind the virus's evolution, and track the activity dynamics of those processes through time. The paper was written by me with assistance from Martha Luka and Megan Saathoff. Martha assisted with the introduction and sections involving regression models and factors behind the SARS-CoV-2 waves (Sections 2.5.1 and 2.5.2). Sections involving mutational signatures were completed by me (Sections 2.5.3 – 2.5.6). The section involving analysis of mutational signatures in the wastewater and immunocompromised datasets (Section 2.5.7) was completed by Megan. The relevant methodology and discussion sections relating to these results were completed by their results author. Richard Orton assembled the SARS-CoV-2 sequence alignments that were used to complete the analysis. Matthew Cotton and My Phan were involved in the reviewing and editing of the manuscript. Ke Yuan and David L. Robertson supervised, edited and helped with the writing of the paper.

1.3.2 LARGE LANGUAGE MODELS CHARACTERISE THE PROTEINS OF SARS-COV-2

Kieran D. Lamb (1,2), Joseph Hughes (1), Spyros Lytras (1,3), Francesca Young (1), Orges Koci (1,4), Jamie Herzig (5), Simon C Lovell (5), Joe Grove (1), Ke Yuan (2,6,7), David L. Robertson (1)

1. MRC-University of Glasgow Centre for Virus Research, School of Infection and Immunity, Glasgow, UK;
2. School of Computing Science, University of Glasgow, Glasgow, UK;
3. Institute of Medical Science, University of Tokyo, Tokyo, Japan;
4. European Molecular Biology Laboratory- European Bioinformatics Institute, Hinxton, UK;
5. School of Biological Sciences, University of Manchester, Manchester, UK;
6. School of Cancer Sciences, University of Glasgow, Glasgow, UK;
7. Cancer Research UK Scotland Institute, Glasgow, UK;

This chapter comes from the pre-print "From a single sequence to evolutionary trajectories: protein language models capture the evolutionary potential of SARS-CoV-2 protein sequences". It shows how protein language models can be used for many different tasks that can help inform about fundamental properties of viral proteins. We show how these representations are useful tools for understanding viral evolution, from performing in-silico deep mutational scanning (DMS) to tracking new SARS-CoV-2 variants and assessing their potential to be a new VOC. This chapter was written by me and most of the analysis was completed by me. Joseph Hughes, Orges Koci and

Spyros Lytras assisted with editing, and gathering and processing much of the sequence datasets used in the analysis. Joe Grove and Francesca Young provided feedback, editing and comments on the chapter. James Herzig and Simon Lovell provided the protein stability changes data. Ke Yuan and David L. Robertson supervised and helped with editing.

1.3.3 Investigating the co-occurrence of mutations in SARS-CoV-2

Kieran D. Lamb (1,2), Stephanie Brown (1), Ke Yuan (2,3,4), David L. Robertson (1)

1. MRC-University of Glasgow Centre for Virus Research, School of Infection and Immunity, Glasgow, UK;
2. School of Computing Science, University of Glasgow, Glasgow, UK;
3. School of Cancer Sciences, University of Glasgow, Glasgow, UK;
4. Cancer Research UK Scotland Institute, Glasgow, UK.

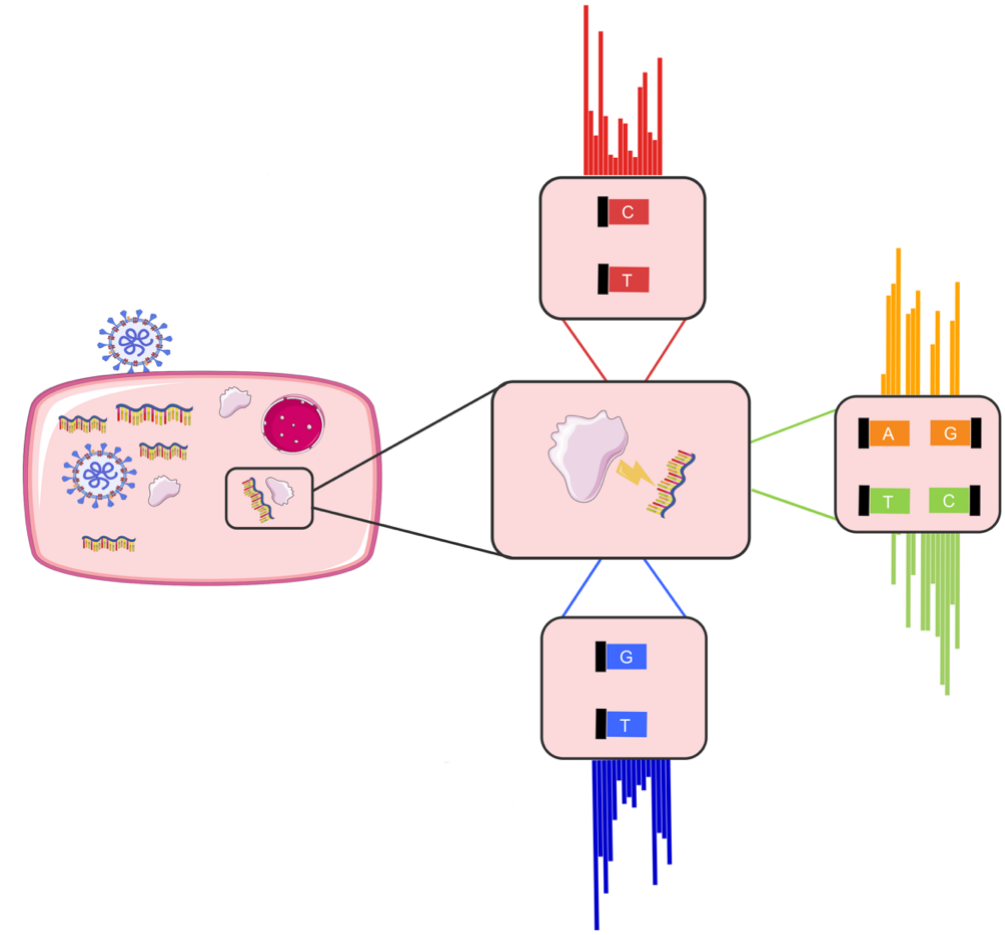
This final chapter looks into the co-occurring mutations that appear throughout SARS-CoV-2's phylogeny. We look into where these co-occurrences appear across the genome, investigate their mutational contexts, and use language modelling to investigate how these mutations may interact within the protein. This chapter was written by me, and the analysis and code were written by me and by Stephanie Brown. Ke Yuan and David L. Robertson supervised and helped with editing.

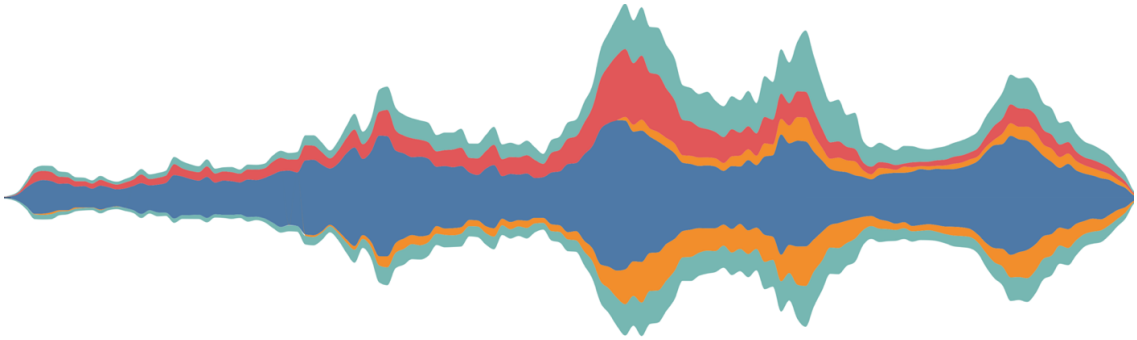
1.3.4 Data Availability

The data, code and observable notebooks for visualisation can be found at:

https://github.com/kieran12lamb/Thesis, version number f000e60.

2 MUTATIONAL SIGNATURE DYNAMICS INDICATE SARS-COV-2'S EVOLUTIONARY CAPACITY IS DRIVEN BY HOST ANTIVIRAL MOLECULES





"Ch-ch-ch-ch-changes turn and face the strange ch-ch-changes"

David Bowie, "Changes" (1971)

2.1 Abstract

The COVID-19 pandemic has been characterised by sequential variant-specific waves shaped by viral, individual human and population factors. SARS-CoV-2 variants are defined by their unique combinations of mutations and there has been a clear adaptation to more efficient human infection since the emergence of this new human coronavirus in late 2019. Here, we use machine learning models to identify shared signatures, i.e., common underlying mutational processes and link these to the subset of mutations that define the variants of concern (VOCs). First, we examined the global SARS-CoV-2 genomes and associated metadata to determine how viral properties and public health measures have influenced the magnitude of waves, as measured by the number of infection cases, in different geographic locations using regression models. This analysis showed that, as expected, both public health measures and virus properties were associated with the waves of regional SARS-CoV-2 reported infection numbers and this impact varies geographically. We attribute this to intrinsic differences such as vaccine coverage, testing and sequencing capacity and the effectiveness of government stringency. To assess underlying evolutionary change, we used non-negative matrix factorisation and observed three distinct mutational signatures, unique in their substitution patterns and exposures from the SARS-CoV-2 genomes. Signatures 1, 2 and 3 were biased to C→T, T→C/A→G and G→T point mutations. We hypothesise assignments of these mutational signatures to the host antiviral molecules APOBEC, ADAR and ROS respectively. We observe a shift amidst the pandemic in relative mutational signature activity from predominantly Signature 1 changes to an increasingly high proportion of changes consistent with Signature 2. This could represent changes in how the virus and the host

immune response interact and indicates how SARS-CoV-2 may continue to generate variation in the future. Linkage of the detected mutational signatures to the VOC-defining amino acids substitutions indicates the majority of SARS-CoV-2's evolutionary capacity is likely to be associated with the action of host antiviral molecules rather than virus replication errors.

2.2 SUMMARY

We show that both public health measures and virus properties are associated with the rise and fall of regional SARS-CoV-2 reported infection numbers with regional differences attributable to the extent of vaccine usage and the effectiveness of public health measures. In our mutational signature analysis, using non-negative matrix factorisation, we detected three distinct mutational signatures that can be putatively attributed to the action of specific host antiviral molecules. Interestingly, we observe a shift in mutational signature activity from predominantly Signature 1 changes to an increasingly high proportion of changes consistent with Signature 2. These mutation patterns influence SARS-CoV-2's evolutionary capacity, the available genetic variation that selection can act on, and so can be linked to the mutations defining the variants of concern responsible for the distinct SARS-CoV-2 infection waves. The dominant types of nucleotide substitutions involved indicate that much of the mutation and hence variation come from the action of the host immune response rather than replication errors since the virus has an error correction system.

2.3 INTRODUCTION

The extensive and rapid global spread of SARS-CoV-2 and its detrimental impact on human health has placed it as the causative agent of one of the most significant pandemics in recent history[88]. Different geographical regions of the world have reported varied infection patterns that are attributed to differences in population demographics and health care systems, diverse government responses[89,90], the emergence of more transmissible variants[91,92] and other viral, human and population factors. Since its emergence, SARS-CoV-2 has undergone significant genetic change such that numerous variants, i.e., distinct genotypes, have been identified[93], many with altered phenotypic properties[58].

The World Health Organization (WHO) and other public health bodies have broadly classified variants that pose an increased risk to global public health as variants of concern (VOCs) and variants of interest (VOIs) [94]. The early SARS-CoV-2 variants to emerge in 2019 and the more transmissible +S:D614G variant followed by the VOCs (Alpha, Beta, Gamma, Delta and currently Omicron) have driven significant and sequential "waves" of SARS-CoV-2 infections internationally. The emergence of each variant showing a clear geographical link[95–97].

Viral mutations arise from a diverse set of processes (principally viral polymerase replication errors and host anti-viral editing processes), which can be identified by the characteristic mutational signatures that they leave on the genome[87,98]. Such characterisation of dominant mutational processes is routinely used in cancer genomics[99]. The catalogue of SARS-CoV-2 nucleotide changes show distinct mutational patterns suggestive of a role for host antiviral mutational processes in introducing changes in the viral RNA[100,101].

These processes potentially dominate in SARS-CoV-2 evolution because point mutations introduced in replication are mostly corrected by the action of a proofreading enzyme.

The generation of virus diversity, the key to virus persistence by generating novel variation and thus evolutionary capacity, is multi-faceted[102], yet our understanding of the relative importance of underlying mutational processes linked to the action of host anti-viral molecules is still very limited. Given that SARS-CoV-2 continues to develop new variants, many associated with sets of previously observed (convergent) and novel mutations[58], it is critical that we improve our understanding of the mechanisms and sources of evolutionary change.

Along with routine surveillance of SARS-CoV-2 infections, there has been an unprecedented global sequencing effort resulting in databases containing many millions of genome sequences, in particular GISAID[25]. Here we examined this data to describe the global molecular epidemiology and evolution of SARS-CoV-2. Using regression models we first examined how viral properties and public health measures have influenced the magnitude of infection waves in different geographic locations. Satisfied that SARS-CoV-2 variants have been an important driver of infections we then used non-negative matrix factorisation to characterise the mutational processes involved in the generation of variants and their changing patterns of activity over time.

2.4 METHODS

2.4.1 DATA

The findings of this study are based on metadata associated with 13,662,759 sequences available on GISAID up to 01 December 2022 and accessible at

https://doi.org/10.55876/gis8.221201qs. Sequences were filtered to remove records from non-human hosts, with lengths less than 20,000 nucleotides, non-assigned lineages, with greater than 30% unknown bases, sequences reported to be collected before 24/12/2019 and those with excessive mutations/deletions. The cutoff for filtering out hypermutated sequences was 175 mutations in coding regions or more than 69 different deletions, the cutoffs were manually determined after evaluation of the mutation/deletion distribution and selecting the point where sequence counts were consistently observed in single digits, this resulted in 1,852 sequences being filtered out.

Publicly available daily SARS-CoV-2 cases, tests performed and total vaccinations per capita were obtained from Our World In Data (OWID) [103] in September 2022. Prior to February 2023, the OWID data was piped from the Johns Hopkins University COVID-19 dashboard [104,105]. Country-level government stringency indices were downloaded from the Oxford COVID-19 government response tracker (OxCGRT)[106]. Government stringency indices are composed of nine indicators: school closure, workplace closure, cancellation of public events, stay at home order, public information campaigns, restrictions on public gatherings, public transport, internal movement and international travel. The index on a given day ranges from 0 to 100 and is calculated as the mean of the nine indicators, with higher indices indicating stricter regulations. If responses vary at sub-national levels, the index at the strictest level is used[106]. Wastewater findings are based on metadata associated with 1,343 sequences available on GISAID and accessible at https://doi.org/10.55876/gis8.230406qg. Wastewater sequences were downloaded from the 'wastewater data' section of GISAID in December 2022.

Sequences for immunocompromised individuals were downloaded from GISAID in November 2022. Analysis of this was based on the metadata associated with 34 sequences available on GISAID and accessible at https://doi.org/10.55876/gis8.230406fb. Sequences were chosen based on the known list of sequences used in Harari et al. [107]. Sequences were aligned to the COVID reference genome before use.

2.4.2 DESIGN

Predictors of SARS-CoV-2 reported cases were explored using a linear model at both country and continent levels. We collected continuous dependent variables reported on a daily basis. These were classified into two groups: (i) public health measures (government stringency, testing capacity and vaccination), (ii) viral properties (diversity and fitness). We examined the data for completeness of predictive variables. In instances of missing vaccination data, we interpreted this as no vaccinations having been given. This was a reasonable assumption for periods prior to the vaccine rollouts in the respective countries. With the exception of vaccinations, variables with less than 70% of the countries reporting data were not included. The number of SARS-CoV-2 diagnostic tests performed was excluded as a predictor due to missing data. We determined the previous burden by summing the adjusted new cases per capita over the past 90 days. Prior infection significantly reduces the risk of a subsequent infection, with a reduction in risk of up to 95% in the initial three months[108]. This was included as a predictor variable in the linear model. Amino acid substitutions were defined against the Wuhan-Hu-1 (GenBank: MN908947.3) sequence. Building on findings from Obermeyer et al.[109], we extracted a list of previously identified fitness-associated

mutations[109]. Mutations were categorised as fit using the PyR0[1] model which uses a Bayesian logistic regression model to determine mutations associated with lineages of the virus that have an increasing predicted relative prevalence[1]. The fitness associated mutations were as the top 20 identified mutations, and are as follows: S:H655Y, S:T95I, ORF1a:P3395H, S:N764K, ORF1a:K856R, S:S371L, E:T9I, S:Q954H, ORF9b:P10S, S:L981F, N:P13L, S:G339D, S:S375F,S:S477N, S:N679L, S:S373P, M:Q19E, S:D796Y, S:N969K, S:T547K,ORF1b:I1566V, M:D3G, S:G446S, S:N440K, M:A63T, S:N856K, ORF1a:A2710T,ORF1a:I3758V, S:E484A, S:A67V, S:K417N, S:Q493R, S:N501Y,S:Y505H,S:L452R, S:P681H, S:Q498R,S:G496S, ORF1a:T3255I, ORF14:G50W, S:P681R,N:R203M, ORF1b:P1000L, ORF1a:P2287S, M:I82T, ORF3a:S26L, N:D63G, N:G215C, ORF1a:V3718A, ORF9b:T60A. Each fit mutation within a sequence was counted and the counts were normalized to the number of sequences per geographical location. Virus fitness was therefore defined as the sum of the frequencies of previously identified[109] amino acid substitutions that increase SARS-CoV-2 fitness divided by the sum of total genomes and the log of total mutations per location. The denominator (i.e. the total sequences per week and the log of the total mutations per week) was used to normalise the weekly fit mutations across sequencing levels, so that viruses found in countries with higher levels of sequencing were not preferentially more fit.

$$Virus\ Fitness = \frac{weekly\_sum\_of\_fit\_mutations}{total\_seqs\_per\_week + \log\left(total\_mutations\_per\_week\right)}$$

Diversity was calculated by dividing distinct lineages by the total number of genomes in a given week. Sequences reported in GISAID were assumed to be representative of the diversity of infections for that continent/country.

2.4.3 LINEAR MODEL

We employed a linear regression model, described by Heo et al.[110], to adjust reported cases per country using the Human Development Index (HDI), which encompasses not just economic growth but also reflects a country's capacity for per capita testing. Countries with higher HDI levels, typically high-income nations, conducted more tests per million people, often leading to more confirmed cases compared to nations with lower HDI levels. Adjusted daily cases were smoothed using a 14-days rolling average to limit possible noise and identify simplified changes over time. For continent-level analysis, data from all contributing countries was used to fit the linear model. To ensure that countries with a large number of cases didn't artificially inflate the results, each country's influence on the continent-level OxCGRT index was adjusted based on its percent contribution to the continent's 14-day average daily case tally.

Pearson's correlation was used to test for correlation among the variables. Multiple linear regression was fitted to evaluate the relationship between infection rate (adjusted daily cases per capita) as the outcome and the public health measures and viral properties as predictors within the different continents. The regression models were fitted on data from 01 April 2020 onwards, as (sequence) data addition remained stable after this. The country-level analysis was carried out for countries with less than 50 days of missing genome data using a similar approach.

1.1.1 PANDEMIC PLOTS

Case numbers and sequence data were aggregated by their respective continents, a 14-day rolling average was used to smooth out daily infection rates and categorical variables were summarised by counts. Proportions of

lineages were calculated in 14-days bins and the most common lineages were visualised per continent.

2.4.4 TREE-BASED REFERENCING

The rapid evolution of SARS-CoV-2 means that the majority of viral sequences are distinct from the early pandemic reference genome Wuhan-Hu-1[28]. Continuing to count mutations against the early reference sequence can result in mutations being allocated the wrong substitution category (i.e., A→T instead of a C→T) where sites have mutated multiple times. Azgari et al.[111] tackled this issue by building a tree of clustered sequences to remove ancestral mutations. However, we utilise the available SARS-CoV-2 tree generated as part of the Pango[93] nomenclature to generate a reference sequence for each defined lineage. This means that sequences from the lineage B.1 are compared against a generated reference sequence for the B lineage rather than the Wuhan-Hu-1 sequence (See Figure 2.1 for diagrammatic description).

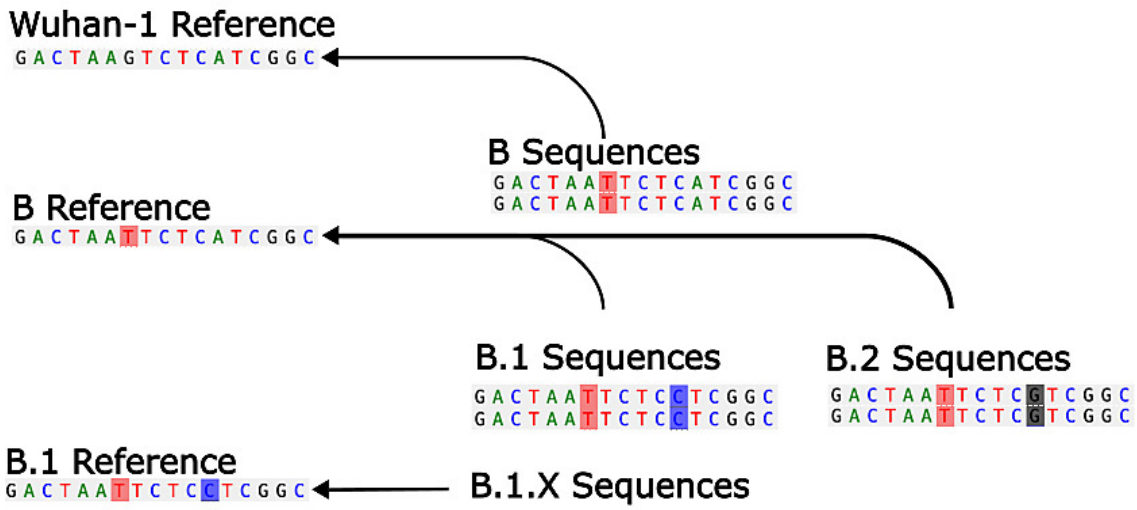


Figure 2.1: Diagrammatic depiction of how tree-based referencing works. Each Pango lineage has a reference generated for it. Arrows show which sequences use which reference sequence, with the arrow tip indicating the

reference. For example, sequences from the B.1 lineage are compared against the reference for the B lineage so that B.1 lineage-defining mutations can be counted.

One reference sequence was generated for each of the Pango lineages in the alignment. A nucleotide was included in the generated Pango reference if it exceeded a frequency threshold of greater than 75% of the samples from the lineage. If this threshold was not reached, the reference nucleotide of the nearest parental lineage was used (i.e., if a mutation in B.1 is ambiguous, the nucleotide from the B lineage reference at that position is used). Building intermediate references also meant that counting inherited mutations could be avoided. Since mutations were identified relative to their nearest parental Pango lineage, inherited mutations are not counted because, relative to this sequence, there hasn't been a mutation. Mutations are also only counted once per lineage set of sequences so that mutations that are observed many times due spread of the virus rather than acquisition by a mutational process are not over-counted. This means that convergent amino acid substitutions can be observed between lineage sets, although they may be undercounted within a lineage. However, this is necessary since it is very difficult to identify convergence within similar sequences (especially at a global scale). Overcounting of the mutations results in mutational signatures that reflect the circulating predominant lineages rather than the mutational processes producing the mutations in those lineages.

1.1.1 PSEUDO-SAMPLING

Mutations were binned into categories composed of their substitution type (e.g., cytosine → thymine = CT) and their mutation context. The mutation

context is the mutated base and the nucleotides at the 5' and 3' positions of the mutated base. There are a total of 192 types of substitution-context matchings that can appear (12 possible single nucleotide changes x four possible nucleotide 5' x four possible nucleotide 3'). Every sequence produces a single count vector of mutation category counts, with the total count matrix becoming the mutational catalogue of the virus. On average, a single SARS-CoV-2 genome sequence has very few new mutations. As extracting mutational signatures when mutation counts are low is unlikely to produce meaningful results, we define each sample as a time-point (all of the sequences collected in an epidemic week) and decompose signatures from the counts at each time-point rather than from each sequence. This means that for each week, the mutations are counted for all the sequences of that week, and summed together to produce one row that contains all of the substitution-context counts. This shrinks the mutational catalogue of the virus from millions of samples down to less than 200 samples, one for each Epidemic Week. The rationale behind this was that NMF on millions of rows is computationally intractable on the available hardware, and performing NMF on individual sequences is problematic as there is not enough signal of mutational patterns from one or two mutations to produce a meaningful result. We also found that weeks produced very similar results to sample days, yet were significantly easier to run, plot, and discuss with regards to pandemic developments in real time.

1.1.1 NON-NEGATIVE MATRIX FACTORISATION

NMF (non-negative matrix factorisation) was used to split the mutational catalogue into two sub-matrices. One matrix represents the mutational signatures, the other matrix represents the exposure of the signatures. These matrices were used to reconstruct the original mutational catalogue with some degree of error. To verify the validity of the identified signatures, NMF was performed 100 times for each value of N, with N representing the number of signatures to extract from the mutational catalogue. The N of 100 was determined from best practices in Islam et al.[153], which is the tool used for de novo mutational extraction by the authors of the Catalogue Of Somatic Mutations In Cancer (COSMIC)[140] which is the gold standard resource for mutational signatures in cancer. Results were also stable at this number, and often even for N's much lower than 100. For this analysis, N was set to 2, ..., 10. For each NMF run, a new mutational catalogue was generated using bootstrap re-sampling of the original matrix and removal of any mutational categories that did not account for more than 0.5% of mutations. Mutational categories are pseudo-sampled down into epidemic week matrices that NMF was run on. The signatures were then clustered together using K-means clustering, with the cluster means forming the new signatures. Clusters were then assessed using the silhouette score[154] to determine the clustering quality. The silhouette score uses the cluster density (how close the points within each cluster are) as well as the distances between clusters to calculate how well a clustering method has performed. Clusters with a score between -1 and 0 are either misassigned (negative values) or overlapping (zero). Clusters above zero indicate increasingly better clustering, with 1 being the best. The silhouette score threshold was set to 0.95 to ensure high quality clusters. Clustering with

Clusters with high silhouette scores are well separated from other clusters and are dense and well-formed. Cosine similarity was used to determine if the signature was reliably extracted from the cluster. The cosine similarity was calculated between signatures extracted from the whole mutational catalogue and the cluster means of the signature clusters. A higher cosine similarity indicates that the cluster mean shows a similar pattern to the initial mutational signature. Again, a threshold of 0.95 was set to ensure high quality, robust signature extraction.

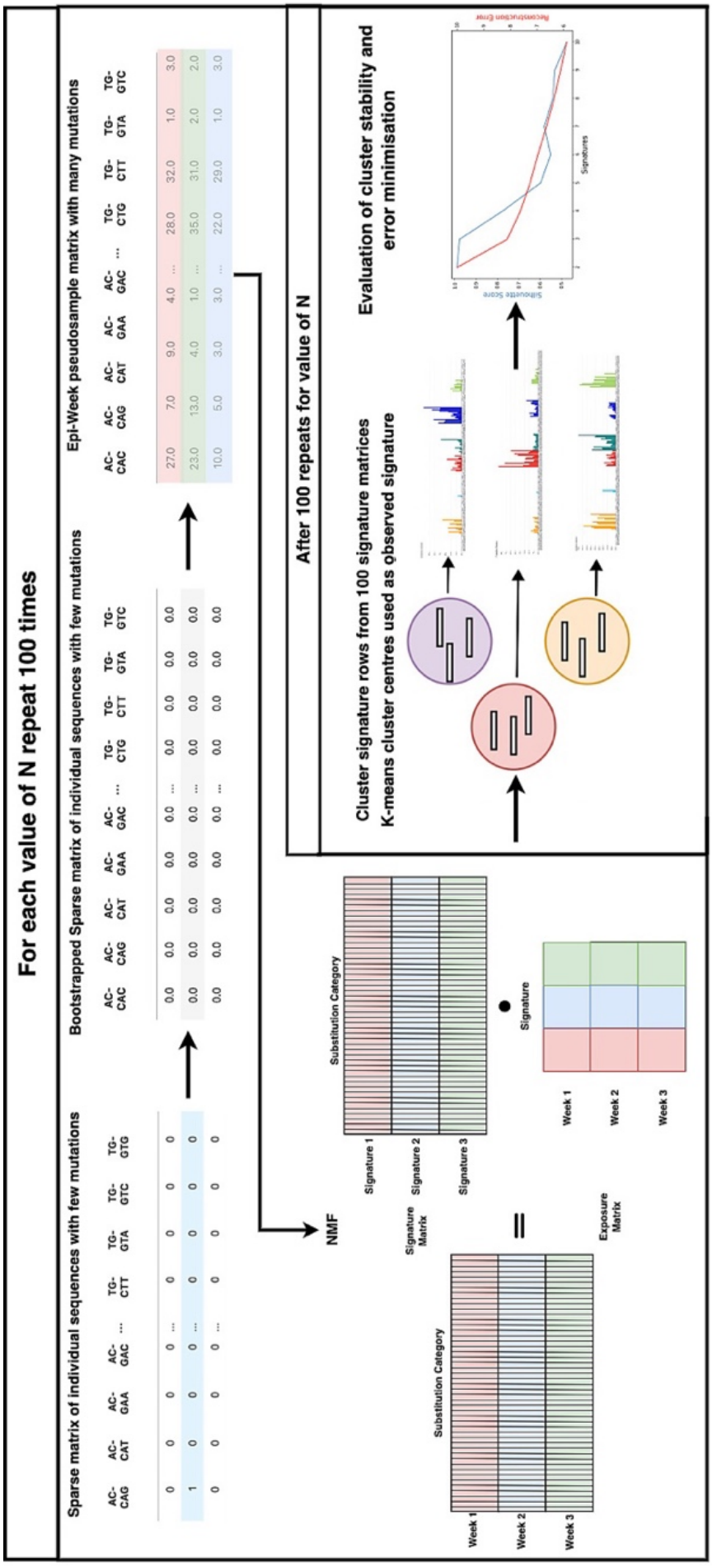

For each value of N repeat 100 times
Sparse matrix of individual sequences with few mutations
Bootstrapped Sparse matrix of individual sequences with few mutations
Epi-Week pseudosample matrix with many mutations
AC-CAG
AC-CAT
AC-GAA
AC-GAC
...
TG-CTG
TG-CTT
TG-GTA
TG-GTC
TG-GTG
AC-CAC
NMF
Substitution Category
Signature 1
Signature 2
Signature 3
Signature Matrix
Signature
Exposure Matrix
Week 1
Week 2
Week 3
After 100 repeats for value of N
Cluster signature rows from 100 signature matrices
K-means cluster centres used as observed signature
Evaluation of cluster stability and error minimisation
Silhouette Score
Reconstruction Error
Signatures

Figure 2.2: Graphical description of the methods for NMF extraction of mutational signatures. For every value of N signatures, the mutational signatures are extracted 100 times for bootstrapped and pseudo-sampled datasets. Once this has been completed, signatures are clustered into N clusters and the stability and density of those clusters are evaluated using the silhouette score. Signatures that have silhouette scores above 0.95 are evaluated as stable signatures. The cluster means become the extracted signatures. The best set of N signatures is selected by picking the value of N that best minimises the reconstruction error and has the best silhouette score (with a minimum of 0.95). A further evaluation is the cosine similarity of the clustered signature means with the signatures extracted by completing NMF on the original pseudo-sampled dataset. Again, signatures must have a cosine similarity of at least 0.95 to be considered.

Again, following the best practices in Islam et al.[153], an N value of three was selected due to the reduction of the reconstruction error plateauing around three and the marked decrease in silhouette score for signatures greater than 3. Reconstruction error is defined as the difference between input matrix to the NMF, and the values of the matrix that can be produced using the NMF components and weights. The average cosine similarity between signatures and clusters was consistently above 0.95 for each cluster and had an average of 0.98 for all three clusters when clustering was repeated 100 times. Silhouette scores for each cluster were above 0.95, suggesting excellent separation and density of clusters (Table 2.1 and Figure 2.3). Signatures can therefore be reliably extracted from the bootstrapped catalogues, are robust and thus are unlikely to be artefacts. Counts of mutations were normalised by the tri-mer

composition of the SARS-CoV-2 reference sequence (dividing the counts by the number of contexts in the reference sequence). Composition biased versions of the signatures were then produced by rescaling the signatures using tri-mer composition.

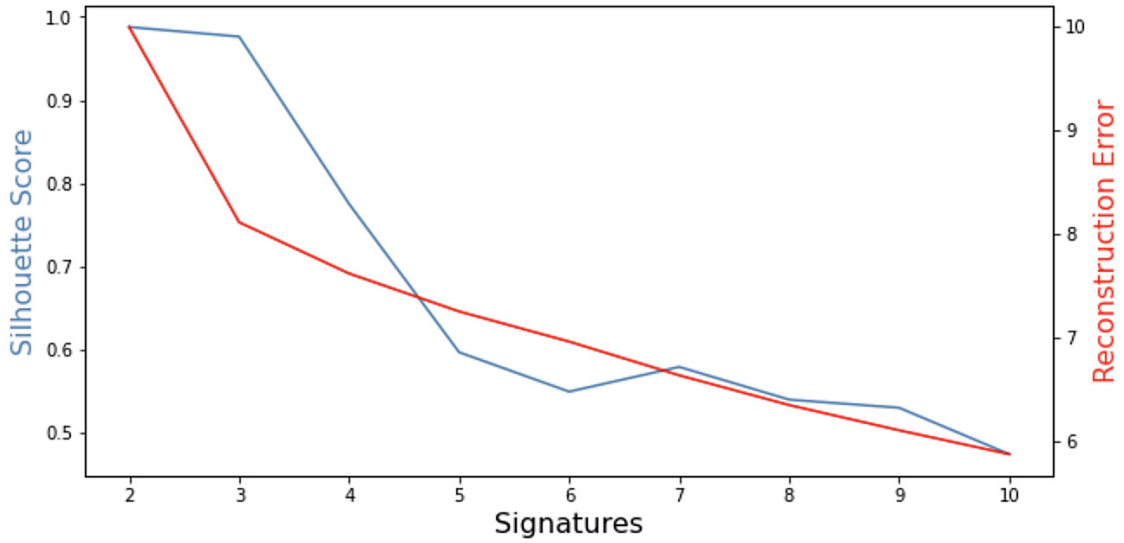


Figure 2.3: Signature evaluation metrics. The number of signatures was selected at N = 3 since this produced an "elbow" for the reconstruction error while having a suitable silhouette score greater than 0.95.

Table 2.1: Evaluation Results for Signature with N = 3.

Signature	Cosine	Silhouette
0	0.99983	0.957438
1	0.997578	0.973717
2	0.998414	0.997209

1.1.1 Non-negative least squares regression

A non-negative least squares (NNLS) Regression was used to produce positive exposure weights for each of the signatures in each of the datasets. The non-

negativity of the regression ensures that the weights of the signatures continue to represent an additive process. The NNLS weights can then represent the exposures of the signatures on each dataset.

2.4.5 Consensus lineage and continent signatures

Mutational catalogues were constructed for each continent and each of the Variant of Concern (VOC) lineages (Alpha, Beta, Gamma, Delta and Omicron). The global signatures were then used to extract exposures for each of the mutational catalogues to determine how processes varied between each mutational catalogue subset. VOC sequence sets were filtered so that weeks with fewer than 100 sequences were excluded.

2.5 Results

2.5.1 Characterising the SARS-CoV-2 waves regionally

This first part of the study reports on global SARS-CoV-2 data from 24/12/2019 to 28/01/2022 only as limited public health measures were in place after this time. We observed 1,544 distinct SARS-CoV-2 lineages from 7,348,178 sequences. 88% of the infections in the global pandemic during this time frame were caused by a subset of 13 Pango and WHO variants (Table 2.2). While there are geographical differences there is a clear dominance of a subset of variants and replacement of these through time (Figure 2.4). This “wave” infection pattern was evident in all geographic locations. Although biased by testing rates, Europe and the Americas had the highest infection rates, reporting up to 450 cases per million population per day (Figure 2.4).

Table 2.2: Proportion of common lineages/variants globally.

Lineage	Sum	Proportion
Delta	4,087,909	0.563
Alpha	1,150,798	0.158
Omicron	647,553	0.089
B.1.2	127,557	0.018
Gamma	121,393	0.017
B.1	111,849	0.015
B.1.177	74,643	0.01
Beta	40,786	0.006
B.1.1.214	18,160	0.002
D.2	13,340	0.002
B.1.621	11,050	0.002
B.1.1.284	9,334	0.001
C.37	9,287	0.001
Total	6,423,659	0.884

The emergence or introduction of VOCs coincided with a steep increase in infection rates globally. For example, cases in Asia showed a steep rise in February 2021, which peaked in May 2021 (Figure 2.4, panel Asia).During this period, Alpha and Delta comprised greater than 75% of the SARS-CoV-2 cases identified in the sequence data. Africa and Oceania on the other hand displayed overall sustained low case numbers. Despite this, Beta dominated the second wave in parts of Africa while Alpha dominated the third Oceanic wave. After its emergence in March 2021, Delta spread to become the predominant variant across all continents. The Omicron variant of concern was first identified in South Africa in late November 2021 and, by January 2022, it had rapidly become the predominant cause of infections worldwide

(Figure 2.4).

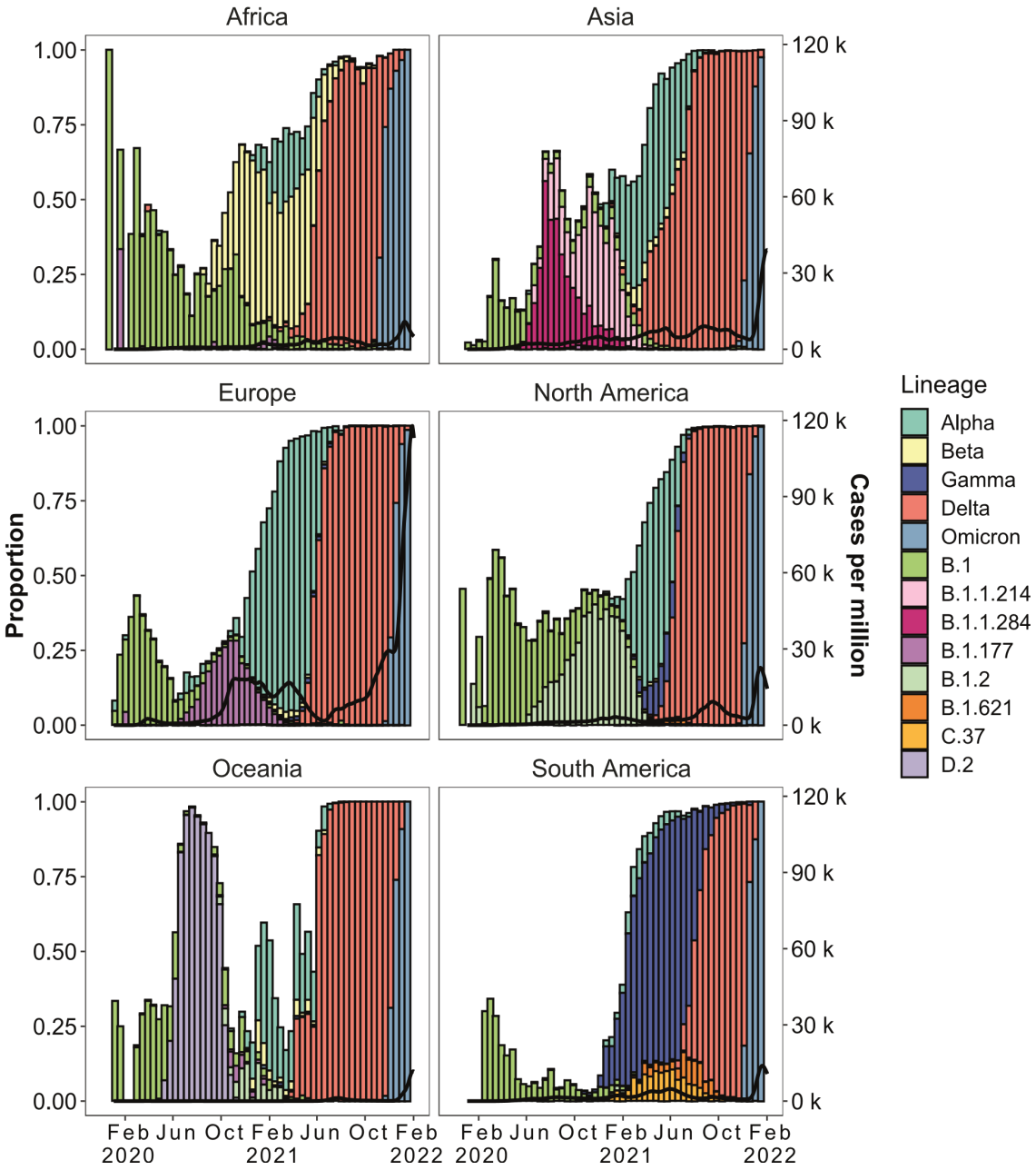


Figure 2.4: Continent-level SARS-CoV-2 lineage dynamics and pandemic curves. Lines show a 14-day rolling average of reported SARS-CoV-2 cases. Bars show the biweekly proportions of common lineages and are coloured by lineage. The white space shows the proportion of sequences from other (non-majority) lineages.

2.5.2 COVARIATES OF THE WAVES

We investigated the degree to which public health measures and viral properties explain continent-specific reported cases of infection. Correlation analysis at the global level showed a significant correlation between infection rates and the predictor variables: government stringency, vaccination, previous infection burden, virus diversity and fitness (Table 2.3).

Table 2.3: Correlation between infection rate and predictor variables across different continents. Virus fitness was shown to be positively correlated in all countries, while virus diversity was always negatively correlated. Stringency was predominantly negative, except for in North America and Oceania.

Continent	Infection Rate Correlation	Virus Fitness Correlation	Virus Diversity Correlation	Stringency Index Correlation
Africa	1	0.66	-0.69	-0.17
Asia	1	0.78	-0.75	-0.59
Europe	1	0.75	-0.47	-0.19
North America	1	0.63	-0.54	-0.75
Oceania	1	0.66	-0.4	0.29
South America	1	0.67	-0.31	-0.14

Regression analysis revealed that the impact of the predictor variables on the magnitude of reported cases were found across all continents. We classified significance levels as follows: no significance for p-values greater than 0.05, weak significance for p-values between 0.05 and 0.001, and high significance for p-values less than 0.001. Our findings indicated that government stringency had a weakly significant impact in Asia, Europe, and South

America, but a strongly significant impact in Africa, Oceania, and North America. Virus fitness, previous infection burden, and vaccination demonstrated a strongly significant impact across all continents.

Table 2.4: Effect of public health measures (government stringency and vaccination) and viral properties (diversity and fitness) on infection rates at continent level.

		Africa	Asia	Europe	North America	Oceania	South America
Diversity score	Estimate	832.4	1503.85	40828.15	15101.25	140.69	-339.15
	p-value	0.1	0.34	1.77E-26	4.89E-24	0.08	0.36
	Std.Error	511.12	1560.94	3663.33	1433.58	79.41	372.11
	t.value	1.63	0.96	11.15	10.53	1.77	-0.91
Fitness score	Estimate	730.01	1036.67	1680.87	506.22	228.52	1042.5
	p-value	3.60E-89	2.68E-23	2.84E-07	1.25E-08	1.42E-100	4.55E-86
	Std.Error	30.55	100.12	323.91	87.76	8.8	43.77
	t.value	23.89	10.35	5.19	5.77	25.98	23.82
Government index	Estimate	37.15	-14.76	28.13	-184.69	7.14	-4.28
	p-value	1.73E-12	0.11	0.41	3.10E-69	4.65E-09	0.43
	Std.Error	5.16	9.1	33.85	9.26	1.2	5.41
	t.value	7.21	-1.62	0.83	-19.95	5.95	-0.79
Intercept	Estimate	-2789.53	1931.43	-9258.87	12349.05	-573.13	1223.64
	p-value	6.31E-19	0.02	4.59E-06	3.99E-68	3.09E-09	0.01
	Std.Error	303.5	831.44	2003.05	625.45	95.26	445.29
	t.value	-9.19	2.32	-4.62	19.74	-6.02	2.75
Previous burden	Estimate	0.01	0.01	0.01	0	0.01	0
	p-value	2.05E-54	5.90E-28	4.32E-68	2.21E-05	3.79E-35	1.03E-12
	Std.Error	0	0	0	0	0	0
	t.value	17.24	11.51	19.75	-4.27	13.2	7.3
Vaccine doses	Estimate	-0.09	-0.01	0.04	0.04	-0.01	-0.02
	p-value	6.37E-22	2.32E-11	5.07E-23	8.25E-17	1.86E-06	3.06E-19
	Std.Error	0.01	0	0	0	0	0
	t.value	-10.01	-6.81	10.28	8.56	-4.82	-9.32
Residual S.E		750.68 on 600 DF	1253.55 on 626 DF	5772.52 on 637 DF	1445.60 on 640 DF	273.20 on 606 DF	713.85 on 536 DF
R squared		0.79	0.73	0.75	0.72	0.66	0.7
Abbreviations: **DF = degrees of freedom** **S.E = Standard error**							

Virus diversity was strongly correlated with high infection numbers in Europe and North America, with a weaker association in Africa, Asia, Oceania, and South America. The R-Squared values, indicating the proportion of variance explained by our model, were greater than 0.5 for all continents, ranging from 0.66 in Oceania to 0.79 in Africa (Table 2.4). Generally, our predictions closely resembled the rise and fall of SARS-CoV-2 infection case numbers (Figure 2.5).

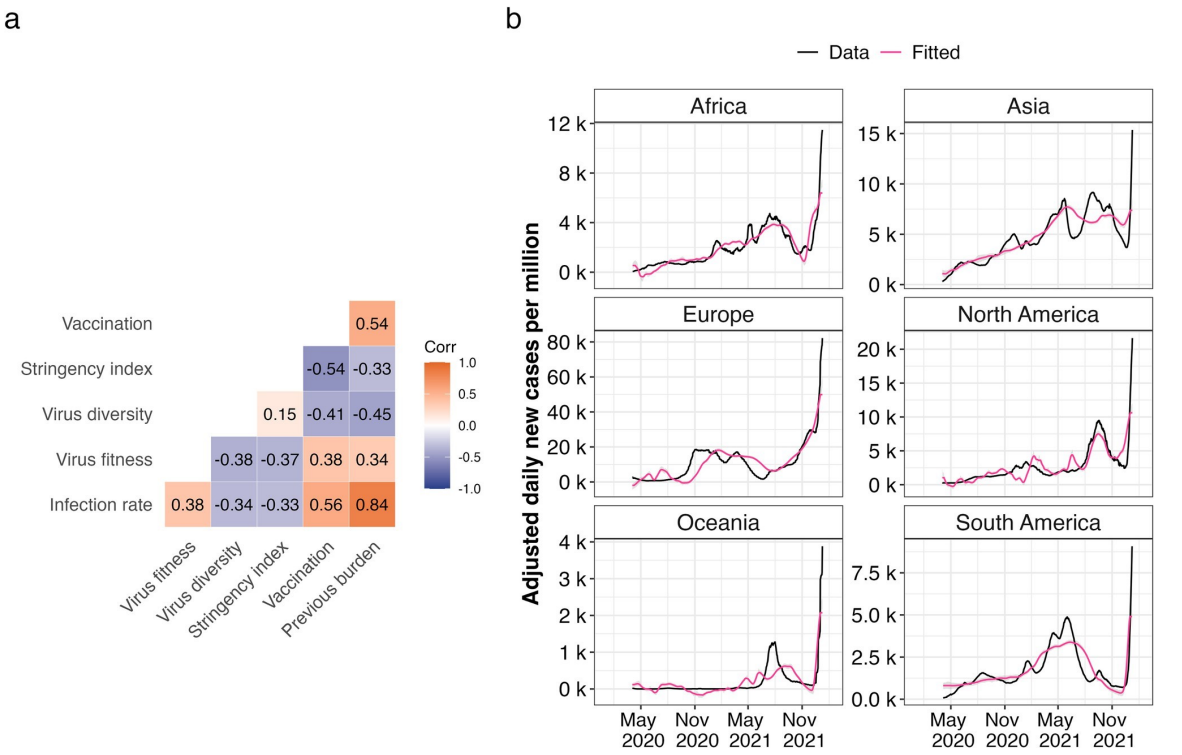


Figure 2.5: Association of SARS-CoV-2 infection rates and predictor variables globally. (A) Pearson's correlation matrix of infection rate and predictor variables. Positive correlations are denoted in orange and negative correlations in blue and colour intensity is is directly proportional to coefficient value. (B) Model fitting using multiple linear regression. Black solid lines show a 14-day rolling average of adjusted SARS-CoV-2 cases. Pink solid lines show fitted mean response values of infection rates with predictor values as input.

For country-level analysis, we included 29 countries from six continents based on the completeness of data (availability of sequence data in every 14 day bin). Pandemic plots were visualised using biweekly bins and multiple linear regression was fitted using the same approach. Different countries had varying lineage dynamics as illustrated in Figure 2.6.

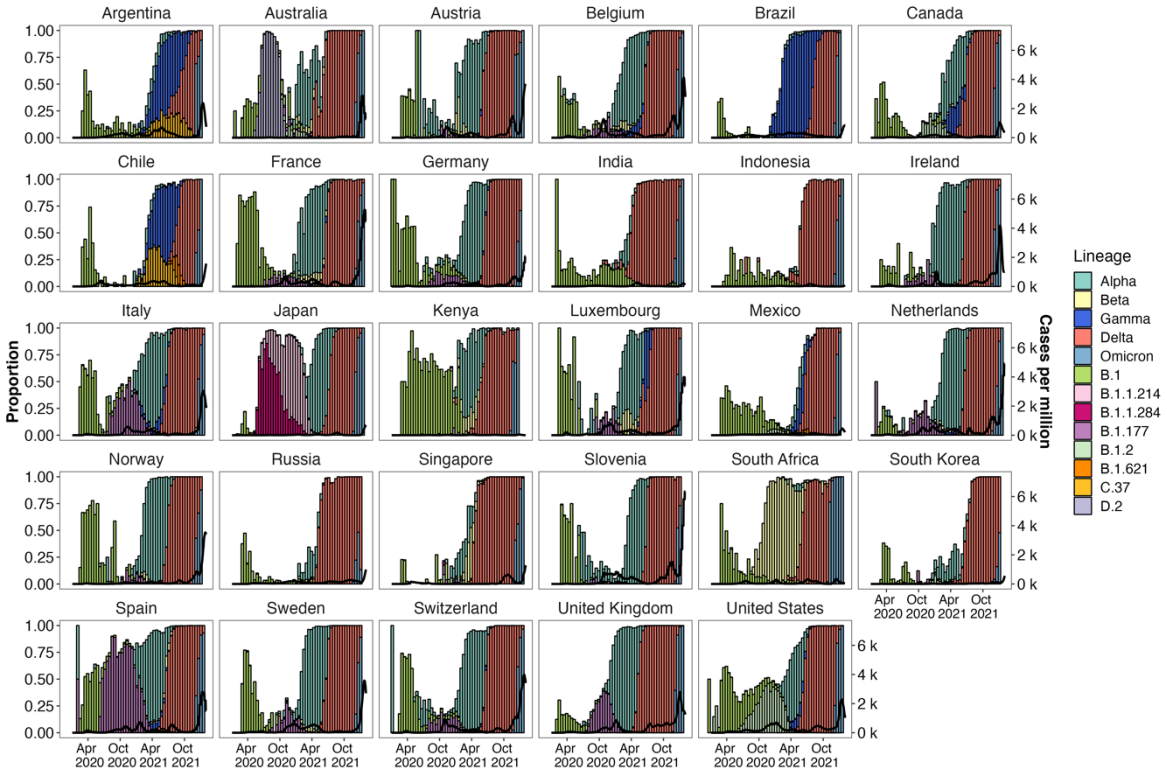


Figure 2.6: Country-level SARS-CoV-2 lineage dynamics. Solid bars show the biweekly proportions of the common lineages. Bars are coloured by lineage and white space shows the proportion of sequences from other lineages. The countries included in this analysis is based on temporal data completeness.

The five predictor variables had varying impacts on infection rates across countries (Figure 2.5). Despite some differences related to the population level processes investigated here, there is a clear variant replacement process taking place. As the generation of novel variants is fundamentally a mutation dependent process we next investigated the underlying patterns of mutations being generated through time. The goodness of fit varied among countries, with the R squared (a measure of how much variation can be explained by the

predictor variables) varying from 0.28 (Japan) to 0.96 (Australia), with a median of 0.69 (Table 2.5). Though our model successfully captured the general infection wave patterns in many countries, it struggled to capture short-term data spikes in specific instances, such as in Belgium (November 2020), India (May 2021), Indonesia (August 2021) and Japan (September 2021) (Figure 2.7).

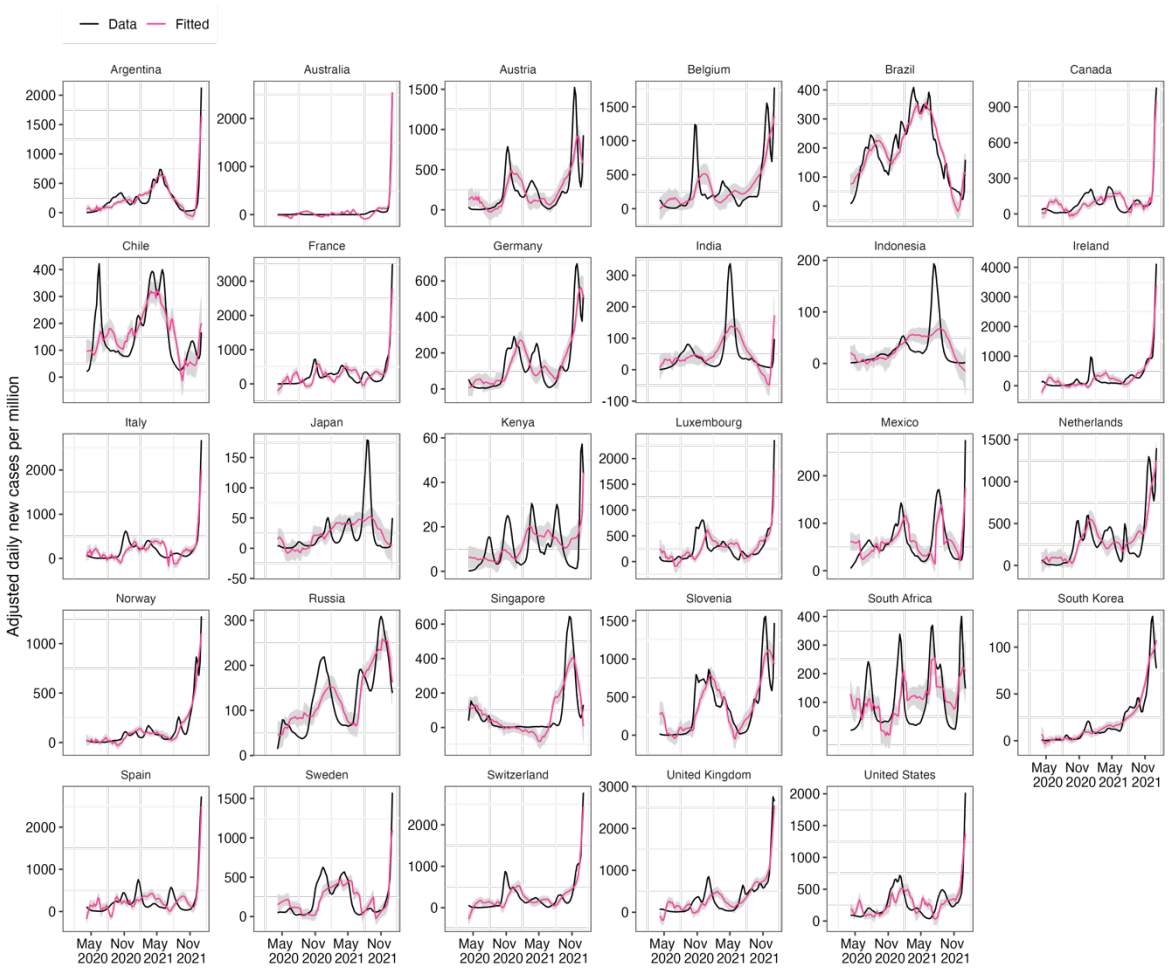


Figure 2.7: Model-fitting of country-level SARS-CoV-2 reported cases. Black solid lines show a 14-day rolling average of adjusted SARS-CoV-2 cases. Pink solid lines show fitted mean response values of infection rates with predictor values as input and grey shaded areas highlight the confidence intervals. The countries included in this analysis is based on temporal data completeness.

Table 2.5: Effect of public health measures (government stringency and vaccination) and viral properties (diversity and fitness) on infection rates at national levels. Coefficients describe how large an impact the variable has on the prediction, with higher magnitude values indicating stronger impact either positively or negatively towards the prediction. R-squared indicates percentage of variation explained by the model, with a max score of 1 and higher values indicating a better model.

Country	Intercept	Government Index Coefficients	Vaccine Doses Coefficients	Fitness Score Coefficients	Diversity Score Coefficients	Previous Burden Coefficients	R-Squared
Argentina	-391.02	2.42	-0.01	318	431.67	0	0.86
Australia	-270.08	2.73	-0.06	111.8	46.23	0.11	0.96
Austria	-318.86	3.29	0.36	-84.88	450.04	0.01	0.54
Belgium	-57.36	-1.17	0.04	4.37	373.74	0.01	0.54
Brazil	-140.05	2.76	-0.01	38.45	25.43	0.01	0.81
Canada	-1011.36	11.85	0.02	175.49	509.21	0	0.8
Chile	550.61	-5.93	-0.03	65.26	-31.82	0.01	0.53
France	-2038.69	9.04	0.06	641.03	2842.56	0.01	0.75
Germany	165.55	-1.99	0.01	-85.54	-42.31	0.02	0.69
India	-28.35	-0.23	-0.03	76.94	116.53	0	0.44
Indonesia	36.03	0.21	-0.01	2.08	-70.45	0	0.35
Ireland	-749.06	0.74	0.01	388.13	1223.4	0.02	0.84
Italy	-1647.1	13.98	0.03	374.47	1419.41	0.01	0.72
Japan	23.36	0.48	0	-0.4	-75.31	0	0.28
Kenya	20.71	-0.14	-0.02	8.81	-4.36	-0.01	0.32
Luxembourg	-791.67	7.3	0.04	224.14	763.4	0.01	0.59
Mexico	-208.26	3.46	0	37.41	-35.13	0.01	0.46
Netherlands	-343.6	1.24	0.33	-45.1	535.95	0.01	0.66
Norway	-337.8	1.91	0.03	12.56	418.25	0.02	0.9
Russia	145.48	-0.76	0.02	-32.88	-72.81	0.01	0.67
Singapore	12.02	2.67	0.02	-97.97	-237.39	0.01	0.69
Slovenia	-471.64	5.5	0.09	-145.49	460.28	0.01	0.73
South Africa	-123.02	4.85	-0.05	39.76	-338.02	0	0.32
South Korea	9.48	0.14	0.01	6.17	-27.66	-0.01	0.89
Spain	-1292.19	2.61	0.07	476.79	2058.39	0.02	0.7
Sweden	-1084.41	18.58	0.47	80.66	155.09	0	0.61
Switzerland	-651.84	0.61	0.05	226.8	1498.23	0.02	0.77
United Kingdom	-417.36	-3.26	0.03	263.37	1717.29	0.01	0.79
United States	-1427.87	16.63	0.06	145.45	1108.11	0	0.62

2.5.3 Identifying putative mutational processes contributing to changes in SARS-CoV-2

New variants of concern have displaced viral lineages that were previously dominant in the population in different geographical regions and in some cases

globally (Figure 2.5). This behaviour has been observed with the original variants of concern (Alpha, Beta and Gamma) and then globally with the Delta and Omicron lineages. We investigated whether these variant wave events (periods of time where infections are dominated by a single variant) were linked to the activity of specific mutationlal processes. Each of the variants of interest/concern has evolved independently such that detecting the patterns of mutations in the SARS-CoV-2 sequence data allows us to observe which processes are most active and could be contributing to the emergence of variants.

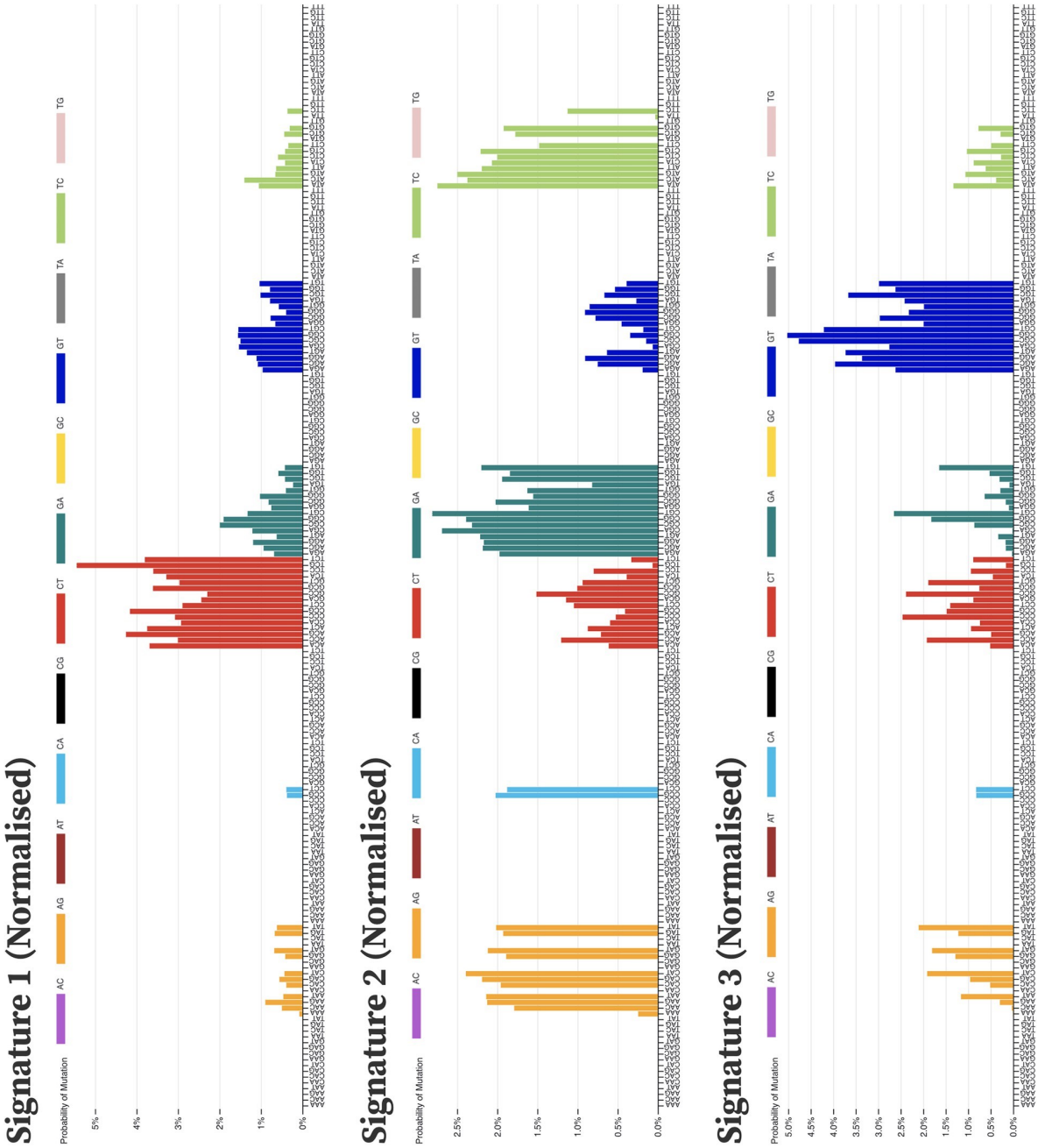


Figure 2.8: Mutational signatures extracted from the SARS-CoV-2 genome sequences by non-negative matrix factorisation. Signatures are patterns of probabilities for each category of substitution in a three nucleotide context. Each bar represents a context and is coloured by the substitution category of the mutation that occurs there. Each signature may represent a distinct mutational process. Signature 1 is heavily biased towards cytosine to thymine (C→T) mutations, particularly in 3' CpG contexts TCG, CCG and ACG. Signature 2 from SARS-CoV-2 is predominantly adenine to guanine

(A→G), guanine to adenine (G→A) and thymine to cytosine mutations (T→C). Signature 3 is strongly guanine to thymine (G→T), a pattern that is thought to be caused by the action of guanine oxidation by reactive oxygen species. Signatures are shown normalised against the tri-nucleotide composition of the SARS-CoV-2 genome. Non-normalised forms in the context of the SARS-CoV2 genome composition are shown in Figure 2.9.

Mutations were called using inferred references for each of the Pango lineages, which we call tree-based referencing (Figure 2.8). The SARS-CoV-2 alignment of 13,278,844 sequences up to 26/10/2022 was used. Of these 13 million sequences 2,195,182 sequences were selected as they contained 5,726,144 newly arisen mutations. Cytosine to thymine mutations (C→T) were the most common and were the primary substitution category for most weeks where sequences were recorded. Note, SARS-CoV-2 has an RNA genome but we refer to uracil as a thymine to match pre-existing DNA mutational signature notations.

Three signatures were identified with distinct substitution patterns using non-negative matrix factorisation (NMF) (Figure 2.8 and Figure 2.9). Signature 1 is heavily biased towards C→T mutations. Signature 1 had a high probability of ACA, ACT and TCT contexts (adjacent nucleotides in the 5' and 3' direction of the mutated site), consistent with what was earlier reported by Simmonds et al.[141] as highly mutated contexts for C→T substitutions in SARS-CoV-2. Signature 2 is predominantly adenine to guanine (A→G), guanine to adenine (G→A) and thymine to cytosine (T→C) mutations. The proportion of A→G and T→C mutations is approximately equal in this signature, which is indicative of

a double-stranded mutational process. SARS-CoV-2 mutations at adenine positions on the negative strand will be counted as thymine mutations due to the negative strand being used to replicate positive sense RNA, with the mutated A→G now pairing with a cytosine on the +sense RNA and replacing the original thymine[155,156]. Signature 3 is predominantly composed of guanine to thymine (G→T) substitutions.

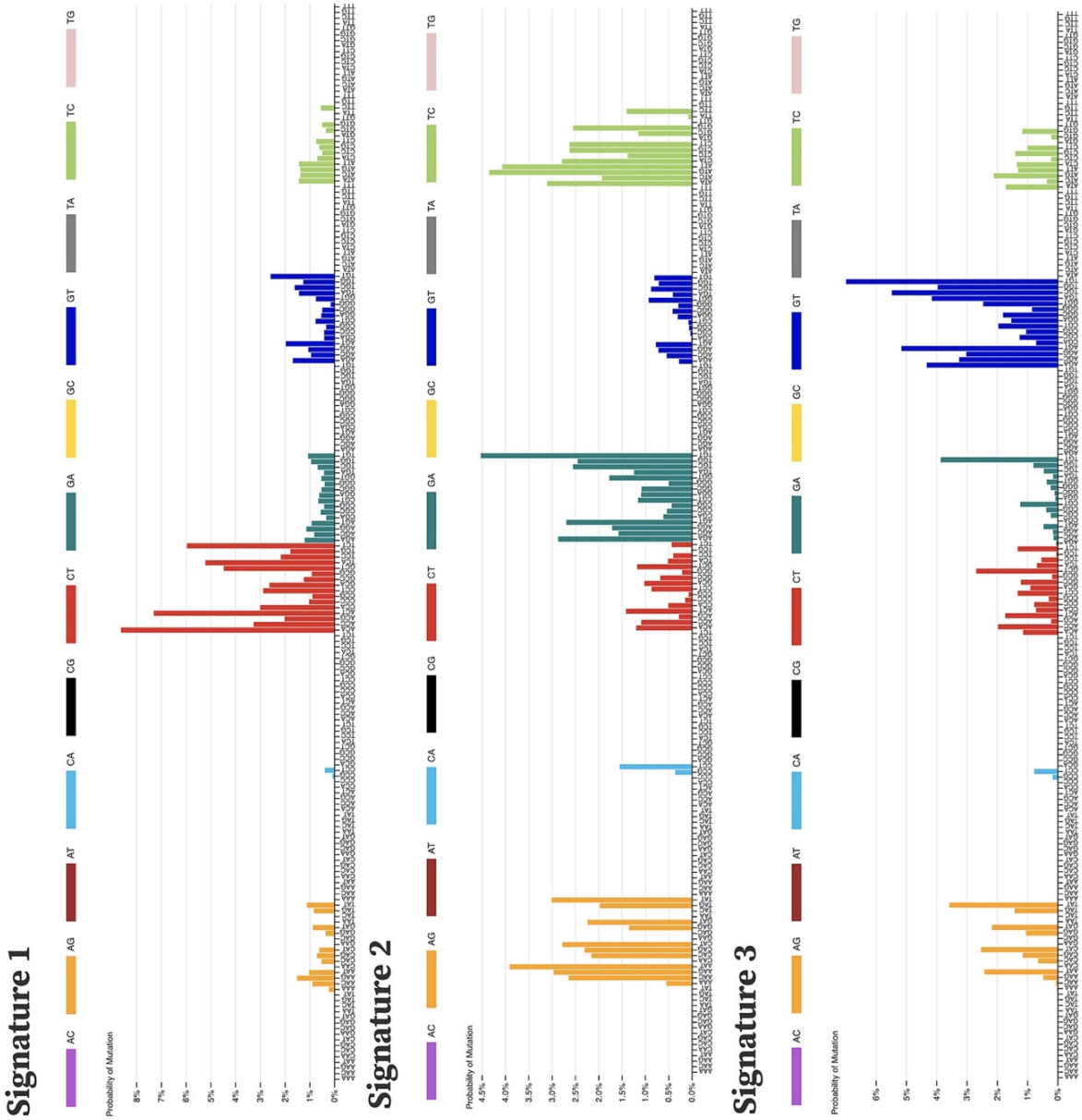


Figure 2.9: Non-normalised mutational signatures for SARS-CoV-2. Signatures were extracted using normalised counts calculated by dividing the mutation counts by the count of the tri-nucleotide context of the mutation context (Figure 2.8). These signatures were then multiplied post-analysis by the tri-nucleotide composition of the reference sequence to produce the non-normalised signatures shown here.

2.5.4 THE DYNAMICS OF MUTATIONAL PROCESSES THROUGH THE PANDEMIC

By using the available SARS-CoV-2 sequences we can measure the mutational signature activity across time as long as our samples are aggregated using time series annotations. Signature exposures (Figure 2.10) show that Signature 1 remained the most prominent signature throughout the pandemic, although following the emergence of Signature 2 its activity reduced proportionally. Absolute exposure values (Figure 2.10 B) show that Signature 1 does not appear to reduce its exposure, rather Signature 2 increases its exposure. Signature 2 establishes itself as a substantial signature after December 2020. It continues to expand after October 2021, just prior to the emergence of the Delta VOC. Signature 3 is by far the least active of the three signatures but remains consistent until after January-February 2022 when it begins to drop towards zero. This is around the time Omicron began to emerge as the dominant VOC. Combined signature activity reached a peak between July and October 2021 (Figure 2.10B) coinciding with the peak number of unique mutations (Figure 2.11, Figure 2.12A and Figure 2.12B). This is around the time the mutational signature dynamics appear to be shifting, with Signature 2 contributing more unique mutations. We can see that this also coincides with the Delta VOC wave, which, between May 2021 and January 2022, was the lineage group showing the greatest number of newly acquired mutations (Figure 2.12). Delta was the first VOC to dominate on a global scale, outcompeting other VOCs like Alpha, Beta and Gamma in their regions of circulation.

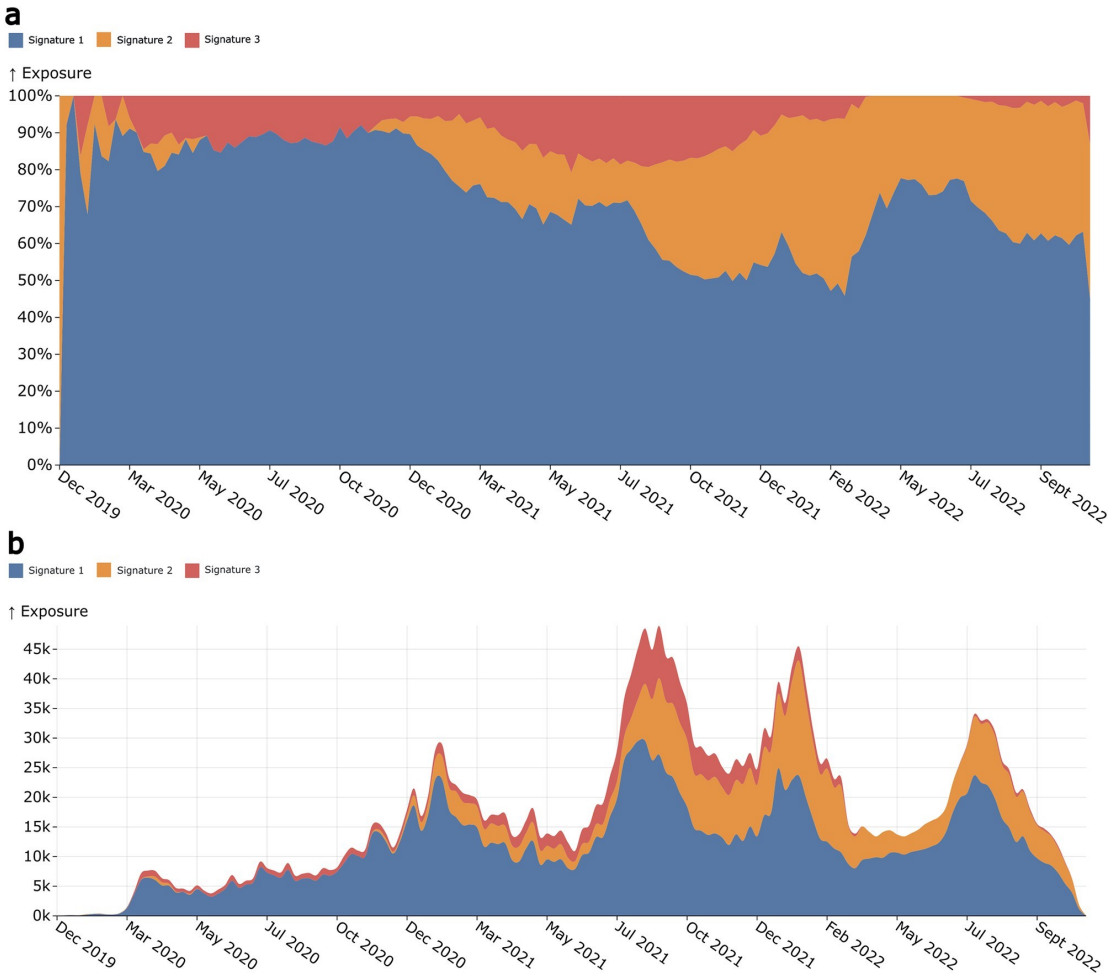


Figure 2.10: Signature exposure plots showing the activities of the extracted mutation signatures over the duration of the COVID-19 pandemic. (A) Shows the percentage activity of the signatures during a given week of the pandemic, with each colour representing a different signature. (B) Shows the signature activities as their absolute values at each epidemic week.

Omicron similarly repeated this phenomenon, almost entirely replacing Delta globally within weeks of its emergence (Figure 2.12B). We also see a marked decrease in the activity of Signature 3 following Omicron's establishment as the dominant variant. A similar decrease in G→T mutations was also observed by Bloom et al.[157] and Ruis et al.[158]. This is different to Delta, where there was an increase in Signature 3 following its emergence. These Signature 3 changes

become particularly apparent when we begin to look at signature activities within variant-defined subsets of the data.

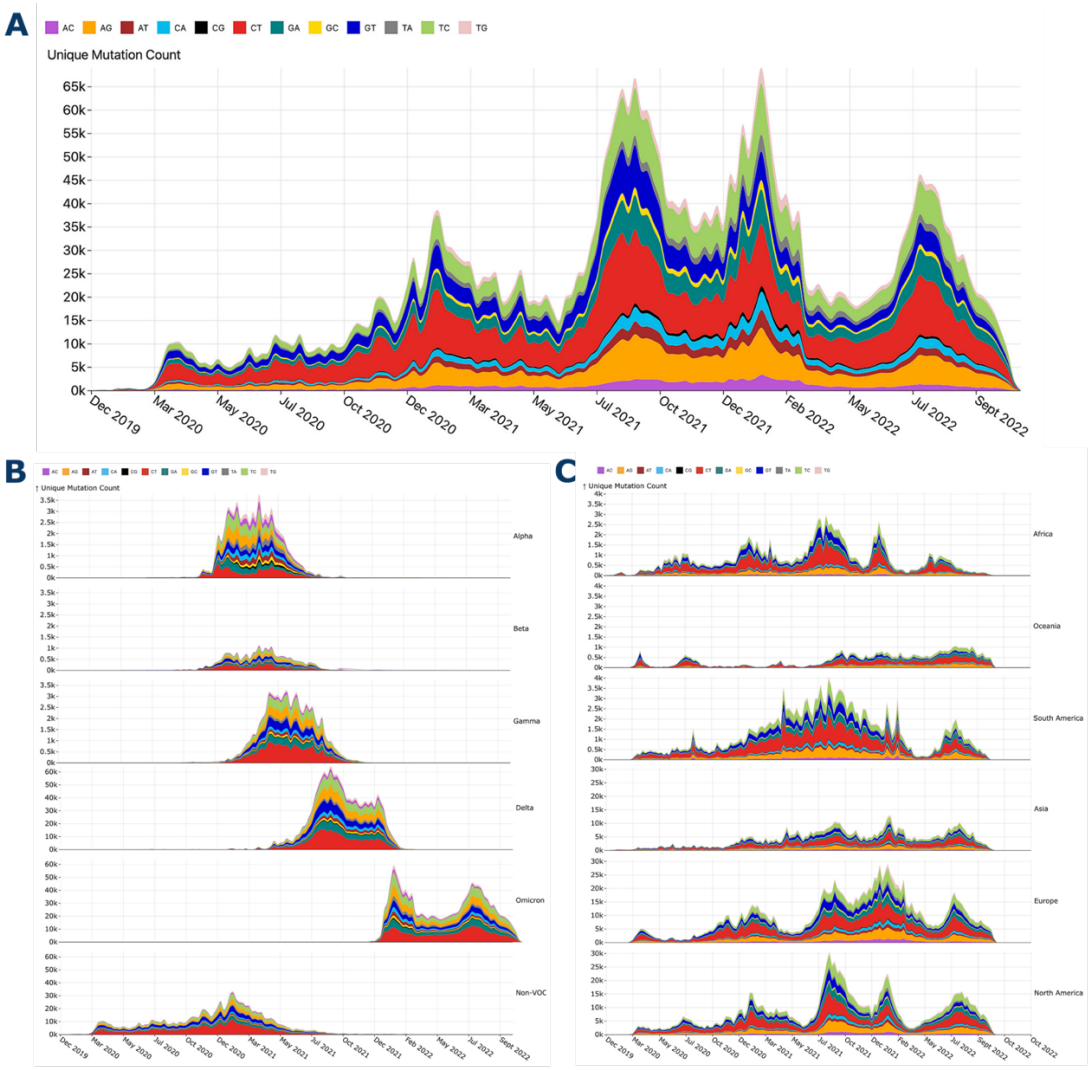


Figure 2.11: (A) Counts of unique substitutions per week of the pandemic. Areas are coloured by substitution category. (B) Counts of unique substitutions per week of the pandemic for each VOC category. Areas are coloured by substitution category. (C) Counts of unique substitutions per week of the pandemic for each continent category. Areas are coloured by substitution category

2.5.5 SIGNATURE DYNAMICS SPATIALLY AND BY VARIANT

After observing changes in signature activity during transitions between dominant variants, we next investigated the differences between signature activities in variant-defined subsets of the data as well as in continent-defined subsets. We used the globally extracted signatures to extract exposures from the subsets using a non-negative least squares regression to retain the non-negativity constraint. This allowed for the measurement of signature activity in each of the subsets of interest. Signature 1 was the most active in almost all the variant-defined subsets as was expected from the global activity. Signature 3 was most active in the Delta subset as well as during the Delta wave in the continent-defined subsets (Figure 2.12). The non-VOC, Beta and Omicron subsets appear to be the least impacted by Signature 3 with almost zero activity in Omicron. Signature 2 also shows low activity in the non-VOC subset but is very active in the other VOC subsets, in particular Alpha, where it appears to be the most active, overtaking the Signature 1 process. Continent-defined subsets of the data also consistently showed the high activity of Signature 1. Signature 2 begins to consistently appear in all continents after 2020, with only small bursts of activity being detected before this (Figure 2.12D), again consistent with what we see in the global data. Signature 3 activity also follows the pattern of the global activity, appearing most prominently during the Delta wave.

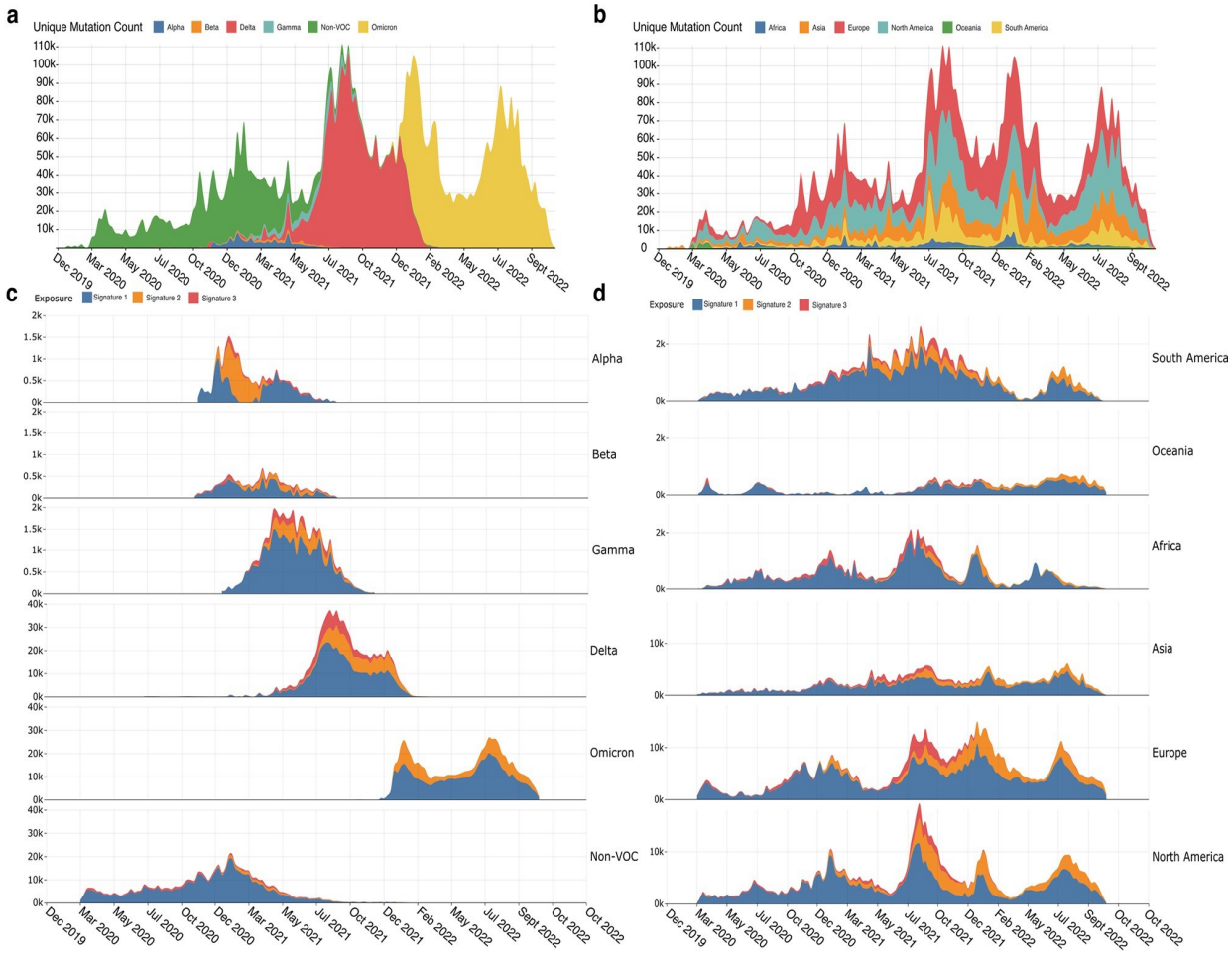


Figure 2.12: (A) Counts of unique SARS-CoV-2 mutations for each epidemic week, with colours representing which continent the mutations came from. (B) Counts of unique mutations per week that are part of the mutational signature substitution-context features (i.e., no indel mutations included). Colours represent which lineage/group of lineages the mutations belong to. (C) Ridgeline plot showing the exposure of mutational signatures in SARS-CoV-2 variant-defined subsets. Exposures are coloured by the signature they have been attributed to. (D) Ridgeline plot showing the exposure of mutational signatures in SARS-CoV-2 continent-defined subsets.

2.5.6 Bridging the gap between mutation signatures and amino acid substitutions

Stratifying non-synonymous nucleotide substitutions by their association with mutational signatures should provide insights into how these mutational processes affect viral proteins. Exposures were calculated by stratifying nucleotide mutations by whether they were synonymous or non-synonymous substitutions for each dataset (Figure 2.13A). The unattributed exposure was calculated using the model error for mutational categories not contained within any of the extracted mutational signatures. The majority of non-synonymous substitutions can be described by the observed mutational signatures. Signature 1 likely produces most of the nonsynonymous mutations, however, Signature 3 is an almost exclusively non-synonymous signature, with particularly high activity during the Delta wave of infections. Signature 2 appears to produce predominantly synonymous mutations.

Using the tree-based references, we can also look at individual lineage reference sequences to observe which mutational processes have probably produced their specific amino acid substitution set. The tree-based references were used since they are equivalent to a high-quality representative sequence and because many of the early real sequences contain sequencing errors. For each variant of concern, mutations were assigned to a signature by calculating the maximum likelihood of the mutation and its context being produced by each of the three extracted signatures. Using the trinucleotide context C[C → T]G as an example, the likelihood function is P(C[C → T]G j Signature), which corresponds to the probability bars for CT-CCT in the extracted signatures. Mutations that contained substitution-context pairs not found within any of the mutational signatures were labeled as “unattributed”.

The Alpha VOC tree-based reference sequence contains eleven Signature 1 changes, six Signature 2 changes and a single Signature 3 change. Signature 1 changes account for 39% of all substitutions within the Alpha tree-reference sequence, with 75% of these mutations being non-synonymous substitutions. Signature 1 was frequently active prior to the Alpha VOC's emergence. The activity plots (Figure 2.10) show that this was the case for much of the pandemic, particularly prior to the Alpha's emergence around September 2020. It should be noted that while Signature 1 mutations are by far the most frequent, only one is found within the Spike protein (producing the S:T716I change). Signature 3 only had one change, which was non-synonymous appearing in ORF:8. Signature 2 mutations were non-synonymous substitutions 83% of the time, with three Spike mutations relating to the process including S:D614G, which is present within all known variants of concern.

The Beta VOC emerged around the same time as Alpha (Autumn 2020) and is defined by a smaller set of mutations. A greater proportion of Signature 1 mutations are non-synonymous substitutions in Beta (66%). Signature 2 mutations resulted in S:D215G and S:E484K, the latter reported to help the virus evade neutralising antibodies[159]. Signature 3 mutations most likely produced S:K417N in spike, which is also reported to aid in antibody evasion[159,160] similar to S:E484K.

Gamma also emerged in Autumn 2020 and has 33 different defining substitutions. Signature 1 mutations account for 11 of these with 54% being non-synonymous. Four are present in Spike including S:L18F, S:P26S, S:H655Y and S:T1027I. Signature 2 mutations resulted in six amino acid

substitutions, with only 75% of changes being non-synonymous. Three of the five mutations in non-synonymous substitutions occurred in Spike. Signature 3 mutations in the Gamma lineage were all non-synonymous except for a single synonymous substitution in ORF1a/b.

Delta was the first VOC to dominate worldwide and replace almost every other lineage in all regions. The initial Delta sequence (Pango lineage B.1.617.2) contains six Signature 1 mutations. 66% of these changes were non-synonymous and none occurred within Spike. Signature 2 mutations were all non-synonymous and displaced throughout the virus ORFs including ORF1a/b, S and M. Signature 3 mutations in Delta are found in non-coding regions and N, with the N mutations both being non-synonymous.

Omicron is the most recent VOC to emerge, quickly replacing Delta globally. Omicron differs from earlier VOCs with a much greater number of Spike mutations relative to the other ORFs. The first identified Omicron variant B.1.1.529 has 40 substitutions of which 32 are nonsynonymous changes. This is almost double that of Delta, which only had 18. Seven of these substitutions were Signature 1 changes, two were Signature 3 and ten were Signature 2 changes. There are four non-synonymous ORF1a/b mutations despite this ORF being substantially longer than SARS-CoV-2's other ORFs. Only one Spike substitution was synonymous out of the 21 total changes. This number is even greater when looking at the major Omicron variants BA.1 and BA.2. BA.1 had 31 non-synonymous substitutions in Spike alone while BA.2 had 28. Between these three Omicron variants, only two Spike substitutions are non-synonymous out of a total of 40. Nine of the 40 changes are from Signature 1, 2 are from Signature 3 and 12 are from Signature 2. This means 23/40 of the

changes appear to come from these three mutational processes. 20 of the 40 substitutions observed in these variants were present in the receptor-binding domain (RBD) of Omicron, with nine of these changes thought to help Omicron evade the immune response or increase its transmissibility[161]. Of these beneficial RBD changes, three are potentially the result of Signature 1 activity, 9 are Signature 2 and one is from Signature 3. The high density of Signature 2 RBD amino acid changes in a variant that has emerged as Signature 2 exposure increased suggests that the mutational process behind Signature 2 may have contributed to the emergence of the Omicron variant.

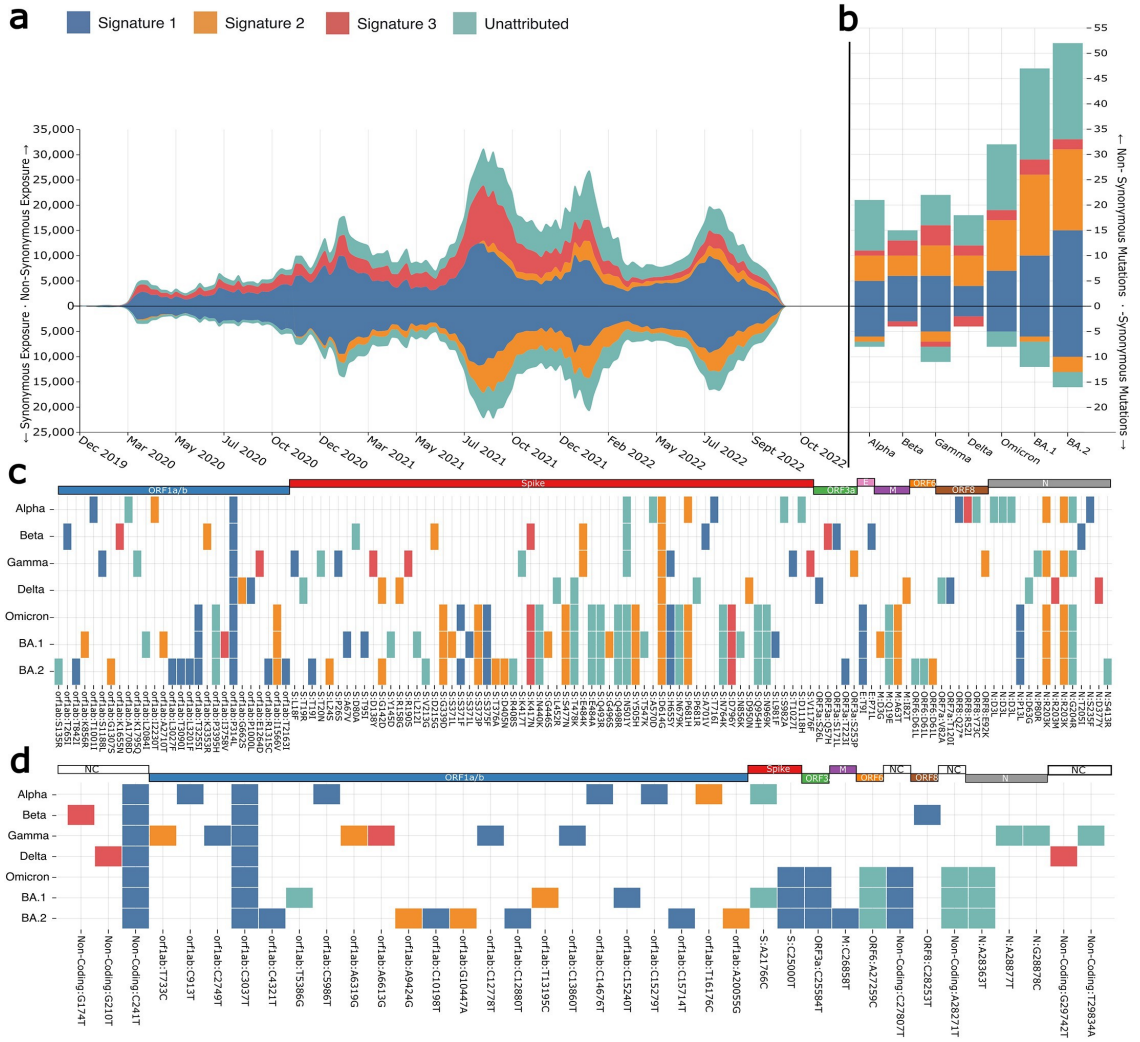


Figure 2.13: (A) Exposures for each of the SARS-CoV-2 mutational signatures for both synonymous and non-synonymous stratified datasets.

Synonymous exposures are below 0 on the y-axis, while non-synonymous exposures are above 0. Each area represents signature exposures across epidemic weeks, with colours representing which signature the exposures are attributed to. (B) Non-synonymous and synonymous mutations in the tree-based references of identified variants of concern. Signature 1 produces the majority of both synonymous and non-synonymous substitutions in all lineages. Signature 3 mutations are more often non-synonymous substitutions in the lineages of concern, with most lineages having few to no changes. Signature 2 non-synonymous mutations appear to have increased in the Omicron lineages (BA.1 and BA.2). (C) Variant of concern associated non-synonymous mutations coloured by the mutational signature with the greatest likelihood of causing the change. (D) Variant of concern synonymous mutations coloured by the putative mutational process that caused the change.

2.5.7 SIGNATURE EXPOSURES AND HIGHLY MUTATED SEQUENCES IN WASTEWATER DATA

Similar trends over time in exposures are seen when the mutational signatures are applied to publicly available wastewater data. Although the trend is seen at a lower resolution than global data, Signature 1 and Signature 3 are gradually replaced by Signature 2 (Figure 2.14A). Although, Signature 2 is not quite as strong as in the global data (Figure 2.10). This suggests trends in mutational processes can be monitored using wastewater, not only sequencing of the infected population. Additionally, at time periods where a high level of virus diversity is expected, there are highly mutated sequences present in the wastewater (Figure 2.14C). This suggests cryptic sequences in wastewater may be used to observe potential upcoming variants, similar to

how known sequences have been back-traced to particular buildings using wastewater[162].

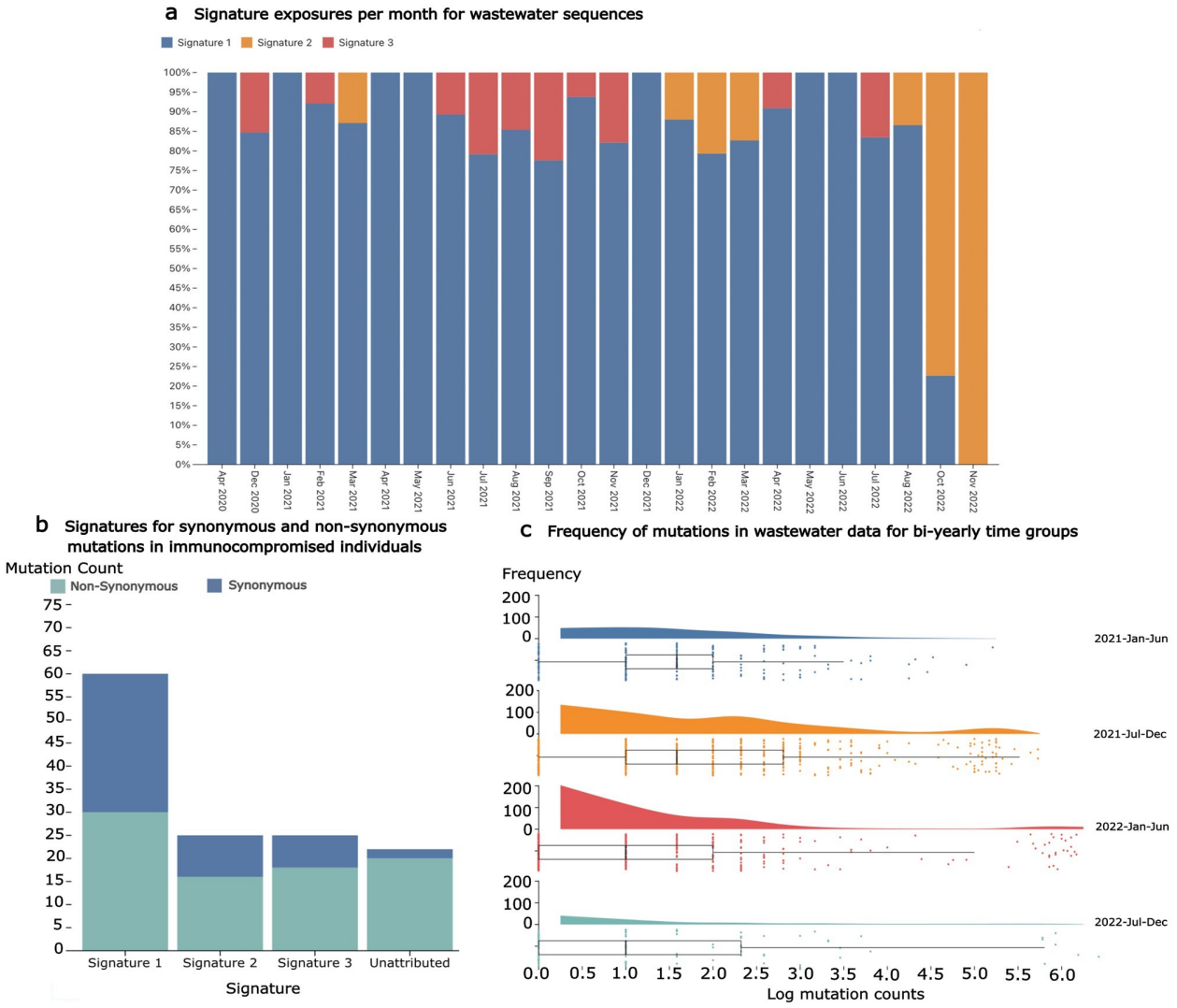


Figure 2.14: (A) Signature exposures per month from wastewater sequences show similar trends in mutational processes as the global data, although at a lower resolution and, interestingly, with a lower Signature 2 exposure. (B) Substitutions in SARS-CoV-2 consensus sequences from infections of immunocompromised individuals contain mutation types corresponding with patterns observed in the distinct signatures. Of note, there are more synonymous mutations present in the chronic infection data than in the global sequences, although it is important to note the sample size for immunocompromised infections is low. (C) Mutation counts

in wastewater sequences for bi-yearly time periods. Highly mutated sequences cluster to the right especially during the 2021 July-December time period, as would be expected when Omicron was emerging.

As chronic SARS-CoV-2 infections are implicated as a major contributor to VOC evolution[148,163], it may be possible to parse highly-mutated cryptic sequences of interest from chronic infections out of wastewater data in the interest of detecting potential VOCs. Unfortunately, this is problematic to deconvolve as sequencing data for immunocompromised and chronically infected individuals is sparse. When sequences from known chronic infections are examined, the distribution of mutation types is consistent with global data, with Signature 1 mutations dominating as expected for samples from January 2022 (Figure 2.14B). Although, due to the low number of chronic infections for comparison this result is not very conclusive, it does demonstrate how mutational patterns can be potentially detected in this type of data. Studying these types of infections, and underlying mutational processes, will be important to understand better the origins of the sets of mutations that contribute to the generation of VOCs.

2.6 Discussion

In this study, we investigated SARS-CoV-2 lineage dynamics and identified temporal variables that are associated with increased numbers of infection cases. Both public health measures and virus properties were associated with the sequential waves of regional SARS-CoV-2 infections cases. These predictors have varying impact in different geographical locations. As more of the global population's immune system becomes sensitised to existing SARS-CoV-2 variants, either through previous infection or vaccination, the virus has

and will continue to undergo changes that enable reinfections. The continued emergence of new variants is thus expected. In some regions, government stringency had limited significant impact on patterns of infection. This could be due to differences in implementation strategies and support, other competing predictor variables, as well as behavioural changes in citizens as a response to the restrictions.

Our analysis highlights the significant role of vaccination in influencing reported COVID-19 case patterns across all continents, even in regions with lower vaccination coverage like Africa. Despite Africa's lower vaccination rates, the continent has seen a relatively low-level of sustained transmission. This phenomenon might be attributed to factors such as the younger median age of the population, lower population density, immune priming due to prevalent infectious diseases, and limited testing capacity[164]. The weak impact of viral diversity on reported cases in Asia and South America may be explained by the emergence and dominance of variants such Delta and Gamma in the regions, respectively. For instance, the Delta variant, initially identified in Asia, quickly became the predominant strain, overshadowing other lineages before spreading globally. Overall, the predictor variables significantly contributed to explaining the rise and fall of infection numbers across different continents, accounting for more than half of the variance in reported cases. The differences in the regression effectiveness can be attributed to intrinsic differences among continents, such as variations in vaccine coverage, testing and sequencing capabilities, and the effectiveness of government stringency measures.

While our model effectively captured the general trends of infection waves, it struggled to accurately represent peaks within short time-frames in some

countries. This discrepancy might be attributed to the omission of certain predictor variables, like mass gatherings, which are known to contribute to viral super-spreading events[165].

In utilising the OWID and OxCGRT datasets, which are arguably among the most comprehensive for addressing our research objectives, we note some limitations. First, there were discrepancies in parameter definitions, such as varying case classifications across regions. Second, positive tests are commonly labeled based on their reporting date rather than "date-of-event"[145]. Lastly, the cases reported in these datasets may not be fully representative of the actual disease burden. Although the Human Development Index (HDI) of a country can act as a proxy to bridge the gap between reported cases and the true disease burden, it does not fully capture the entire complexity.

The extracted signatures from the global SARS-CoV-2 dataset show clear and distinct patterns describing mutational processes acting on the viral genome. The most prominent of these signatures, Signature 1 (Figure 2.8 and Figure 2.9), shows a marked bias towards C→T mutations, a signal indicative of the APOBEC family of cytidine deaminases[141,142]. APOBEC enzymes have been shown to cause extensive C→T editing of DNA and RNA in human and viral genomes. However, it is not yet clear whether they are the cause of this pronounced C→T bias in SARS-CoV-2 despite a number of other studies also observing other APOBEC-like mutational patterns[152,166–169].

Di Giorgio et al.[167] used single nucleotide variant counts from 8 samples of RNA sequencing data to identify possible RNA editing via host RNA editing mechanisms. They suggest APOBEC and ADAR as mechanisms due to the

substitutions observed. Graudenzi et al.[166] and Aroldi et al.[169] both identify the same mutational patterns and attribute to the same mutational mechanisms with an additional G→T signature identified and labelled as reactive oxygen species (ROS). Aroldi et al.[169] expand on the results from Graudenzi et al.[166] by investigating mutational signature exposures on the continents and for VOC samples, briefly mentioning how a lower APOBEC signal was observed in the Omicron sequences. Graudenzi et al.[166] and Aroldi et al.[169] both do mutational signature analysis using NMF on raw reads in order to quantify the intra-host signal. While our analysis comes to similar conclusions about a number of these points, our methods differ from each prior study in a number of ways. Firstly, we look at consensus sequences rather than raw sequencing reads which means we are tracking the mutations produced by these mechanisms that are viable to virus survival and contribute to its evolutionary history. Secondly, we do proper signature analysis using non-negative matrix factorisation as described by Nik-Zainal et al.[170] rather than simply counting mutations. Finally, we do comprehensive analysis of the exposure matrices, characterising the effect of the mutational signatures through time for both continents and variants of concern. We characterise the important mutations likely caused by the identified mutational processes, show the preponderance for each signature to pick synonymous or non-synonymous sites, and delve into the wastewater and immune compromised samples. We expand in meaningful ways upon each of the prior studies in several respects, while also reproducing many of their results, suggesting a robust set of analyses with novel insights.

Cytosines flanked by either an adenine or thymine in both the 3' and 5' direction appear to be the most pronounced targets of Signature 1. APOBEC editing was shown to have contexts outside of the traditional TpC when structural features of the nucleic acid such as hairpin loops are present[171]. Outside of structural features, APOBEC3A is thought to be the predominant cause for TpC changes and is found to be expressed in lung tissue[172]. ApC changes are considered to be caused by APOBEC1, which in cell models was shown to efficiently edit SARS-CoV-2 RNA[172]. APOBEC1 is found predominately in the liver and small intestine, tissues reported to be infected by SARS-CoV-2[172,173]. Another prior paper from Yi et al.[168] also discovered this APOBEC1 like pattern, although they were uncertain of its meaning given that the liver and small intestine are not the primary site of infection for the virus. However, several studies have reported that SARS-CoV-2 does infect both the liver[174] and the small intestine[175–177] . Live virus has also been isolated in at least 2 cases[178,179] from faecal samples, suggesting a possible faecal-oral transmission route. Infectivity of the virus was markedly reduced in an experiment where virus was added to human derived stool samples[180], and it is unclear how much virus derived from these samples would comprise an infectious unit. As such, it seems further investigation into the potential for these transmission chains is necessary, particularly given the highly mutated "cryptic lineages" observed during wastewater sampling[92].

3' CpG nucleotide contexts are the most targeted, in particular TCG, CCG and ACG. CpG suppression is a well-known dinucleotide bias. In RNA viruses, this appears to be a result of selective pressures exerted from the presence of host CpG sensing molecules such as Zinc-finger Antiviral Protein (ZAP). ZAP relies

on host CpG suppression to allow it to specifically target non-host genomic material (such as viral RNA) with higher CpG content[181]. This allows viruses with lower CpG content to better evade restriction by ZAP since it more closely resembles the host CpG composition. While ZAP does not induce C→T changes, it may help explain why C→T sites in a CpG 3' context are preferentially edited relative to other 3' contexts. ZAP has been shown to restrict SARS-CoV-2 despite pre-existing CpG depletion[101]. ZAP isoforms have been shown to prevent necessary translational frame-shifting for SARS-CoV-2 ORF1b protein production[182]. The non-normalised form of Signature 1 (Figure 2.9) shows that when tri-nucleotide bias is not accounted for 3' CpG's are lower than the normalised signatures, yet 5' TpC and ApC contexts remain the most prevalent(Figure 2.9). The most targeted contexts do shift to ACA, ACT and TCT, likely reflecting their comparatively high abundance within the SARS-CoV-2 genome relative to 3'CpG contexts. These non-normalised contexts are consistent with what was earlier reported by Simmonds et al.[1]).

Signature 2 (Figure 2.8 and Figure 2.9) has a nearly identical proportion of A→G and T→C mutations. These are a known target of the ADAR family of adenine deaminases. ADAR enzymes typically operate on double-stranded RNA and convert adenine into inosine. Inosine forms base pairs with cytosine, which after another round of replication causes guanine to replace the inosine and complete the A→G change. As ADAR operates on both strands of dsRNA, the mutational signature resulting from the process is expected to contain an equal proportion of A→G and T→C mutations, which is the case for Signature 2. Signature 2 also contains a number of G→A mutations, which could be caused by low-level C→T activity on the negative sense RNA strand. Due to

the cellular strand biases present between the positive and negative sense RNA[167], C→T mutational processes acting on ssRNA are much less likely to produce a mutation on the negative strand (resulting in G→A substitutions) than C→T changes on the positive strand. The negative strand will only be present during the replication phase of the virus while the positive strand will be present both on cell entry and on exit as the new viral particles are packaged to infect further cells. This could explain why the negative sense Signature 1 changes are present in Signature 2, since it may be operating at a similar level to Signature 2 on the negative strand. The non-normalised form of Signature 2 (Figure 2.9) does have different targeted contexts, just as with Signature 1. However, the main attribute of Signature 2 is its equal contributions of A→G and T→C substitutions, which still remain equal. Signature 3 (Figure 2.8 and Figure 2.9) is dominated by G→T substitutions. A putative mechanism for this is Reactive Oxygen Species (ROS) in the cell. Increases in oxidative stress as part of a ROS 'burst' have been associated with viruses during the early stages of infection [166,183]. Guanine nucleotides are known to be vulnerable to oxidation, with the product 7,8-dihydro-8-oxo2'-deoxyguanine (oxoguanine) pairing with adenine bases rather than cytosine[183,184]. Similar to inosine causing A→G changes, this change to oxoguanine will result in a G→T mutation after a replication cycle. The lack of C→A changes in the signature also suggests that the mechanism is most active on the positive single-stranded RNA rather than the negative single-stranded RNA. The initial positive single-stranded RNA is found in the cytoplasm, meaning it can be easily accessed by ROS and other mechanisms of mutation. Viral replication is thought to take place within membrane-bound

environments that aim to protect the RNA. The presence of double-stranded RNA within these environments strongly suggests that this is the case[83] and may explain the relative lack of negative strand mutations in SARS-CoV-2 signatures.

The non-normalised G→T signature (Figure 2.9) seems to display a context preference of TpG and ApG nucleotides, although this contextual bias is changed to CpG and ApG following normalisation. These contextual biases mean that the signature could be some other as yet unknown editing mechanism on the viral RNA, although normalisation changing this context so heavily suggests that this bias perhaps has more to do with genome composition. The increased CpG context shift post-normalisation could also be another ZAP-induced effect, where CpG depletion is selected for to help the virus evade ZAP. Curiously, this G→T bias has been observed in other coronaviruses, but not widely among RNA viruses[185]. ROS has a verified cancer mutational signature [130,186] although the context preferences do not match the signatures (normalised or non-normalised) observed here. However, there are a multitude of differences between viral RNA and human DNA that make these signatures difficult to compare.

It is important to note that while SARS-CoV-2 does have an error correction mechanism resulting in fewer replicase-induced errors, this mechanism will not catch all changes. A number of the mutations picked up from the set of sequences (and included in our mutational signatures) will be derived from replication errors. However, the clear and repeatable extraction of the signatures indicates that despite this potential contamination, the extracted signatures do appear to be predominantly other mutational processes. While a

replication error associated mutational signature may be identified in future, this signature is too diffuse to identify as a distinct process. Similarly, a high proportion of mutations are not accounted for by the extracted mutational signatures. These mutations were not present in large enough quantities to enable effective extraction from the data. Future methods may be able to tease out the more subtle mutational mechanisms that almost certainly exist to induce these less common mutation types.

Signature activities clearly change in both the global dataset and in the various subsets of the data for VOCs and continents. In the global data (Figure 2.10) Signature 1 is dominant throughout the pandemic. Signature 2 only begins to appear around November 2020, after which it appears consistently active for the remainder of the pandemic. This is approximately when variant of concern lineages began to emerge, as well as the beginning of the first vaccine rollouts. This is particularly apparent in the Alpha subset where Signature 2 is the most highly active mutational process (Figure 2.12), with a large depletion of Signature 1 activity as well.

Alpha was shown to increase sub-genomic RNA expression of several immune-antagonist viral proteins including nucleocapsid (N), ORF9b and ORF6[77,79,187,188]. N is thought to shield dsRNA from detection by RNA sensors, which trigger downstream antiviral response pathways [187–190]. ORF9b antagonises TOM70, a protein required for the activation of mitochondrial antiviral-signalling proteins (MAVS)[187] while ORF6 inhibits the transportation to the nucleus of inflammatory transcription factors[191]. Combined, the cumulative immune inhibition may have resulted in an observable change in the mutational processes that we observe within the Alpha lineage. Beta and

Gamma (both VOCs that emerged around the same time as Alpha) gained amino acid substitutions that helped evade the immune system primarily via antigenic change. Alpha's reliance on attenuating immune pathways rather than antibody binding may be why we see a different signature exposure pattern in this VOC relative to the others. This could be due to the attenuated pathways being involved in signalling for the mutational processes behind Signatures 1 and 3, while not inhibiting Signature 2 as much.

This Alpha pattern is not observed in the other VOC datasets, although Delta and Omicron have a high level of Signature 2 exposure as well, despite Signature 1 remaining the dominant process in those subsets. Signature 3 appears to be most prominently found in the Delta subset and remains consistently at low levels in the global data until January 2022 when it appears to disappear almost entirely. The Omicron subset has little to no exposure for Signature 3 and this happens to be the VOC almost exclusively circulating after January 2022. Why Omicron appears to have so little Signature 3 exposure is unclear, although unlike previous VOCs, Omicron differs in its preference of cell entry mechanism. Previous variants of the virus typically enter the cell using membrane fusion, where the viral membrane fuses with the cell membrane via the action of ACE-2 receptor binding and TMPRSS2 cleavage of the spike protein. Omicron instead favours an endosomal route of entry whereby the viral particle binds to the cell using ACE-2 and is enveloped by endocytosis into the cell. Cleavage of the spike protein then occurs via the action of Cathepsin L, which allows for the release of the viral RNA into the cytoplasm of the now-infected cell[54,192].

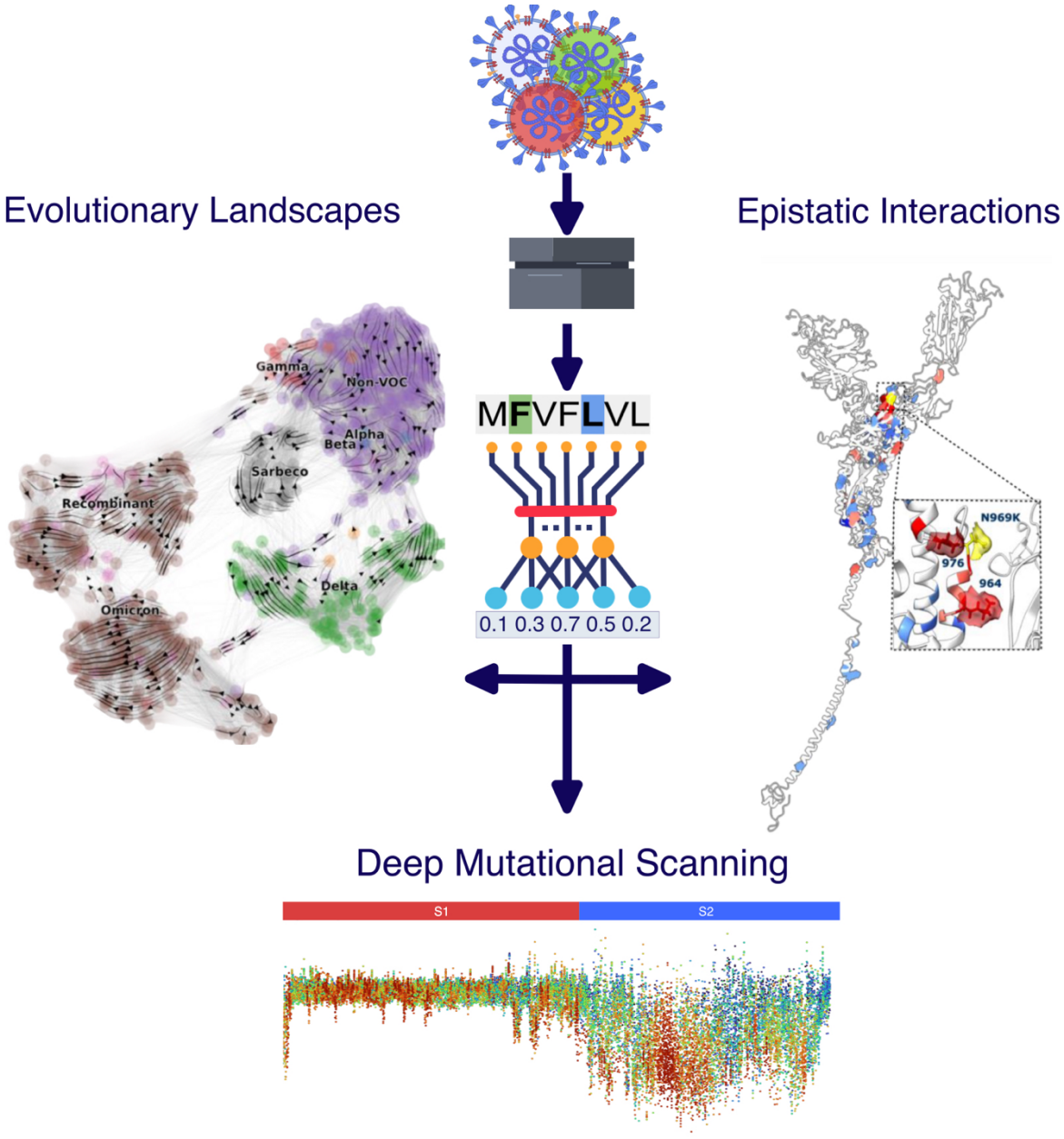
Signature transitions from Signature 1 to Signature 2 changes occur from December 2020 onwards in the global dataset and appears consistently in the VOC and continent-defined subsets around this time point as well. Alpha underwent a major shift to Signature 2 mutations early in its time as a VOC, although Signature 1 returned as the predominant set of changes towards the end of its wave of infections. The non-VOC subset appears to be the least impacted by Signature 2 changes. However, this can mostly be explained by the number of non-VOC sequences quickly declining after the emergence of the VOC lineages. Delta underwent a dramatic increase in Signature 2 and Signature 3 exposure from July 2021, with Signature 2 becoming the predominant signature towards the end of Deltas wave. Signature 2 changes continue into Omicrons introduction, although it does decrease after the initial Omicron(BA.1) wave from December 2021 to March 2022. It seems clear that while Signature 1 mutations have dominated in contributing to the evolutionary capacity of SARS-CoV-2 throughout the pandemic, this mutational environment is beginning to change. Such shifts in mutational processes are potentially evidence of changing interactions between the viruses and the immune systems of the hosts they circulate within. For example, changes in population-level immunity via vaccination or previous infections may influence the mutations that we observe in the data. Changing mutational process activity in consensus sequences from infections is unlikely to fully reflect the true activity of each process, but they are likely to show which processes are contributing mutations that eventually make it into circulating viruses.

All variants of concern we assessed show predominantly non-synonymous mutations and all mutational signatures are associated with more non-synonymous than synonymous changes. More synonymous substitutions in the lineage references were found in ORF1a/b, which is expected due to it being the longest ORF. However, this pattern is not observed with non-synonymous mutations as these are mainly located in the spike protein (Figure 2.13C and Figure 2.13D). This is consistent with spike being under intense immune pressure since it is the main glycoprotein for SARS-CoV-2. As such, spike must change in order to escape the host immune response, while maintaining its main function of binding and entry into host cells. Signature 1 changes are the predominant source of mutations in all SARS-CoV-2 VOCs that we analysed, followed by unattributed mutations, Signature 2 changes and Signature 3 changes. Signature 3 changes were unlikely to be synonymous mutations with only Beta, Gamma and Delta containing very few such changes (Figure 2.13D). This is also reflected in the global synonymous/nonsynonymous exposures where Signature 3 appears completely inactive in the synonymous mutation subset (Figure 2.13A). Signature 2 exposure appears the most likely to be synonymous mutations (Figure 2.13A) but this does not seem to be observed in the VOC lineages where most Signature 2 changes are non-synonymous mutations (Figure 2.13B).

In conclusion, mutational signature analysis reveals important processes contributing to SARS-CoV-2 genetic variation and serves as a tool to track the dominant changes over time and to generate hypotheses about the main mechanistic processes in play. Specifically, host antiviral molecules as opposed to replication errors appear to be the main generator of mutations (confirming

earlier computational studies), a result that requires experimental confirmation. Despite limitations in potential biases, our findings contribute to a better understanding of the complex dynamics driving the evolution of SARS-CoV-2 and the emergence of VOCs.

3 LARGE LANGUAGE MODELS CHARACTERISE THE PROTEINS OF SARS-COV-2



"With great power there must also come—great responsibility"

Cf. S. Lee and S. Ditko, Amazing Fantasy No. 15: "Spider- Man" (1962)

3.1 Abstract

Protein language models (PLMs) like Evolutionary Scale Model 2 (ESM-2) use millions of protein sequences to learn the properties of amino acid sequences. ESMFold can now fold protein sequences into their 3D structure, with no alignment or other information about the sequence necessary. The PLM derived metrics (grammaticality and semantic scores) have previously been shown relate to important structurally related features such as antigenicity. PLMs therefore appear to contain information about protein structure and potentially evolutionary constraint.

Here, we describe how ESM-2 can be used to represent meaningful information about a novel virus from sequences at various stages of an outbreak. We show using the SARS-CoV-2 pandemic how these models could have been applied; both in the early stages before experimental observations and in later stages for monitoring and horizon scanning. Using PLM enabled in-silico deep mutational scanning (DMS) we demonstrate that PLMs describe fundamental properties of viral proteins. In combination with traditional metrics from protein structure and evolutionary contexts, we show that the model understands changes to the protein surface, protein stability and evolutionary considerations. We show that PLMs can differentiate between phenotypically different viral protein sequences and even demonstrate their ability to identify mutational epistasis. These model outputs can supplement the available information, and we make a case for their future application and use in outbreaks or pandemics to come.

3.2 INTRODUCTION

Natural language processing (NLP) has been applied not just to written language but also to biological sequences like DNA and proteins[100,118,193]. NLP is a field that has long tried to model the complexities of human language using machine learning and statistical methods. Much of this effort has coalesced around the distributional hypothesis; that the meaning of words can be largely determined by the context they keep[194]. Several different methods have shown that this hypothesis shows promise[114,117]. Biological sequences represent physical molecules which are represented by arbitrary characters. This is convenient since these characters, each representing an amino acid, can be processed in the same way as words in natural language. Hie. et al[193] show how, much like natural language, biological sequence context can be used to create informative embeddings that reflect complex properties of the sequence, and[193] show how concepts like grammar and semantics that are used to assess changes in natural language can be applied to changes in protein amino acid sequences.

Changes to a protein sequence, amino acid substitutions or insertions/deletions, can alter its structure enough to change functional and antigenic properties. This can be thought of as changing the meaning (semantics) of the protein and can be estimated by measuring distances between proteins in the embedding space. A grammatical sequence in natural language represents a sequence where the rules of the language are followed. In the model, this is captured by the probability of the protein sequence. A more likely sequence includes contexts that are often seen together in the set of known proteins. Contexts not seen are assumed to be absent because they break the rules of the protein "language". Hie et al[193] implemented these

concepts of protein grammaticality and semantics in combination to predict escape mutations in various viral proteins.

While it is commonly assumed that impactful missense mutations (non-synonymous substitutions resulting in amino acid changes) in human genome data primarily leads to disease-associated changes, the alterations in amino acids observed in virus data (as is also the case for human data) are often associated with adaptive evolution[77,148]. Alternatively, these changes can be deleterious and may result in misfolding and/or decrease fitness or be "neutral" and have little to no consequence on viral fitness. Approaches for assessing these different impacts are premised on comparative genome sequence data being available to assess features like the relative proportions of non-synonymous to synonymous substitutions (dN/dS) or evolutionary conservation at individual sites. Both require sequence alignments. Similarly, genome sequence-based surveillance methods often require comparative sequence data to assess the relative growth rate of a pathogen lineage[195,196] ideally detected through real-time sequencing. While experiments can be focussed on an unusual variant when first detected, these can take months to complete. Protein-based assessment, e.g., disruption of stability/folding energy, can be used to put mutations in their three-dimensional context, however, these only assess one aspect of changes in a linear sequence that impact proteins. The emergence of SARS-CoV-2 and its subsequent variants of concern (VOCs) Alpha, Beta, Gamma, Delta and Omicron caused large waves of infections during the pandemic. Predicting their evolutionary advantage just from their first available sequences, before the viruses were in widespread circulation, has been difficult to impossible.

Language-based models could be used to meet this challenge of rapid characterisation of individual pathogen genomes. With the currently available models unaware of the COVID-19 pandemic, this represents an opportunity to test how these LLMs can be used on a novel virus.

Evolutionary Scale Model (ESM) 2[100,118] is an example of a protein language model (PLM). It is trained on approximately 65 million unique protein sequences gathered by sampling across 43 million UniRef50 clusters and 138 million UniRef90 sequences. Sequences are made of amino acid characters, with each possible amino acid represented as a unique token in the model. The largest instance of the model contains 15 billion parameters and marks an enormous leap in size compared with the previous largest model ESM-1b with 650 million parameters. ESM is trained using the Masked Language Model (MLM) objective whereby the model is tasked with predicting the amino acid present at a given masked-out position of the protein sequence. This objective allows the model to learn the contextual requirements of a given amino acid, i.e. given a sequence, what's the likely amino acid at a position. Similar transformer-based language models such as BERT have used this objective to produce meaningful textual embeddings that capture complex features of sentences. Earlier iterations of ESM learned to identify which amino acids were in contact, but advances in the model architecture allow ESM-2 to predict full 3D structures of proteins from sequence alone.

Here we demonstrate how ESM-2 can be used with individual SARS-CoV-2 spike protein sequences to identify variants of interest. We show how language models can provide useful information and actionable information about sequences. We show that metrics like the semantic score and grammaticality

can reveal characteristic properties of the spike protein sequence and assess how they can be used in the context of variant and sequence horizon scanning. In combination with traditional metrics from protein structure and evolutionary contexts, we show that the model understands changes to the protein surface, protein stability and evolutionary considerations. We show that PLMs can differentiate between phenotypically different viral protein sequences and even demonstrate their ability to identify mutational epistasis. These model outputs can supplement the available information, and we make a case for their future application and use in outbreaks or pandemics to come.

3.2.1 AN INTRODUCTION TO THE GRAMMATICALITY AND SEMANTIC SCORES

The use of language models to interpret biological sequences like nucleic acids and proteins has meant the use of natural language terms to describe characteristics of these sequences has become more common. Hie et al.[193] make big strides in this area, equating the likelihood of mutations in biological sequences to a measure of the correctness of grammar in a natural language sequence. They also equate increasing distances between words in a language model embedding (often used as a measure of word meaning/semantics as made famous by word2vec[114]) to be equivalent to a change in the antigenic properties of the sequence. They demonstrate rather convincingly that by training models on the key immunogenic proteins of three viruses (SARS-CoV-2, HIV and Influenza), this grammaticality appears to represent protein fitness (i.e. its ability to exist as a folded structure) while the semantic score identifies possible antigenic change. The combination of these two scores together can be used to conduct constrained semantic change search (CSCS),

where mutations are identified with higher antigenic properties that are also sufficiently "fit" such that the protein should still function.

Much of this chapter is focussed on the discussion of these methods, except we use a new different model to calculate the values. We use ESM-2[100], a protein language model that is trained on many different proteins across the domains of life rather than a specifically trained model of the SARS-CoV-2 spike. This means that we cannot make the same inferences that Hie et al.[193] make, since it is not clear that distances between ESM-2 embeddings have the same meaning (i.e. antigenicity) as the Hie et al.[193] model appears to represent, since not all of the proteins used to train ESM-2 are immunogenic glycoproteins[100]. In principle, there are no bounds to either score, since log likelihoods range from between 0 and -infinity, while the semantic score is simply a distance. However, for the purposes for discussion below, since we are dealing with single position mutations and our grammaticality scores are ratios between the reference and the mutated amino acid, semantic scores range from between 0 to 13, while relative grammaticalities range from -22 to 4. For relative grammaticalities, a positive value indicates that the mutated position is more likely than the reference position according to the model. When looking at score for full sequences (rather than single mutations), scores range from 0 to -237 for relative grammaticality, and 0 to 9 for the semantic score.

3.3 METHODS

3.3.1 ESM-2 AND EVO-VELOCITY

SARS-CoV-2 spike proteins were acquired by filtering the GISAID database (available from available from https://doi.org/10.55876/gis8.240620pm) for the earliest sequences from each PANGO lineage with a fully intact spike protein sequence. The sequence was then embedded in the ESM-2 model to produce an embedding for the sequence. For the DMS data and for the embedding scores, the ESM-2 3 billion parameter variant was used. The semantic score is equivalent to the L1(Manhattan) distance between the embedding of the reference sequence (the spike protein from the original Wuhan SARS-CoV-2 genome) and each of the PANGO spike proteins. The grammaticality of a sequence is calculated as the product of the probabilities of each amino acid at each position in the spike protein. The probabilities come from a softmax of the last layer of the embedding and range between 0 and 1. However, many of the probabilities are small and for numerical stability the probabilities are represented in the log space. The relative grammaticality is the same as the grammaticality, except the probability of the reference is subtracted from the probability of the mutant sequence so that the score is relative to the original SARS-CoV-2 sequence. The sequence grammaticality represents the summed log-likelihoods of every reference position in the sequence, rather than just the mutated positions. For the deep mutational scan, a sequence was produced for every potential amino acid at every position in the spike protein. Each sequence was then embedded to calculate semantic and relative grammaticalities for every mutation. We then used the evo-velocity package[119] to embed the initial sequences using the ESM-2 650M parameter variant, and then performed velocity analysis and spearman's rank for the sample dates

against the pseudo-time. Evo-velocity uses the probabilities from the language model to produce a vector field of transition states between observed sequences. This field maps the transitions between sequences and determines an ordering between the different sequence states using a diffusion process through the graph. This identifies a root sequence that determines the directionality of the evolutionary trajectory, which is the summarised by a UMAP plot, with arrows showing the general direction of "evolutionary" pseudotime through the sequences. The idea is to identify probable mutational trajectories using just sequence probabilities. The 650 million parameter model was used primarily due to hardware limitations of using the larger model (the required GPU memory was not available), however the smaller model performs similarly with regards to structural prediction as shown by Lin et al.[100]. However, Lin et al.[100] show that additional parameters can be better in certain circumstances which was why the 3B parameter model was used where possible. The 15B model outperformed all of the smaller models in their structural tasks, however the results were often again comparable suggesting certain proteins require the additional parameters to make good predictions.

3.3.2 EPISTASIS EXPERIMENTS

The epistasis experiments used an Omicron(BA.1) SARS-CoV-2 spike protein and embedding this using the ESM-2 3 billion parameter variant. Due to the 3 amino acid insertion "EPA", the logits (probability values produced by the model) for this position were subsequently removed to map logits to the 3D structure. The Omicron(BA.1) sequence contains several mutations relative to the Wuhan-Hu-1 SARS-CoV-2 reference spike sequence. Each of these

mutations was reverted one by one back to the reference position, and the likelihood differences upon reversion were recorded. To eliminate noise, changes less than 2 standard deviations from the mean across all reverted mutations were removed and deemed not significant.

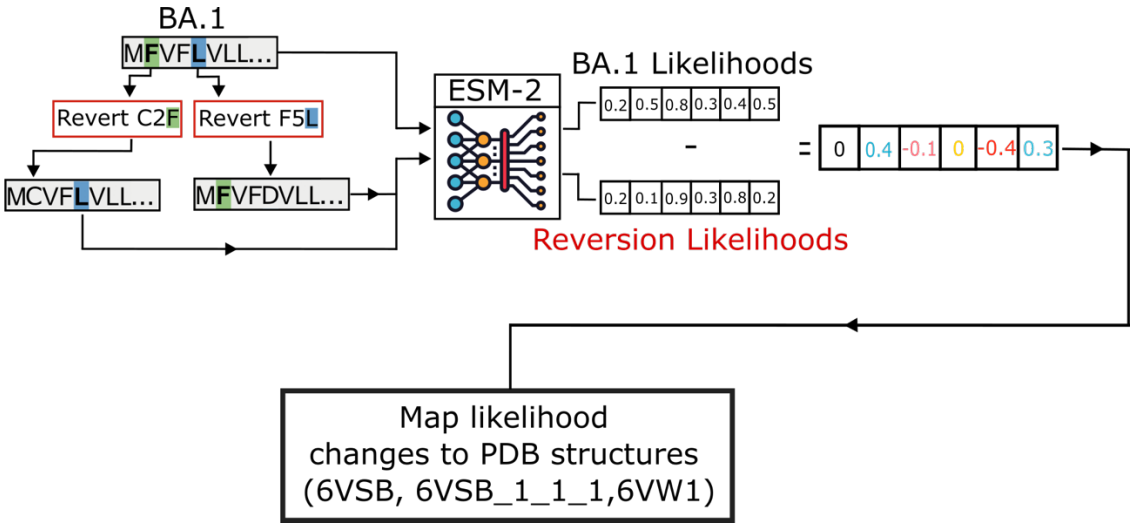


Figure 3.1: Identifying epistatic interactions involves reverting the mutations from a SARS-CoV-2 variant (here Omicron(BA.1)) and measuring the effect this has on the other likelihoods. This is achieved by subtracting the reference likelihoods from the mutant likelihoods (we show 2 example mutation sites C2F and F5L), before mapping onto a PDB structure of the Spike protein (6VSB, 6VSB_1_1_1 and 6VW1 were all used in this chapter).

3.3.3 DYNAMIC EMBEDDINGS AND HORIZON SCANNING

Dynamic embeddings were computed by first gathering UK GISAID data (available at https://doi.org/10.55876/gis8.240621ma) and clustering sequences into haplotypes with 99.9% similarity. A haplotype here is a representative sequence from a cluster where other cluster members are at least 99.9% similar. These haplotype clusters produced 11272 sample date labelled haplotypes with a sequence returned for each cluster. Each haplotype

embedding was measured against a mean of the embeddings from the prior 3-month period using the L1 distance (i.e. the semantic score). For sequence grammaticalities, mean sequence grammaticalities were calculated in a similar way and differences measured.

3.3.4 ASSESSING EMBEDDINGS SCORES WITH KNOWN METRICS

The language model's metrics were first assessed using a spearman's rank against several known biological scores. Next, Support Vector Regression (SVR)[197] was used as a simple model to fit the model scores as well as the embeddings and logits to the biologically relevant metrics. Models were fit to the data using 5-fold cross validation, and a linear kernel for the SVR. Model results were reported as the average spearman's rank between the folds, with the error bars as +/-1 standard deviation from the mean (Figure 3.10).

3.3.5 SELECTION ANALYSIS SIGNALS

Signals of ancestral evolutionary selection in the animal (bat and pangolin) Sarbecovirus most closely related to SARS-CoV-2 referred to as the "nCoV" clade (defined in Lytras et al.[198]) were inferred on a set of 167 Sarbecovirus genomes, accounting for recombination by inferring selection separately in each non-recombinant segment. These results are published in Martin et al.[199] and presented in more detail in the following Observable notebook: https://observablehq.com/@spond/ncos-evolution-nov-2021. Briefly, sites under negative selection were inferred using the Fixed Effects Likelihood (FEL) and sites under positive selection using MEME[200] by testing on internal branches of the nCoV clade. FEL calculates the per-site synonymous and non-synonymous substitution rates[201]. Sites denoted as conserved have the same amino-acid residue among all Sarbecovirus sequences in the analysis. In

addition, the evolutionary "flexibility" of the site in the SARS-CoV-2 sequence was obtained as the entropy of the predicted distribution of credible evolutionary states.

3.3.6 Antibody accessibility, spike protein stability and Deep Mutational Scanning

Structure-based epitope score, referred to as Accessibility, which approximates antibody accessibility for each spike protein amino acid position, was calculated using BEpro software[202] for a Woo *et al.* model of the Wuhan-Hu-1 spike protein sequence. Scores relating to substitution probabilities, namely, Environment Specific Substitution Table (ESST)[203] probability, Log Position Specific Scoring Matrices (PSSM) and predicted ΔΔG, were obtained for every possible single amino acid substitution for the 6VXX SARS-CoV-2 spike structure (note that only values for Chain A are included in the results as data is generally identical across all three chains). ESST probability values were calculated using the Environment Specific Substitution Table[203] after local structural environments were calculated by JOY[204]. ESST substitution tables aim to incorporate the local structural environment of amino substitutions, taking into account the secondary structure, accessibility by solvents and hydrogen bonding between adjacent amino acids[205]. Log Likelihood substitution values were calculated using PSSM with the DELTA-BLAST[206] algorithm in BlastX. The PSSM is a substitution matrix constructed by querying a database of conserved domains using the input sequence (the Wuhan-Hu-1 spike protein), aligning the domains and computing the substitution probabilities for each position using the alignment[206]. It should be noted that this is a sequence-based method, so residue numbering does not match the numbering in 6VXX and values are available for residues not

described by the 6VXX PDB file. ΔΔG values were predicted by FoldX5[207] software. FoldX uses empirical energy functions to predict the energetic effect of mutations to protein stability. The predicted ΔΔG quantifies the change in the free energy of unfolding between the wild-type and mutated structure. The 6VXX structure was first repaired with the *RepairPDB* function to fix residues with bad torsion angles, van der Walls' clashes or total energy. Substitutions were then performed using the repaired structures on all three chains simultaneously using the *BuildModel* function, giving the change in free energy of unfolding, with negative values implying stabilising mutations.

3.3.7 DEEP MUTATIONAL SCANNING DATA

The receptor binding deep mutational scanning data was taken from experimental DMS studies of the SARS-CoV-2 from Yisimayi *et al.* (2024) and Starr *et al.* (2022)[208,209]. The Wuhan-Hu-1 scores were taken as is, while the variant average score was calculated by averaging the scores for each position between each of the SARS-CoV-2 variant specific DMS results[208] (Wuhan-Hu-1, Alpha, Beta, Delta, Eta). For the RBD mutational escape values we utilised the high-throughput mutation antibody escape profiling results presented in Yisimayi *et al.* (2024)[209] This study used a panel of 1,350 monoclonal antibodies against all possible RBD substitutions. The backbone virus used was the SARS-CoV-2 BA.5 variant instead of Wuhan-Hu-1, however this should still provide the most comprehensive dataset of unique substitutions' effect on antibody escape. Our mutational escape metric is the average of the raw antibody escape values for each substitution on each site of the RBD across all tested monoclonal antibodies. The full spike DMS data was taken from Dadonaite *et al.* (2023)[210]

and DMS scores (Binding, Escape and Entry) are from a BA.2 and XBB.1.5 spike backbone and averaged together to give the presented score[210].

3.4 Results

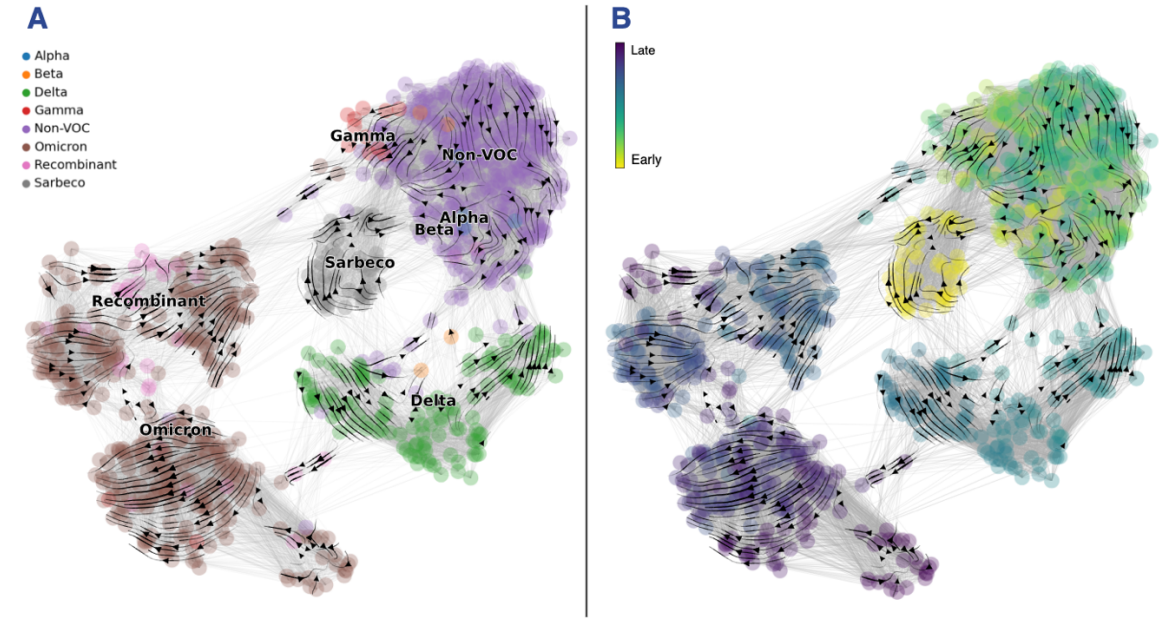
3.4.1 Language models capture the mutational landscape of SARS-CoV-2

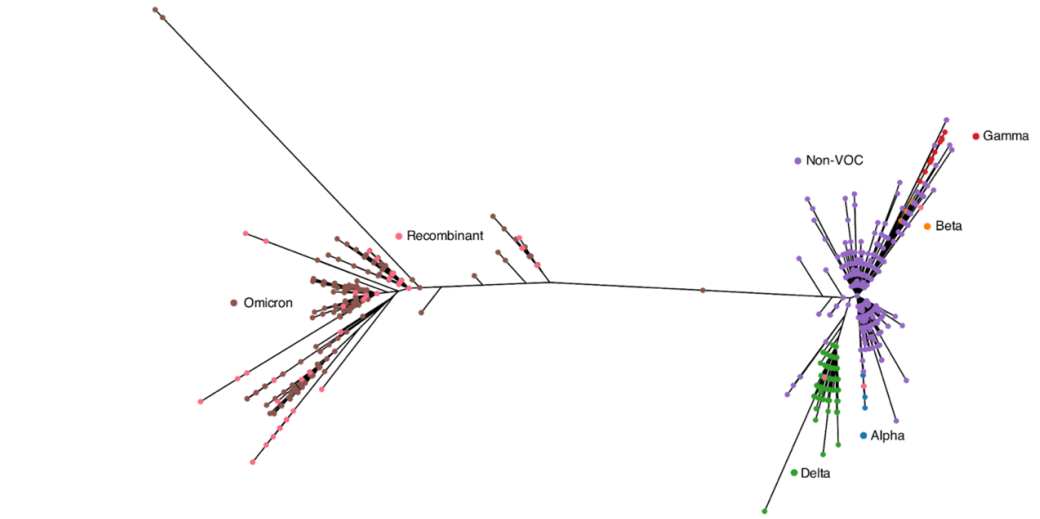
The earliest known SARS-CoV-2 sequences for each of the PANGO lineages were extracted from the global SARS-CoV-2 data (retrieved from GISAID) and their spike protein sequences embedded using ESM-2[100,118]. The evo-velocity package was then used to produce a UMAP and evo-velocity plot for the sequence embeddings. Evo-velocity assigns a putative directionality between the embeddings that describes the flow of evolution through the UMAP embedding space. Firstly, evo-velocity[119] shows that the language model embeddings capture information that can distinguish between meaningfully different spike proteins (Figure 3.2). Secondly, it illustrates that embeddings can be used to understand the evolutionary landscape of the protein and the directionality of its evolution.

The evo-velocity accurately describes the evolution of SARS-CoV-2 moving out from the non-VOC clusters into VOC clusters as was shown during the pandemic (Figure 3.2A and B). Omicron and Delta in particular form distinct clusters separate from the non-VOC sequences and the earlier VOCs, while Gamma forms a concentrated cluster on the fringe of the non-VOC cluster. On the other hand, Beta and Alpha are less homogenous. The Sarbecovirus sequences (which include bat viruses related to SARS-CoV-2) form another cluster separate from the SARS-CoV-2 sequences. Interestingly, they are correctly identified as the earliest sequences **(**Figure 3.2B and Figure 3.3**)** and

are placed close to the early non-VOC SARS-CoV-2 sequences. Recombinant lineages, i.e., spike proteins that have been produced by combining nucleotide sequences from two distinct SARS-CoV-2 genomes, have occurred between two or more VOCs. Many of these recombinants occur between two or more Omicron VOC sequences, however, several sequences have recombined between different VOCs i.e., generating an Omicron and a Delta recombinant. The nucleotide sequence information used in the phylogeny (Figure 3.2C) is "higher resolution" than the protein sequence due to redundancy in the genetic code that allows for multiple nucleotide codons to encode the same amino acid. The additional information provided at the nucleotide level can be necessary for identifying the correct placement of very similar sequences, particularly when at the amino acid level they are identical. SARS-CoV-2 spike sequences are a good example of this, where most sequences are only a few amino acids changes apart, and convergent mutations have happened repeatedly. The velocity recapitulates the topology of the representative sequence phylogeny in Figure 3.2C, with early VOCs more closely related to non-VOC sequences than Omicrons and recombinants. This supports that the protein sequence embeddings do capture meaningful representations that can identify differences between distinct lineages containing many mutations. The tree provides further confirmation that ESM-2 evo-velocity recapitulates the evolutionary history of spike, due to this matching of tree topology and evo-velocity determined UMAP structure. The package uses diffusion analysis through the UMAP nearest neighbours' network to identify root and endpoint sequences, before running a pseudotime simulation to determine the order of evolution between the two points (Figure 3.2B and Figure 3.3). When diffusion

analysis is used to select the root and endpoint sequences, pseudotime achieves a spearman's correlation of 0.86 against sampling time with a p-value of 3.24e-296, confirming that the model has inferred the evolution of the sequences in the correct order. We can infer from this that ESM-2 embeddings can reproduce the evolutionary landscape of a protein without fine-tuning or prior knowledge of the specific protein.


A
Alpha
Beta
Delta
Gamma
Non-VOC
Omicron
Recombinant
Sarbeco
Gamma
Non-VOC
Alpha
Beta
Sarbeco
Recombinant
Delta
Omicron
B
Late
Early


C
Recombinant
Omicron
Non-VOC
Gamma
Beta
Alpha
Delta

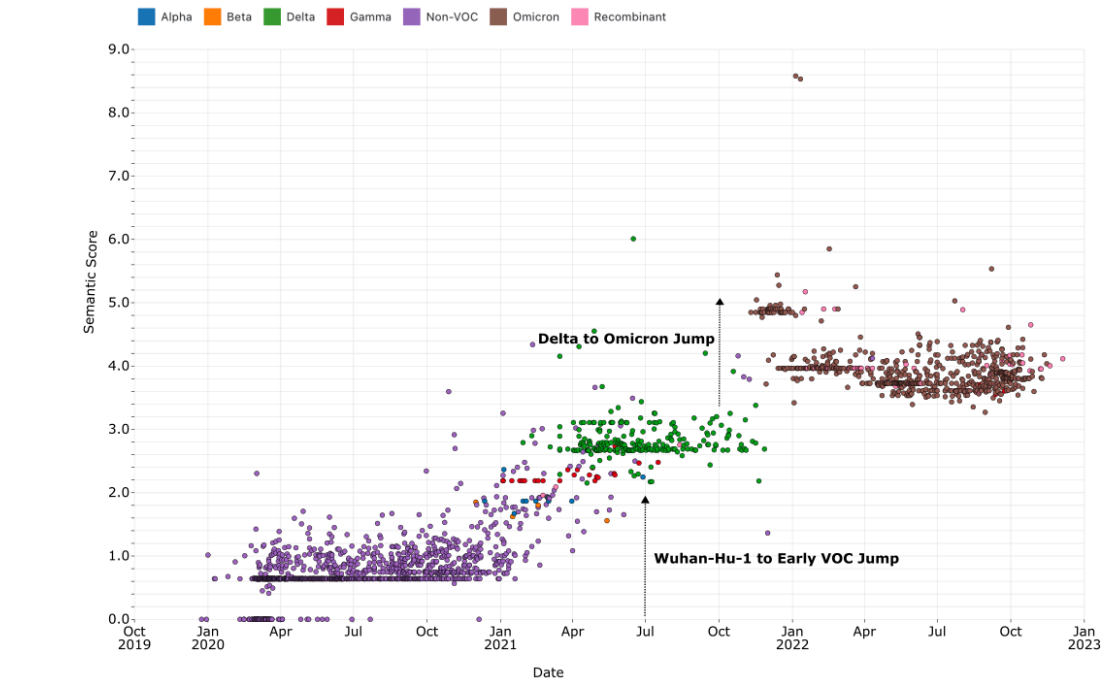

D
Alpha
Beta
Delta
Gamma
Non-VOC
Omicron
Recombinant
Semantic Score
Delta to Omicron Jump
Wuhan-Hu-1 to Early VOC Jump
Date
Oct 2019
Jan 2020
Apr
Jul
Oct
Jan 2021
Apr
Jul
Oct
Jan 2022
Apr
Jul
Oct
Jan 2023

Figure 3.2: (A) UMAP of initial spike sequence embeddings for SARS-CoV-2 PANGO lineages and a selection of other known Sarbecovirus spike sequences. Each lineage is represented by 1 spike embedding. Points are coloured on VOC classification. Arrows represent the evo-velocity through the embedding space, which shows a "directionality" of evolution. (B) shows the sequences coloured by pseudotime inferred using sequence embedding probabilities to order sequences in time using an inferred root and an endpoint. (C) Shows an unrooted nucleotide phylogenetic tree of the spike sequences, coloured again by VOC. (D) shows the spike protein sequences plotted using their sample date and semantic score coloured consistently to Figure 1A.

Even using summary metrics (grammaticality and semantic score) derived from the full embedding, we can largely separate out variants of concern into distinct groups (Figure 3.2D). Three main clusters can be seen in Figure 3.2**,** a cluster containing the majority of non-VOC sequences, a cluster containing early VOC sequences (Alpha, Beta, Gamma and Delta), and a cluster containing Omicron sequences. Unlike the UMAP (Figure 3.2A and Figure 3.2B) which is a dimensionality reduction and projection from high-dimensional space, grammaticality and the semantic score are simple distance metrics and log-likelihoods calculated from the embedding. As such they are significantly simpler than UMAP dimensions yet still recapitulate much of the information shown. The relative grammaticality shows a decreasing trend over time and the semantic score an increasing trend.

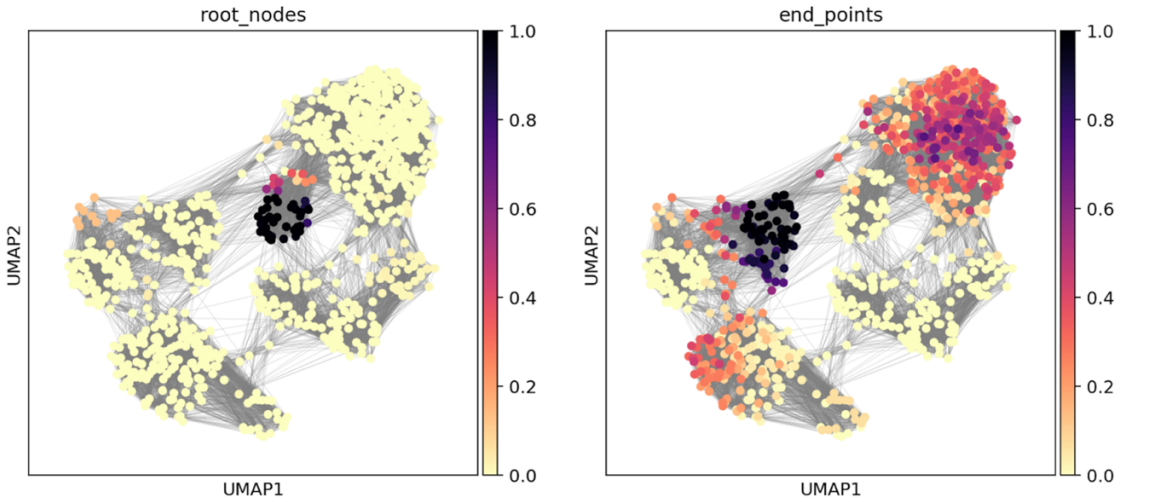


Figure 3.3: Predicted root nodes and endpoints identified by running Markov diffusion process over the weighted edges of the evo-velocity network. The root nodes are correctly identified as the Sarbecovirus spike sequences, with Omicron VOC sequences predicted as the end nodes.

The VOC lineages show a characteristic "jump" in both semantic score and grammaticality (Figure 3.2B). Alpha, Beta, Gamma, and Delta form a cluster away from the non-VOC sequences while Omicron separates into its own cluster after another step jump in semantic score and grammaticality. The earlier VOCs may have required fewer changes to compete since much of the population during these waves was naive to infection or vaccine-derived immunity. As such, small improvements to immune evasion and or intrinsic virological characteristics such as replication efficiency could confer major fitness advantages relative to other circulating lineages. Omicron meanwhile emerged after massive levels of both vaccination and infection, which may be why it required a larger jump to outcompete the Delta variant at the time. Omicron contains many more mutations in the spike protein than previous variants, it changed its entry mechanism preference and managed to largely evade previous immunity which resulted in an extended vaccination regimen

of three doses being reccomended[211,212]. Omicron sequences decrease in their semantic scores after establishing unlike the other VOC groups. This can initially be attributed to the emergence of BA.2 which had fewer mutations than the initial BA.1 wave of Omicron infections. However, despite subsequent variants increasing their number of mutations to above BA.1 levels, we see further decreases in semantic score within Omicron. This further suggests that the semantic score is not simply a proxy for mutation count and suggests that future variants may be successful without increasing their semantic scores.

3.4.2 In-silico deep mutational scan of the SARS-CoV-2 spike protein

To assess how well the model understands individual mutations, we conducted an in-silico deep mutational scan using embeddings for every possible single substitution in the spike protein. Traditional deep mutational scans involve experimentally substituting an amino acid with every other amino acid for every position in a protein[209,210,213,214]. With a language model, we can do this by embedding each of these substituted sequences using the model and then calculating grammaticalities and semantic scores for each.

ESM-2's scores correlate well with the structure of the spike (Figure 3.4). Using the DMS approach to produce relative grammaticality and semantic scores, we immediately see that the protein has 2 main regions that broadly align with the two main subunits of Spike S1 and S2 (Figure 3.4A and C). Relative grammaticalities dramatically decrease when substitutions are introduced in the S2 subunit of spike. This region is composed of alpha helices that form the core of the spike structure when the different spike monomers trimerise. As such, changes in this region could disrupt the formation of a

stable and functional spike protein by preventing or inhibiting interactions between spike monomers. On the other hand, mutations in the N-terminal Domain (NTD) and RBD regions of spike appear to have lower relative grammaticalities. Despite these regions being important to spike function, they also need to be flexible to facilitate interactions and accommodate mutations to evade host immunity. Alignment entropy (specifically Shannon entropy) is a measure of the variance at a particular site in the protein based on a multiple sequence alignment, and in Spike, the positions with higher entropy are found more frequently in the S1 rather than the S2 (Figure 3.4B and Figure 3.5). This provides further evidence that ESM-2 has correctly identified the S1 as a more variable region than the S2 and suggests that ESM-2 has learnt this property of spike, despite not being explicitly trained or fine-tuned on it.

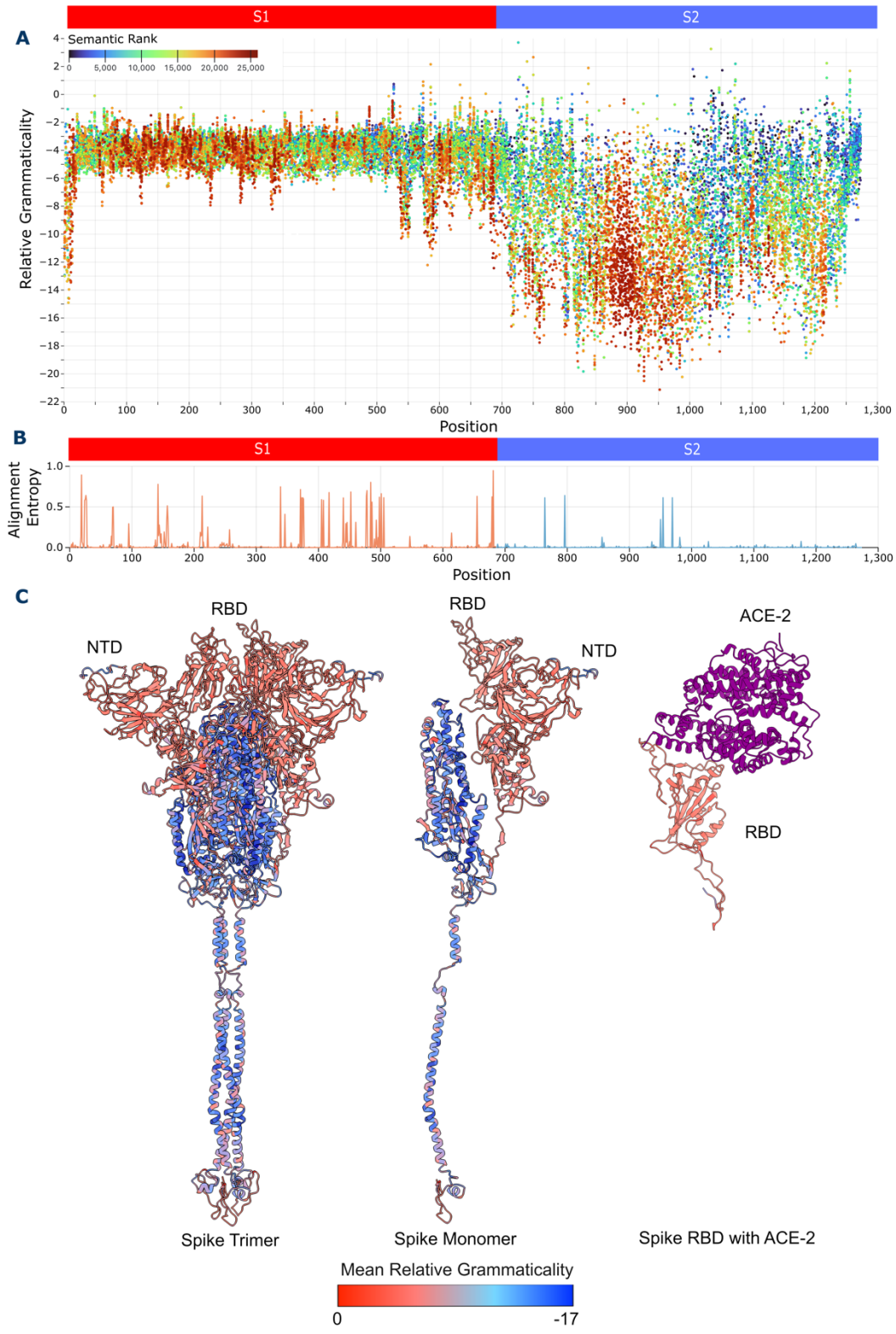


Figure 3.4: (A) Scatter plot of spike protein DMS. Relative grammaticality is shown on the y-axis, with the amino acid position on the x-axis. Points are coloured on the semantic rank of each change. (B) shows a line graph of the entropy at each position in the SARS-CoV-2 spike. S1 contains the

majority of the sites with high entropy, while S2 contains only a few. (C) Average relative grammaticality at each position on the spike protein plotted on 3 structures (6VSB_1_1_1, 6VW1)[215], the full Spike protein, the spike monomer, and the spike receptor binding domain (RBD) bound to the ACE-2 receptor in purple.

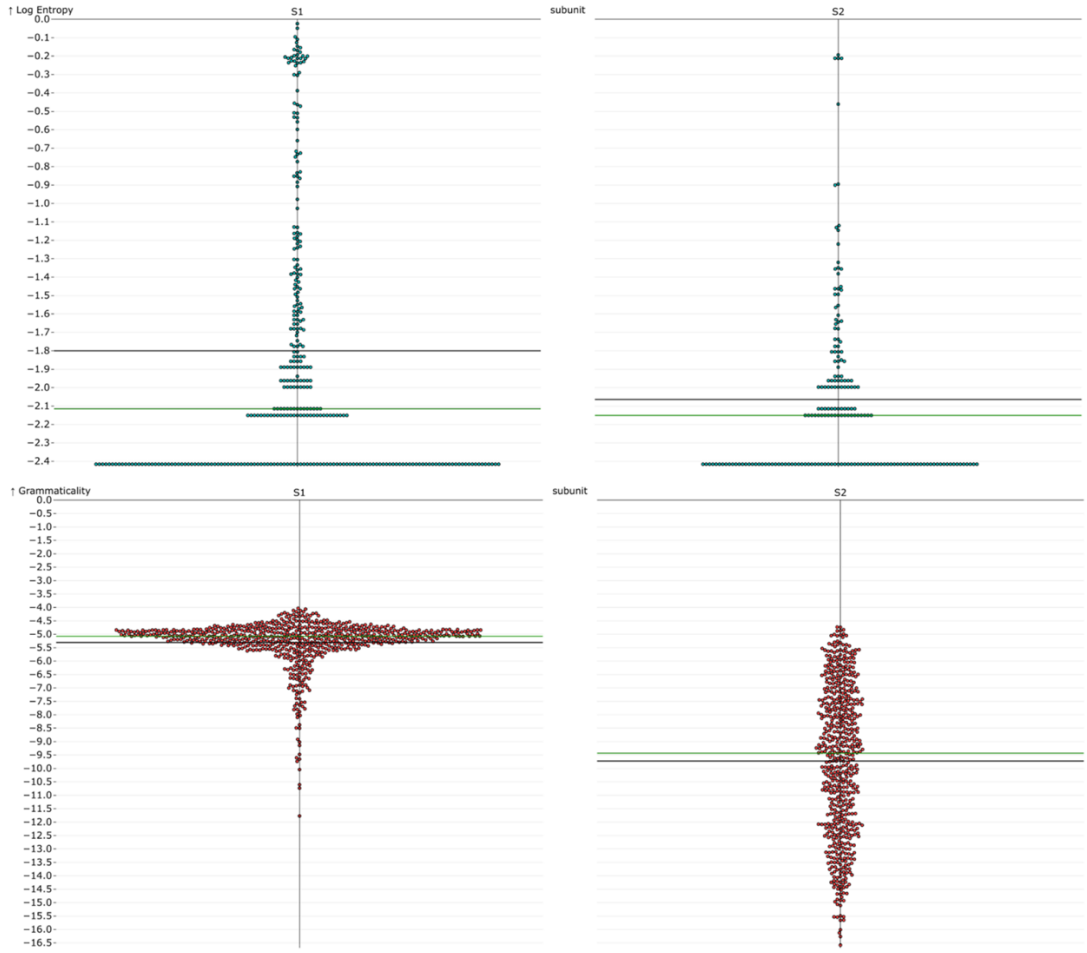
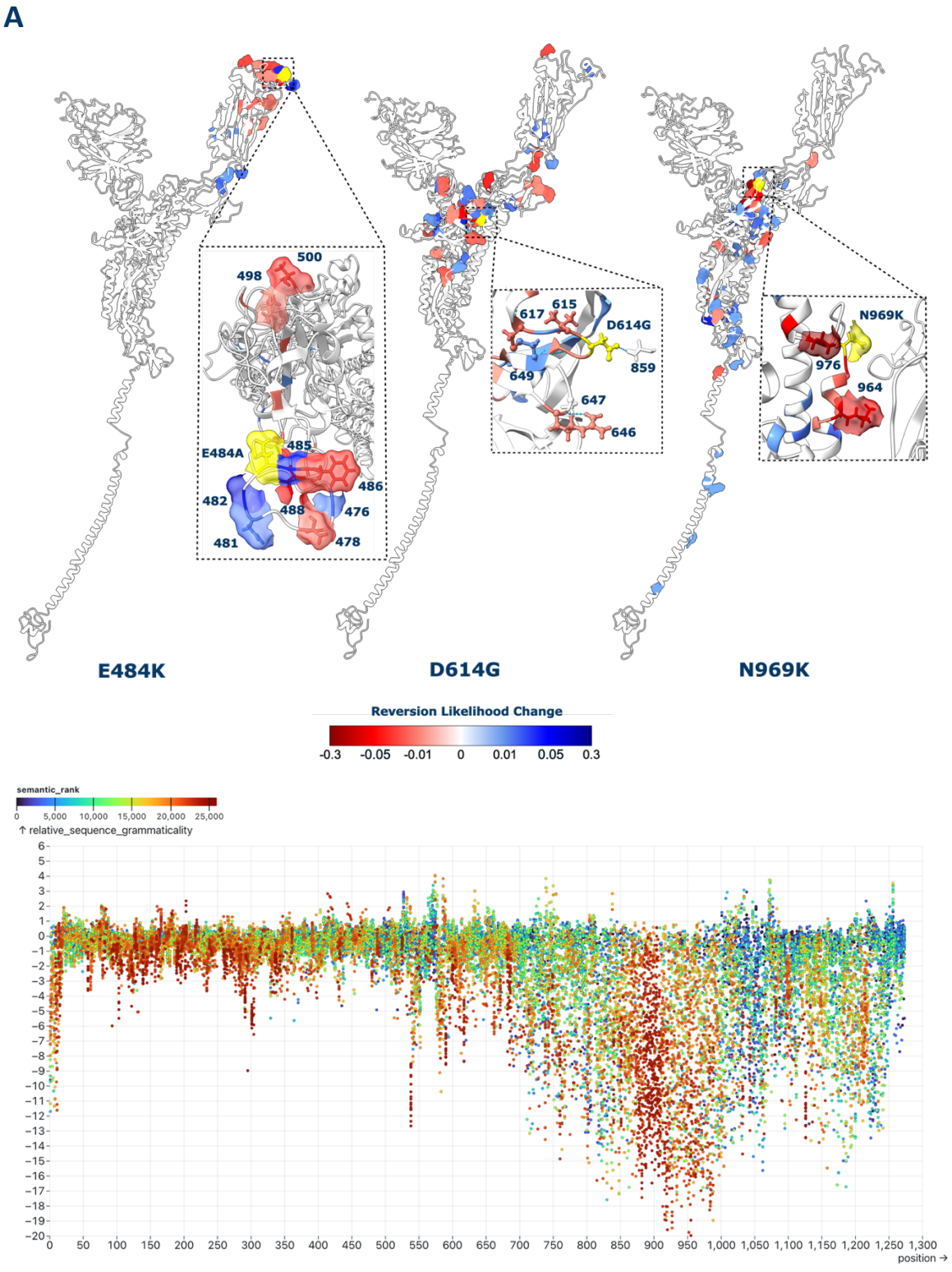


Figure 3.5: Swarm plots for the distributions of entropy and grammaticality for each of the spike subunits. The black line shows the mean while the green shows the median. For both entropy and grammaticality, the S2 subunit has on average lower scores compared to the S1.

3.4.3 LANGUAGE MODELS CAPTURE EPISTATIC EFFECTS OF MUTATION

Since ESM-2 scores appear to correlate well with the spike protein properties (Figure 3.4), we investigated whether the protein language model "understands" the epistatic interactions among mutations within spike. Measuring the model probabilities for reference positions will help to identify the epistatic interactions induced by new mutations. Language models produce likelihoods for each amino acid in a sequence based on the context of the rest of the sequence. This means that the likelihoods depend on the other amino acids in the sequence, and therefore change if other positions mutate. Positions that are unaffected by the mutated site would be expected to change very little outside of stochastic noise, whereas those affected may shift their probabilities up or down depending on whether the mutation is complementary or not. The BA.1 Omicron VOC contains over 30 mutations in the protein that make significant changes to the structure. By reverting the changes of the mutations from the Omicron(BA.1) sequence back to the reference sequence amino acid, we can measure the effect this reversion has on the probabilities of all other Omicron(BA.1) amino acids. We did this for each of the mutated sites in Omicron(BA.1) and describe in detail here the interactions of three key mutations E484A, D614G and N969K.


A
500
498
E484A
485
486
482
488
476
481
478
615
617
D614G
649
859
647
646
N969K
976
964
E484K
D614G
N969K
Reversion Likelihood Change
-0.3 -0.05 -0.01 0 0.01 0.05 0.3
semantic_rank
0 5,000 10,000 15,000 20,000 25,000
↑ relative_sequence_grammaticality
6 5 4 3 2 1 0 −1 −2 −3 −4 −5 −6 −7 −8 −9 −10 −11 −12 −13 −14 −15 −16 −17 −18 −19 −20
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1,000 1,050 1,100 1,150 1,200 1,250 1,300
position →

C

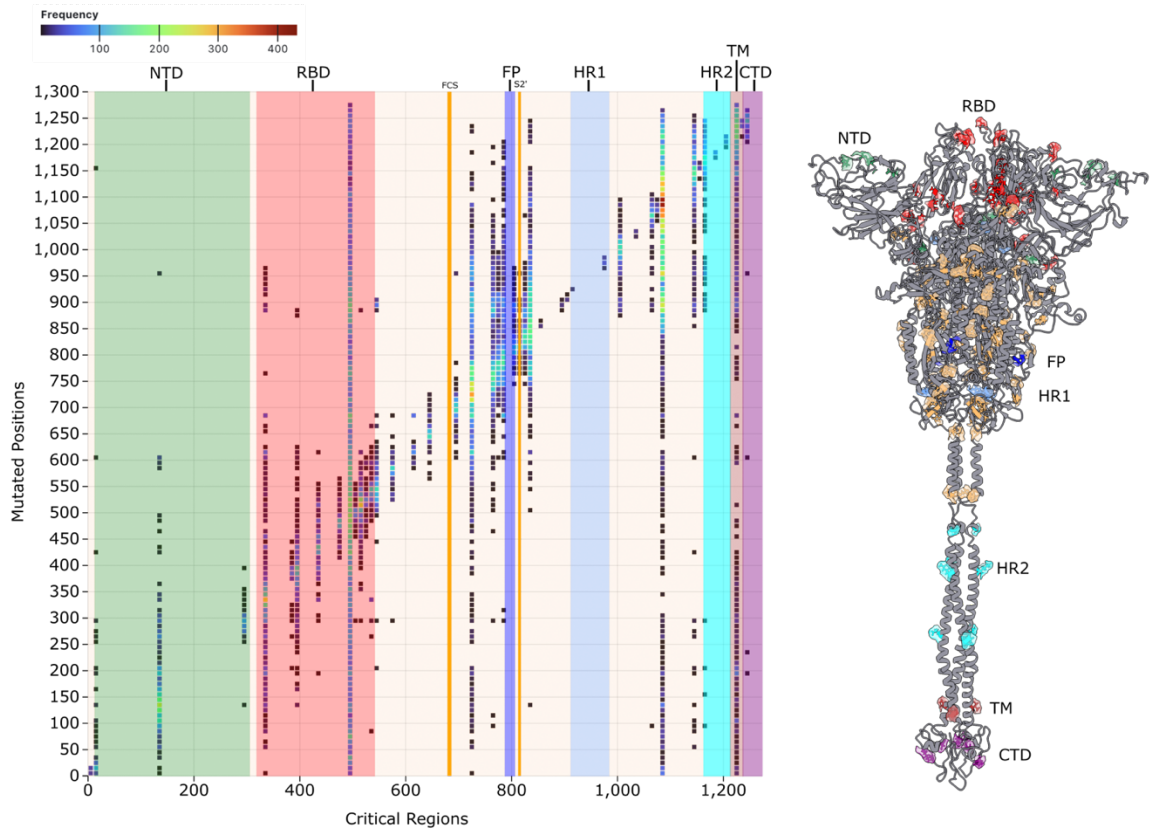


Figure 3.6: (A) Monomeric structures of the spike protein showing the changes in probabilities for 3 mutations: E484A, D614G and N969K. The mutation site is coloured yellow, blue sites increase in probability while red sites decrease. Mutation probabilities were only shown if they were outside two standard deviations of the mean change. (B) Relative sequence grammaticalities, the product of each amino acid likelihood rather than just the mutations, against the amino acid position. Amino acids are coloured on the semantic rank, which is a ranking of the semantic scores of all positions from highest to lowest semantic score. (C) The significantly changed logits across the whole DMS were identified, with positions that were repeatedly identified (called critical sites) as being affected by the in-silico mutations counted. These were then mapped onto the spike structure and coloured on their domains.

E484A is an amino acid change that occurs in the S1 RBD region of the protein which is necessary for ACE-2 binding. Reverting position 484 from A back to E produced shifts in the likelihoods of the amino acids primarily within the RBD (Figure 3.6A). Since position 484 is on the RBD surface, this is unsurprising since it is unlikely to be in contact with many other positions in the protein outside the domain, unlike an internal change. S1 reversions also appear to have shorter distances to affected positions in sequence space and in 3-dimensional structural space (Euclidean distance between structural atomic coordinates) (Figure 3.7).

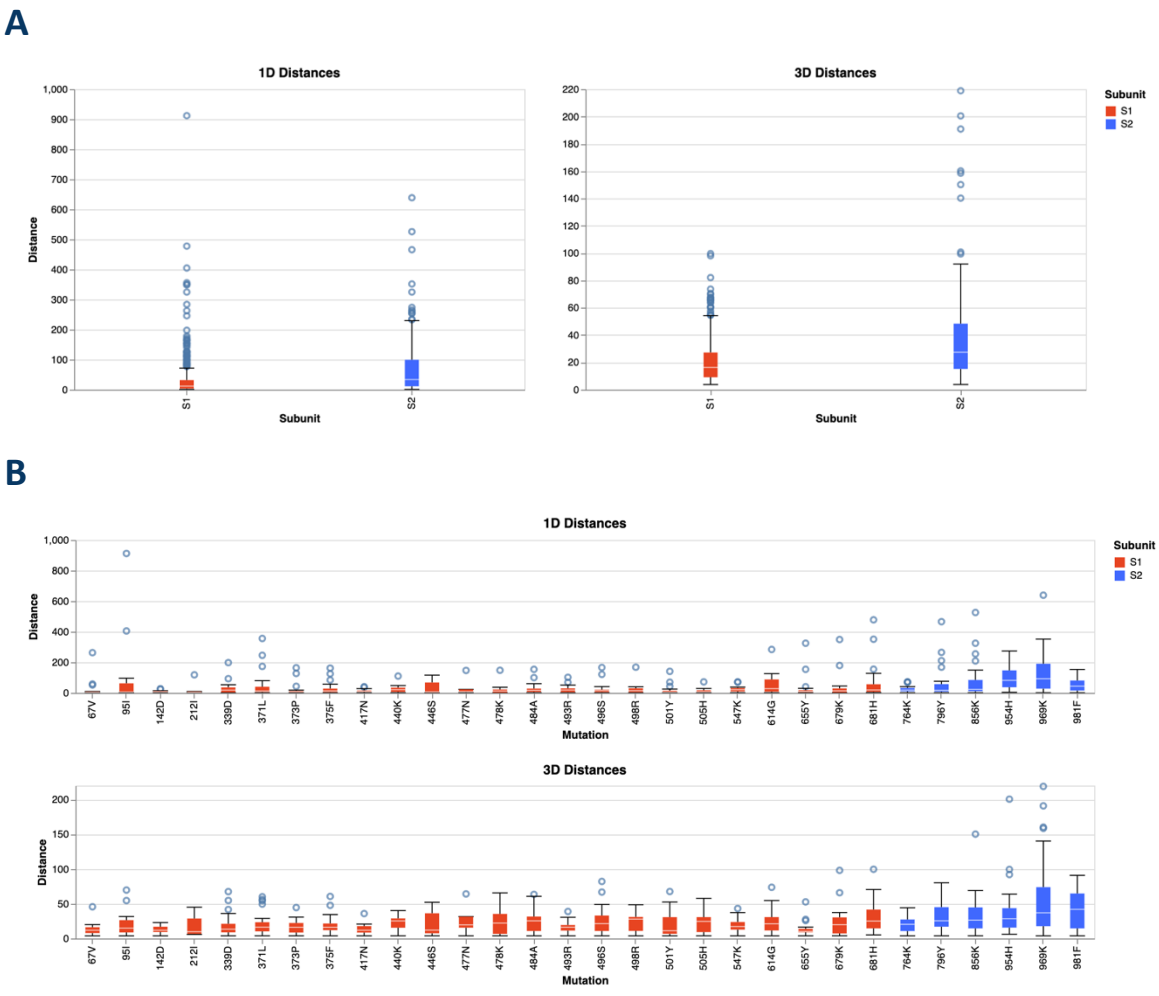


Figure 3.7: (A) Boxplots showing the distribution of distances between mutations and the positions affected by mutations in each subunit. (B)

Boxplots showing the distribution of distances between mutations and the positions for each mutation.

However, its impact on the other surface amino acids in the RBD is intuitive and may help explain why the site is a potent site of immune evasion. The E484A change primarily impacts local sites in the RBD, at positions 488, 485, and 486 having the largest absolute changes in probability.

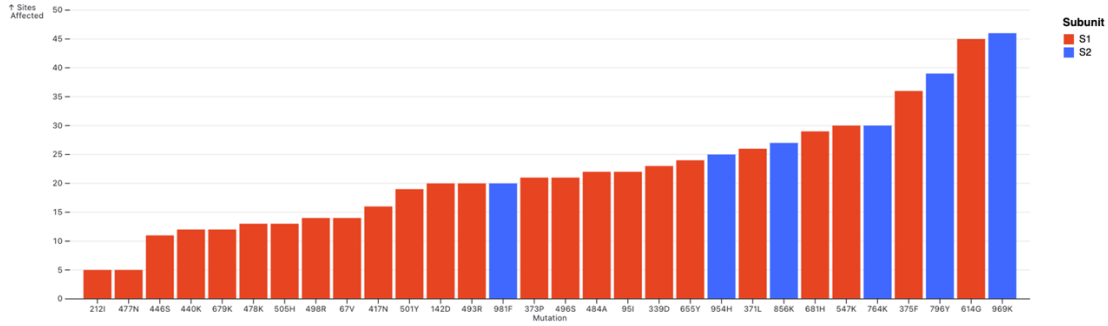


Figure 3.8: Number of sites with a significant (+/-2 deviations from mean) change in probability for each Omicron(BA.1) reversion change.

The D614G mutation emerged in spike early during the pandemic[78,216,217], defining the B.1 lineage, and is now present in all circulating lineages, making it particularly interesting to study. It is located near the end of the S1 domain and is therefore more central in the protein. It is strikingly the amino acid change that has the second largest number of detected interactions with 45 positions affected (Figure 3.8). Nine of these are in the RBD, with a further 30 in S1 and the other 6 in the S2 region. The largest changes occurred in positions 615, 617 and 649. 649 is an internal amino acid located directly behind 614 while 615 is in direct contact with the site and 617 is slightly downstream. The D614G reversion revealed that sites 617 and 615 both reduced their likelihoods, suggesting that the D614G acquisition was a complementary mutation for those sites. Much like E484A, these sites are in

direct contact with the site which explains why these are the most affected positions. Interestingly, position 649 appears not be complemented by D614G, with a likelihood increase of 3.7% on reversion.

N969K is an amino acid change that occurs on the S2 subunit of the protein and has the next largest number of detected interactions with 46 positions affected (Figure 3.8). It has some of the longest-range interactions both in positions along the sequence (more than 600 amino acids away) and in angstroms (more than 200 angstroms away) (Figure 3.7A and B), which would appear to indicate that it was a very consequential mutation. N969K also has the greatest number of affected sites compared with the other Omicron(BA.1) mutations. It is situated near the top of the S2 region and is right in the centre of the protein. Two positions have an order of magnitude larger shift in likelihood resulting from the N969K reversion: 976 with a 21% decrease and 964 with a 15.9% decrease. This would suggest these two positions are very accommodating of the N969K mutation and are strongly affected by this change. N969K is positioned in the middle of a loop connecting two alpha helices at the top of the S2. 976 and 964 are positioned at the end of the first helix and near the start of the second. The impressive shift in likelihood suggests that N969K must be interacting with these positions in some way, whether it be directly or indirectly. N969K is in the HR1 region of Spike which is involved in membrane fusion and its corresponding conformational change. The changes are predominantly in S2, with only 2 sites (330 and 617) out with this subunit.

Site 330 was consistently found to decrease its likelihood in these three reversions, but also in a number of other mutations in Omicron(BA.1). This

prompted an investigation into whether certain sites were more prone to fluctuations than others, and whether these sites were of functional relevance or importance. Using the data from the DMS, differences in the likelihoods between the reference and mutant reference positions were again filtered for significance and counted across all mutants to identify recurring positions. Figure 3.6C shows that the language model detects several sites in the protein that consistently fluctuate their likelihoods in response to changes elsewhere. Strikingly, positions 499**,** 723**,** 1087 appear to be affected by substitutions across the whole protein while other positions such as 330 tend to have effects at closer ranges. The consistently affected amino acids in the NTD and RBD regions of spike appeared to be mostly prolines and cystines, while in other regions there is a larger spectrum of amino acid types 8 of which contain aromatic side chains (F and Y) (Figure 3.9).

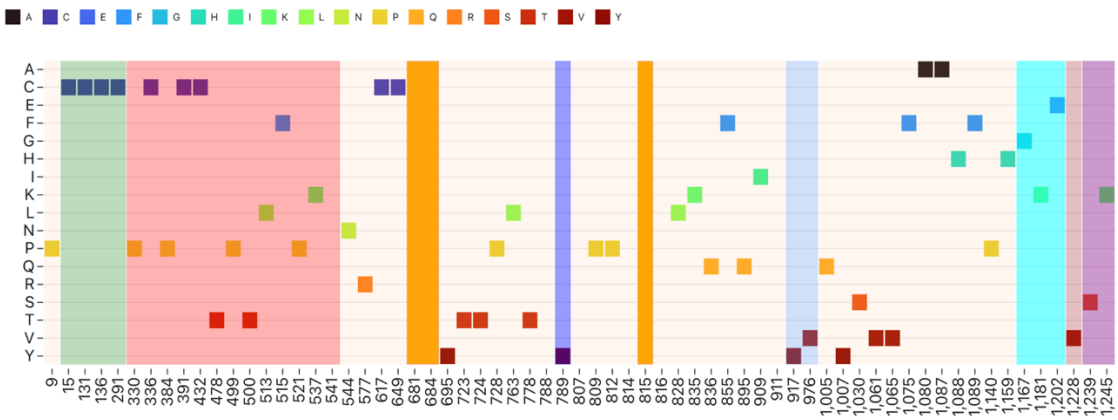


Figure 3.9: Amino acids of each consistently affected reference residues from the DMS data. The NTD and RBD appear to contain mostly Prolines(P) and Cystines(C), while the rest of spike has a wider distribution of amino acids.

Using the grammaticality of the whole sequence rather than just its mutations allow for an estimation of the overall epistatic effect of the mutation on the

sequence, rather than the likelihood of a mutation occurring given the sequence (Figure 3.6A and B). This relative sequence grammaticality is the product of the mutated sequence logits (a pseudo-log likelihood[218]. While some mutations are unlikely within the language model, selection pressure may allow these mutations to arise and propagate if beneficial. We see in Figure 3.6D a similar distribution to Figure 3.4A, however many more positions are now positive indicating that the epistatic effect of the mutation has made the sequence more likely than the reference. Several important mutations that have been observed within VOC sequences are among theseTable 3.1. Interestingly, all these mutations were found to be less likely than the reference with the relative grammaticality. D614G has a negative relative grammaticality of -3.07, yet its sequence grammaticality is slightly positive at 0.105, suggesting that the overall sequence is fitter with 614G than 614D. E484A similarly appears to be a positive mutation in the context of the sequence while the mutation has a negative grammaticality. N969K however is still negative even with the sequence context, suggesting that its effect on the other positions still make it less likely than the reference sequence.

Table 3.1: VOC mutations that have negative relative grammaticality scores, yet positive relative sequence grammaticalities.

VOC Mutation	**Relative Grammaticality**	**Relative Sequence Grammaticality**
L18F	-3.249494	0.33822632
T19I	-5.546568	0.5583191
D138Y	-3.0650895	0.29580688

R190S	-2.4990146	0.6814575
D405N	-2.806141	0.08947754
K417N	-3.1624873	0.06442261
K417T	-2.6595893	0.29382324
G446S	-2.9321728	0.18954468
E484K	-3.051447	0.012939453
E484A	-2.5325012	0.15188599
D614G	-3.07005	0.10519409
H655Y	-2.001164	0.9272156
P681R	-2.175136	0.13928223
T716I	-5.641353	0.5302124

3.4.4 CONTEXTUALISING EMBEDDING SCORES OF STRUCTURAL AND EVOLUTIONARY METRICS

While the embedding metrics appear to be meaningful, using other experimentally or evolutionary-derived metrics may help to contextualise the metrics and ground them in prior knowledge. First, we selected the embedding metrics and calculated spearman's rank correlations between the scores and several meaningful traditional metrics.

Experimental scores were used from three *in vitro* DMS studies, two of which were performed on the RBD while the other was performed on a full spike protein. Escape, Entry and Binding were determined from a full spike DMS by Dadonaite et al. (2023)[210] and correspond to: immune escape from human sera,

cell entry (measured using luciferase expression from pseudotype infected cells), and binding affinity to soluble ACE-2 respectively. RBD Escape was produced by Yisimayi *et al.* (2024)[209] using an RBD only DMS. The Wuhan and Variant binding and expression scores are from an RBD only DMS from Starr *et al.* (2022)[208] where binding to ACE-2 and expression of the protein were measured across five spikes. The Variant score is made from an average of both measurements across all five spikes (Wuhan-Hu-1, Alpha, Beta, Delta and Eta).

We also calculated several computational metrics that are commonly used for the analysis of protein sequences and estimation of evolutionary constraint. The crystal structure of the spike protein was used to calculate Accessibility, B-Factor, ΔΔG and the ESST. Accessibility measures the antibody accessibility at every position. The B-Factor is the temperature factor derived from the protein crystallography experiment used to determine the 6VXX structure and is a measure of the local fit of the structure to experimental data; it is often increased if there is static or dynamic disorder. Protein stability change (ΔΔG) was assessed with the FoldX software. The ESSTs were taken from Mizuguchi et al.[204] and provides a statistical estimate of substitution likelihood based on the observed frequency of substitutions at amino acids in similar local amino acid structural environments. We also calculated two sequence-based measure of substitution likelihood. The position-specific scoring matrix (PSSM) was calculated using DeltaBlast and calculates the log likelihood of substitutions occurring based on their frequency in related sequences[219]. "*Entropy"* was used to provide an estimate of the variability present at a site.

We observe that the language model scores (semantic, grammaticality and sequence grammaticality) do not strongly correlate with any of the tested experimental or computationally derived traditional metrics. However, at least 1 of the language model metrics significantly correlated with all the traditional metrics (Figure 3.10A). They instead appear to represent something different entirely.

The semantic score had the strongest negative correlation with likelihood metrics (ESST, PSSM) as well as RBD based DMS results (Figure 3.10A). Grammaticality scores performed similarly or slightly worse for the probability metrics although the scores were inversely correlated versus the semantic score. B-Factor had the best correlation with relative grammaticality and relative sequence grammaticality which both having a positive correlations greater than 0.4. This suggests that grammaticalities (which align well with conservation) may also be a reasonable proxy for protein flexibility. All metrics struggled with the full spike DMS scores, with correlations of less than 0.2 for Binding and Escape. Entry was more correlated however, with correlations of greater than +0.3 and less than -0.3 for grammaticality and semantic score respectively. Correlations were also below 0.2 for all metrics both for Escape (full spike) and the RBD Escape. While model metrics clearly correlate with known features, this experiment shows they are not simply a new and alternative way to describe existing metrics.

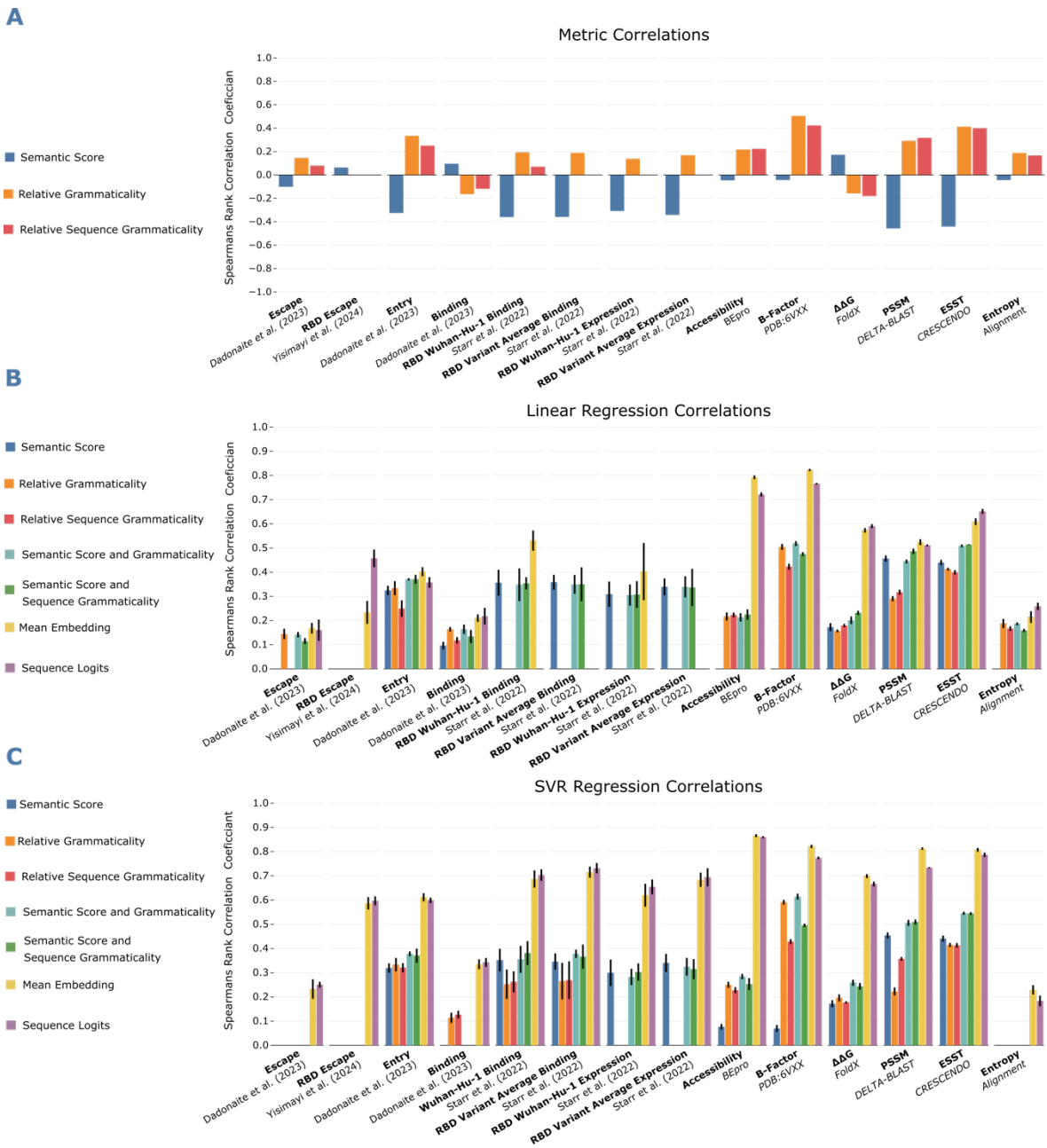


Figure 3.10: (A) Spearman's Rank correlations between the semantic score, grammaticality and relative sequence grammaticality and traditional metrics. Bars not present in a metric category means the correlation was not found to be significant after a Bonferroni correction. (B) Spearman's Rank correlations between the language model metric and the traditional metric. Each pair was fitted using a grid search and a linear regression model, with 5-fold cross validation. Bars represent the mean of the correlations, with the error bar +/- 1 standard deviation of the correlations.

Bars not present in a metric category means the correlation was not found to be significant after a Bonferroni correction. (C) Spearman's Rank correlations with a support vector regression model using an RBF kernel and 5-fold cross validation.

Next, we used the embeddings, logits and score combinations to fit a linear regression and a support vector regression (SVR) to each of these traditional metrics and scored the regressions fit using a spearman's rank correlation. When they are found to be significantly correlated, linear regression predictions for logits and embeddings either match or drastically outperform correlations using the model metrics or their combinations (Figure 3.10B). This is particularly apparent for accessibility and B-Factor with both logits and embeddings achieving correlations of >0.7 for both metrics. ΔΔG also improves dramatically with correlations of >0.5 for logits and embedding features. The RBD Escape data also achieves greater correlations with linear regression predictions using the logits, reaching >0.45 compared with less that 0.1 for the score correlations alone. Combinations of scores don't appear to improve correlations much over individual scores.

Fitting SVR regression results in more consistent and better correlations across the board, with only binding, escape and entropy achieving correlations of less than 0.5 with logits and embeddings as features (Figure 3.10C). The RBD Escape scores also improve reaching a correlation of ~0.6, although notably the full spike DMS score remains low at ~0.2.

Fitting models using sequence logits and embeddings produces much better correlations with metrics than simple model derived scores. This is expected since the embeddings and logits provide significantly more information about

each sequence than can be distilled into a singular score. While the SVR here performed well for many of the metrics, it seems likely that some parameter tuning and testing of other model types will produce even better results.

3.4.5 Horizon scanning of UK SARS-CoV-2 sequences

Once regular sequencing and surveillance are underway, a pipeline for analysing newly sampled sequences is useful to identify outlier sequences that may become future VOCs. As we can see from scanning lineage sequences, language model scores appear to distinguish between the VOC classes well, but a bigger score does not always mean a variant will take over. As such, we looked at creating a dynamic moving average embedding that represents the average sequence circulating at a given time. This way, we can check whether sequences are both divergent from the SARS-CoV-2 reference, and whether sequences diverge greatly from what is currently or recently circulating. By looking at a UK subset of the GISAID sequences, we can assess the usefulness of this approach thanks to the high sequencing capacity of the UK as well as the well-defined lineage waves that were experienced there. UK sequences were clustered to 99.9% similarity and a representative haplotype sequence for each cluster was embedded each using the model. Sequences were measured against a mean embedding of the sequences from the prior 3 months.

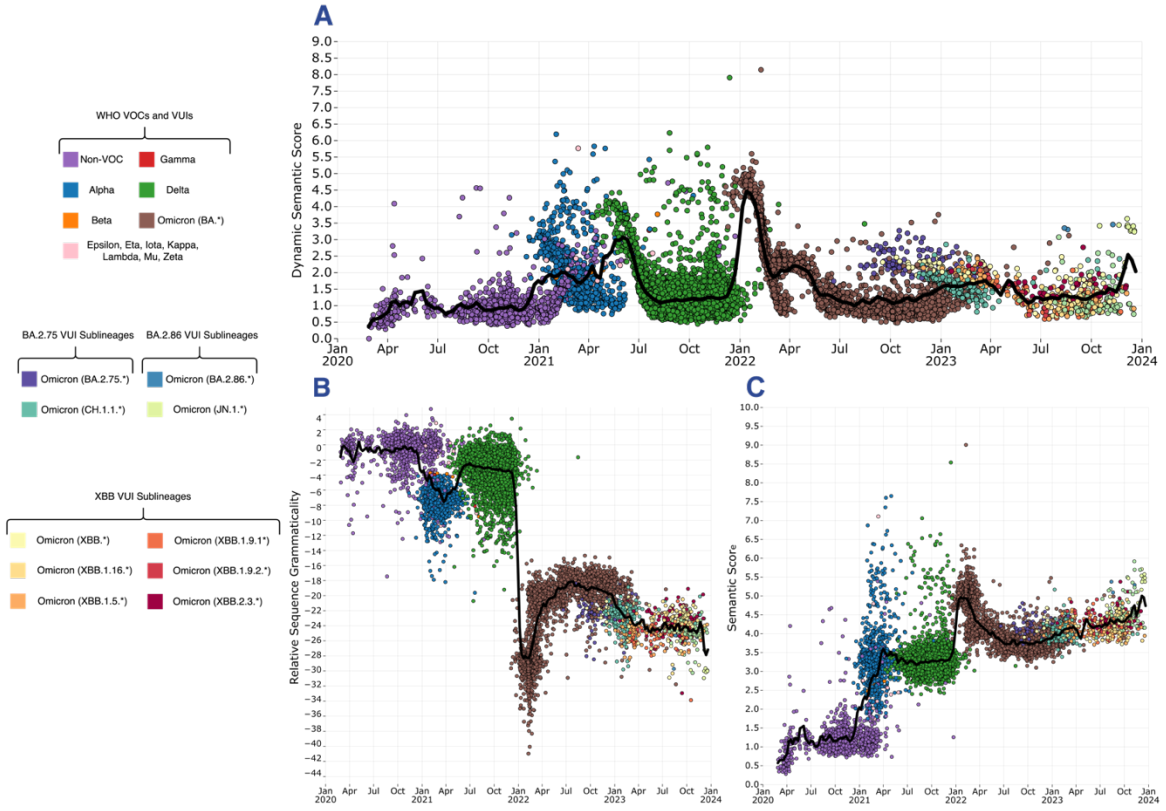


Figure 3.11: (A) UK SARS-CoV-2 spike sequences through the pandemic. Each point represents a sequence cluster with 99.9% sequence similarity. Dynamic semantic scores were calculated for each sequence cluster, with the black line showing the mean sliding score. (B) Relative sequence grammaticalities for each of the haplotype spikes. (C) Semantic scores for each of the haplotype spikes.

Waves of semantic change take place between all major VOC waves in the UK, i.e., Alpha -> Delta -> Omicron(BA.1) ->Omicron(BA.2) (Figure 3.11). The non-VOC->Alpha transition resulted in sequences deviating by semantic score of >2 although this quickly increased to almost 3 after the emergence of Delta. Following BA.2, the semantic changes in Omicron have been more incremental, with a gradual increase up to March 2023, a decrease until July 2023. There have been several smaller waves in the Omicron section, which

suggests repeated dominance and replacement of multiple distinct sublineages throughout this period. The haplotype embeddings also uncover the large diversity of semantic scores contained outside just the Pango representative sequences. Alpha and Delta in particular have a large range of semantic scores, equalling some of the most divergent Omicron sequences despite pre-dating them by months (Figure 3.11C). To a lesser extent, the Non-VOC sequences also have some very highly semantic sequences relative to the average semantic score during the Delta wave. What is clear is that as a new VOC began to take over, sequences diverge very quickly from the average circulating sequence embedding.

Alpha subvariants appeared to heavily decrease their sequence grammaticality, which might be expected due to its increase in semantic score (Figure 3.11B and C). Delta appears to have maintained a high sequence grammaticality despite having a similar number of mutations to Alpha, which might indicate why it outcompeted Alpha. Omicron (BA.1) gains a modest increase in its semantic score while sacrificing a large amount of sequence grammaticality. However, this is restored by Omicron(BA.2) and later Omicron derivatives until August 2022, after which point sequence grammaticality begins to decrease and semantic score begins to rise. There is a small cluster of Omicron(JN.1) sequences towards the end of sampling that appear to be increasing the semantic score (dynamic as well as normal) and decreasing the relative sequence grammaticality.

3.5 Discussion

Protein language models represent a new approach for understanding protein sequence properties by learning from the high numbers of available protein

sequences. PLMs are a powerful tool for compressing a significant amount of evolutionary information and act by transforming a sequence into a numeric vector representing the models understanding of the sequence's intrinsic properties. Like traditional language models, the real-world beneficial use cases for these models are still being explored. However, the ability to predict structure demonstrates the impressive tasks these models make possible. Here, we describe several different ways to use these models that either supplement existing information about a novel virus (SARS-CoV-2) or expand directly upon it. The use cases for each of these techniques can be summarised into early, mid, and late stages of an outbreak which we will discuss here.

3.5.1 EARLY

3.5.1.1 DMS

Following the sequencing of the virus, ESM-2 can be quickly used to produce DMS data for each protein in the virus genome. Here, we focused on Spike glycoprotein, but protein language models can be used for any of the other proteins, and almost certainly should be, to gain a more holistic view of the virus and its properties. ESM-2 appears to understand spike protein constraints since grammatical changes are shown to be very unlikely in the S2 subunit of spike while regions heavily targeted by host antibodies (NTD, RBD)[220–224] are shown to be much more likely and have higher grammaticalities. The relative grammaticality score, therefore, seems to align well with structural conservation. Amino acid changes in the core of molecules are more likely to be deleterious and in the case of viruses, targeting of amino acids on the surface of the protein by antibodies is a driving force for the fixation of novel changes. In the presence of a protein structure

(experimentally or computationally produced), the scores can be mapped onto the structure to gain further insights. This is particularly useful when variants emerge, and the semantic scores and grammaticalities for the variants mutations can be mapped rather than an average as in Figure 3.4C. The semantic score (and the semantic rank) identifies regions of interest in both the S1 and S2 regions of Spike (Figure 3.4A). Hie et al[193] use the semantic score as a proxy for antigenic change, however, this is difficult to reason with at positions that are not involved in any sort of antibody binding. Since ESM-2 is trained on many different proteins, it makes more sense to view the semantic score more like a shift in structure/function rather than specifically antigenic change. Through this lens, the score makes more intuitive sense, since a large part of S2 is internal and would not be antigenic as such. We see regions of high semantically ranked changes in the NTD region of S1 as well as in the centre of the S2 subunit between positions 850 and 950 (Figure 3.4A). The NTD contains what is known as a "supersite" which contains several epitopes targeted by potent neutralising antibodies[223]. NTD also remains an area of low relative grammaticality and appears to be accommodating for mutations. With many very semantic positions present in NTD, there may be plenty of opportunities for the region to make big structural changes to alter its antigenic properties without greatly affecting function.

3.5.1.2 Mutational Epistasis

In the prior SARS-CoV-2 variants Beta and Gamma, the 484 position changed to E484K and helped escape several RBD-neutralising antibodies[80,159]. Substitutions of A, D, G and K at 484 were all shown to confer some level of

resistance to human convalescent sera[80,222]. E484A appears to have decreased immune evasiveness relative to 484K, yet this was thought to be overcome by epistatic interactions from the nearby 501Y and 498R mutations[225]. These interactions helped to boost E484A's immune evasive properties to 484K-like levels. The strongest of these epistatic interactions was found to be between 498R, which is one of the epistatic interactions picked up by the model. 498R has a lower likelihood upon E484As reversion, suggesting that the acquisition of an A at the site is complementary to 498R. While 501Y is not identified, position 500 is one of the most complementary sites identified with a likelihood change of -3%, again suggesting the E484A mutation has a positive interaction with this site. The largest change in likelihood comes from position 488, which is mostly internal in the RBD. It does have small surface pockets, however, and is in direct contact as a result with 484, which explains the magnitude of the change. It is also directly behind 484 in the RBD structure, so likely in contact internally as well.

614G was found to increase the stability of the spike protein[217] by preventing premature cleavage of the S1 subunit from the S2. It also appears to slightly increase both the infectivity of the virus, as well as susceptibility to neutralising antibodies[216,217]. The mutation in the wildtype spike appears to have two main effects: the first is to remove a hydrogen bond between 614 and 859 on an adjacent spike monomer in the S2 domain, the second is at 647 which is now closer to 614 and as a result may reinforce the structure of the C-terminal domain. ESM-2 does not detect the 859 changes, although this might be expected due to the nature of current language models. ESM-2 accepts a single sequence, yet the Spike protein exists as a trimeric multimer. While the

model may be indirectly aware that spike is trimeric due to its amino acid composition, the second and third monomers are not present during model inference. As such, there may be more limited interpretability with inter-protomer interactions. Training models that more explicitly model multimeric proteins or protein-protein interactions (PPIs) could be key to unlocking more implicit interactions and better effect prediction. This has become apparent with other models such as AlphaFold, with specifically trained multimeric versions of being trained to improve prediction[99,226]. The interaction with residue 647 is not directly picked up either, although curiously positions 646 and 649 are identified as significant changes, suggesting the model has detected this region as being involved with the D614G change despite missing the exact position.

N969K is one of the few Omicron mutations that appear in the S2 region and has a large impact when reverted in the Omicron(BA.1) backbone (Figure 3.6A, Figure 3.7 and Figure 3.8) and has also been linked to the change in Omicrons entry approach[227]. As such, it is clearly an impactful mutation and due to its positioning is also likely to have inter-protomer interactions. This has been shown to be the case with the mutation forming electrostatic contacts with the Q755 residue on the adjacent protomer in the pre-fusion state[228], as well as interacting with and displacing the HR2 backbone in the post-fusion spike structure[229]. Similar to D614G, these contacts are not observed potentially due to the lack of the other protomers in the ESM-2 embedding. N969K has no obvious monomeric interactions, yet the 975 and 964 positions were found to be strongly affected by the reversion and could be as yet undiscovered and important interactions.

Assessing mutations using the relative sequence grammaticality (pseudo log-likelihood) is also meaningful. Building on the epistasis experiments, it is apparent that understanding how a mutation impacts the likelihoods of the other positions in the sequence can be important in interpreting its effect. This is clear when comparing the relative grammaticalities to the relative sequence grammaticalities, since many mutations are predicted to be less likely than the reference for the former yet more likely with the latter (Figure 3.4A and Figure 3.6B). This is particularly interesting since a number of VOC mutations appear to follow this pattern, suggesting that reference positions may help better identify VOC mutations. The VOC mutation N969K does not change its prediction between scores but in this particular case the mutation may be a fitness trade off as it appears to destabilise the proteins postfusion conformation[229] while stabilising it perfusion context[228].

The sites identified as consistently affected by mutated positions elsewhere in the protein are predominantly amino acids with important structural features. Proline is a rigid amino acid due to its sidechain binding to its amine group forming a ring and thus reducing its flexibility [230,231]. This sidechain binding also prevents prolines from forming the N-hydrogen bonds necessary to form stable α-helices, except for the first 4 residues which do not require hydrogen bonds[230]. As such, prolines are often found in flexible loops, initiators of α-helices, or where sharp turns are necessary for structure. Cystines are also highlighted and are another amino acid with a unique structural property. The presence of a thiol (sulphur-hydrogen group) containing sidechain allows cystines to form strong covalent di-sulphide bridges between adjacent

cystines[232]. Cystines are therefore often critical to protein structure, and several of the identified cysteines in the spike protein critical sites were cysteines involved in di-sulphide bridges (Figure 3.6C and Figure 3.9). This may explain why they are well conserved throughout proteins, and particularly when they are involved in di-sulphide bridges[232]. It appears that the language model is identifying the constraints that these positions are under due to their functional importance, based on its understanding of protein sequence.

To summarise, using just a single spike sequence (and a structure where available but not necessary), we were able to determine the likely regions of variability in the Spike protein, and identify regions where mutations were most likely to impact the structure and function of the protein.

3.5.2 MID

3.5.2.1 Validation of computational methods

With available experimental data, we can validate that the embeddings contain relevant biological and evolutionary information (Figure 3.10). ESM-2 is a good predictor of many of the metrics we commonly use to understand how mutations impact protein structures. While none of the model scores correlate strongly with any one feature, this is expected and indeed encouraging. The language model scores appear to be telling us something different about the proteins from what we can already determine with existing methods.

Language models are now being used for DMS and variant effect prediction tasks across an impressive array of datasets, and often outcompete other methods[233]. They are also being used to improve the effectiveness of antibodies[234] and even generate new and functional proteins[121]. Our results

here affirm that these models are useful and can be applied to great effect in the scenario of understanding a novel viral pathogen.

3.5.2.2 Variant of concern sequence embeddings

ESM-2 combined with the evo-velocity managed to successfully capture the differences between variant of concern sequences and cluster them appropriately. More impressively, it managed to recapitulate the topology of the nucleotide sequence phylogeny and reconstruct the direction of evolution accurately (Figure 3.2B and Figure 3.3) from the diverse bat sequences up to the latest Omicron sequences. Like evo-velocity, the language model scores are effective at separating out the different lineages into their VOC categories (Figure 3.2D and Figure 3.13). Beta and Gamma appear to have the lowest of the semantic scores (approximately 1.6 and 2.18 respectively), while Alpha (between approximately 1.8-2.3) and Delta are slightly higher (1.8-3.4), although Delta has some outliers that reach to almost 6. The semantic jumps (Figure 3.2) from Wuhan-Hu-1 like viruses (typically below a semantic score of 1) to early VOCs corresponded to regional sweeps by each of the VOCs (Alpha in Europe and North America, Beta in South Africa, and Gamma in South America[35]). Similarly, Delta also produced a worldwide sweep, mostly wiping out other VOCs. Each had a similar number of mutations (approximately between 8-12), yet Delta managed to outcompete them despite also having a similar number of mutations, yet a higher semantic score. The semantic score appears to do more than simply count mutations, conveying some structural that has an impact. However, it is hard to attribute a high semantic score directly to antigenicity, since Delta had a number of other features including better entry efficiency[137] and enhanced pathogenicity[235] which both appear to

be spike mediated. Since ESM-2 is trained on a wider range of proteins, the semantic score relating to these other confounding features should not be ruled out.

Omicron is the most semantic of all the VOCs with scores between 3.7 and 8.5, although interestingly it begins to decrease its semantic score over time. Since the protein still needs to remain functional, there may be limits to how different a sequence might be to retain its function. Crucially, sequences might be different semantically from the reference, but with shifting immune landscapes this may get increasingly less relevant. A real consequence of this can be seen through the need for more up to date vaccines, since more recent SARS-CoV-2 sequences tend to evade previously neutralising antibodies[79,236]. This was one of the reasons we investigated the use of a dynamic embedding. The Omicron cluster separates into 3 "levels" with decreasing semantic scores at each level and is reminiscent of the subclusters from the evo-velocity. This suggests that the virus does not necessarily always get more semantically different as time progresses. It's also worth noting that the score does not indicate the direction or type of difference that is being measured, meaning 2 semantic scores of 6 could also be 6 away from each other. This might relate to two different functional spikes with similar phenotypes like early VOC sequences, where spikes had similar semantic scores yet were antigenically distinct[237]. This was a big motivation for a dynamic score, since this dynamic reference helps explain differences between later sequences better. The three levels primarily correspond to major sublineages of Omicron i.e BA.1, BA.2 and BA.4/5. This is encouraging since BA.4 and BA.5 only differ outside Spike[236] , while BA.1 and BA.2 and BA.4/5 are distinct from each other

bar a set of common ancestral mutations. A caveat to these results is that so far, all VOC sequences have been significantly more mutated than prior circulating sequences. However, within Omicron, BA.1 Gammas outcompeted by BA.2 despite having fewer spike substitutions than BA.1. BA.4/5 spikes return to a semantic score more reminiscent of highly semantic Delta sequences yet have fewer spike substitutions again relative to Omicron(BA.1) or Omicron(BA.2). Deletions may have a role to play here, but also differences in the mutations between variants may have different combinatorial effects within the model. Increasing the number of mutations does not linearly increase the sequence's overall semantic score either. Later Omicron lineages often have more mutations than the earlier sequence but as yet have not reached higher semantic scores than Omicron(BA.1) sublineages consistently. Changes in the semantic score can often cancel out or amplify the effect of individual mutations in a way that mirrors the combinatorial effects of real mutations. In this way, the semantic score may be more interesting when observing the interactions between many mutations.

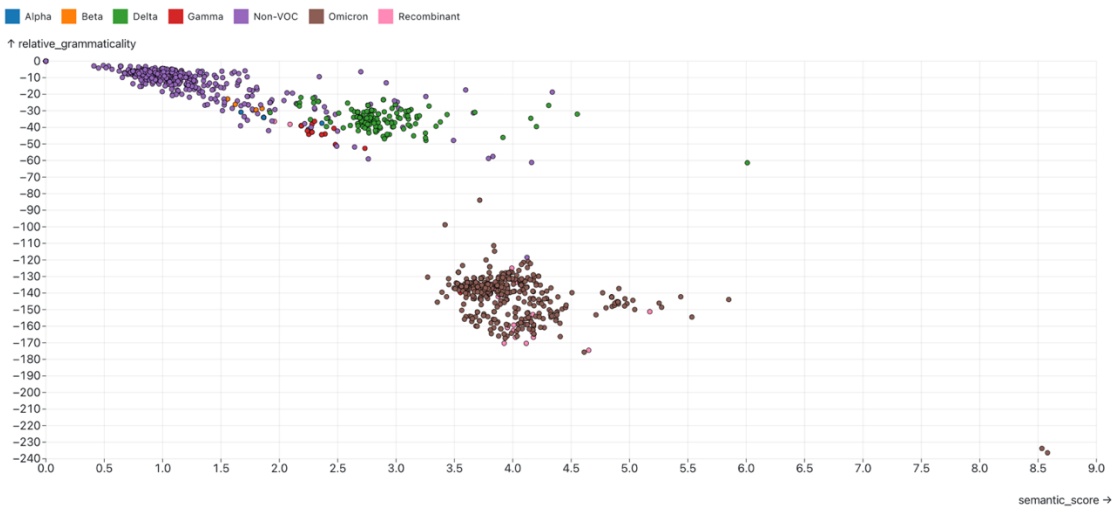


Figure 3.12: Pango representative lineage sequences plotted by their semantic scores and relative grammaticalities.

Relative grammaticality also produces a similar separation between lineages (Figure 3.12) although the Omicron sequences appear more homogenous, especially between BA.2 and BA.4/5. Relative sequence grammaticality however changes things quite a bit, with Delta sequences resembling the sequence grammaticalities of Non-VOC sequences and sometimes being greater than the reference SARS-CoV-2 sequence (Figure 3.6B). Alpha, Beta and Gamma all appear with lower sequence grammaticalities than Delta and are less semantically different which may explain why Delta managed to outcompete them globally. Omicron again shows differences between sublineages like those observed using the semantic score.

The metrics show a number of sequences that appear to be outliers that could be future lineages of concern. While these sequences clearly were not, many potentially concerning sequences might be sampled yet never spread due to reasons outside the sequence properties such as epidemiology. As such, the scores can be used to alert concerning sequences, but cannot alone predict whether the sequence will become one of concern. This is something that will be relevant for all methods both experimental and computational.

3.5.3 LATE

3.5.3.1 Horizon Scanning

The use of a dynamic embedding reference is clearly useful given the lack of circulating Wuhan-Hu-1 SARS-CoV-2. With the emergence of antigenically distinct variants already occurring, it seems likely that many will be repeatedly infected by SARS-CoV-2 over their lifetime[149]. By dynamically updating a reference embedding that estimates the properties of the currently circulating virus, we can move away from referring to a reference sequence that doesn't exist towards measurements that better reflect the immune

landscape of the population. We observe variant shifts in the data typically involve a jump in the semantic score and the grammaticality (Figure 3.2D, Figure 3.6B and C and Figure 3.13**)**

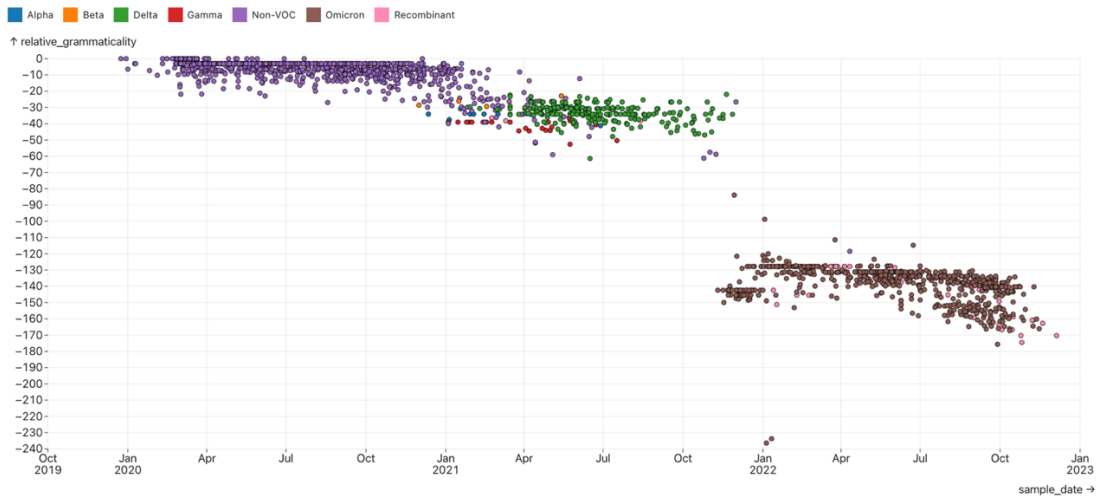


Figure 3.13: Pango representative lineage sequences plotted by their relative grammaticalities against sampling date.

The dynamic embedding also captures this behaviour when the average distances rapidly increase, which maps well to these transitions. We see shifts between the major VOCs (Alpha, Delta and Omicron) as well as between Omicron sublineages (notably BA.1 and BA.2). Encouragingly, we also see interesting transitions at later time points where sequencing has significantly reduced, particularly towards the end of our analysis. We see a marked increase after October 2023 which coincided with the emergence of the BA.2.86 sublineages. JN.1 is descendant from BA.2.86, a lineage that much like earlier VOCs emerged with a huge number of additional substitutions. It did not appear to spread well until it gained an additional L455S substitution[238], after which point it went on to expand rapidly and is now one of the dominant lineages circulating. Its identification by ESM's model scores highlights the approach as a potentially viable horizon scanning method. The method also

gets better when implemented across many different countries, since the populations will differ in their exposures to SARS-CoV-2 lineages which may show as different lineage waves using dynamic embedding scores.

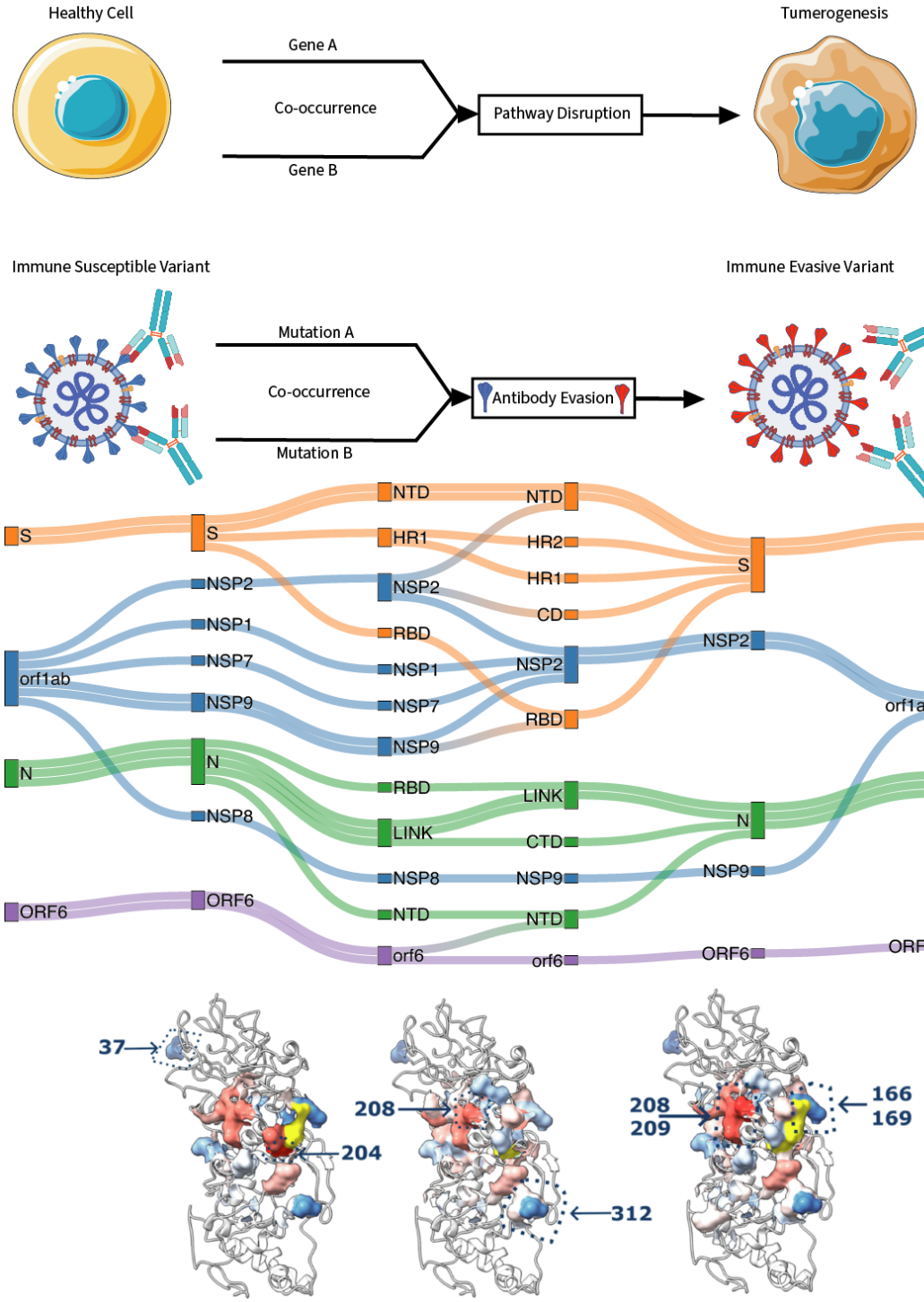
3.5.4 EXTENSIONS AND FURTHER THOUGHTS

There are clear areas for improvement when it comes to using language models to understand protein sequences. Finding ways to explicitly model protein-protein interactions that occur between multimeric and known interacting proteins rather than relying on the models implicit understanding could be key in improving performance. Most proteins exist as multimers[239], and those that don't are still likely to interact with other proteins. While tools like AlphaFold Multimer[226] are enabling multimer prediction, this is not quite at the level of performance that AlphaFold can achieve with monomer predictions. ESMFold can also predict multimer complexes, however these are typically worse than its monomer prediction as well. New pre-training tasks, inclusion of new tokens, expanded model architectures and more over the coming years are likely to advance on this. Despite this, there is already evidence that these models already can distinguish between monomeric and multimeric proteins even at this early stage of language model adoption[240]. Fine-tuning is another approach commonly used in NLP tasks that can improve performance by providing domain specific knowledge to a general sequence model to improve performance on domain specific tasks. A relevant and recent example from Ito et al.[241] uses Sarbecovirus sequences, DMS data and reproduction number estimates to improve upon the base ESM-2 model when predicting SARS-CoV-2 variant fitness.

ESM-2 is learning from a vast quantity of proteins in nature that are not necessarily subjected to the same evolutionary pressures that spike is. Across nature, similar proteins may use the same sequence to produce the same fold in a protein with a completely different function. These proteins might be subject to completely different pressures. For example, viral proteins could be pressured to evade the immune system whilst a host enzyme is unlikely to be. Fine-tuning large language models may be useful here as well, since focusing on proteins that are subjected to similar pressures might allow the model to learn more about where those pressures may exert their effects i.e. the RBD in receptor binding proteins. Fine-tuning on viral glycoproteins for example could help ESM to better identify conservation of structural and functional elements in these proteins without overfitting on the limited evolutionary information we might have for a single virus.

Currently, there are several ways protein language models can be improved or extended in the coming years. However, PLMs in their current state are clearly informative for many different tasks whether that be structural prediction, variant effect analysis or understanding evolutionary constraints. While imperfect, they provide a low cost, quick and actionable set of outputs that can be applied to a diverse spectrum of proteins and used in scenarios including the emergence of novel pathogens. Continued efforts should be taken to improve our understanding of these models, and work to improve their capabilities. Despite this, these models are already incredibly powerful tools that can and should be used now to supplement and expand our current understanding of proteins.

4 Investigating the co-occurrence of mutations in SARS-CoV-2



"Coincidence? I think NOT!"

Pixar Animation Studios: "The Incredibles" (2004)

4.1 Abstract

Over the last 4 years SARS-CoV-2 has evolved to produce thousands of lineages associated with many mutations. Thanks to the massive global sequencing effort, these changes have been tracked in detail and at a level not seen in prior viral datasets. This impressive resolution both in time points and sequence quantity means that mutational patterns such as those discussed in Chapter 2 can be interrogated. Another such mutational pattern is the co-occurrence of mutations. Here, we use a strict definition of co-occurrence that is defined as two mutations occurring in the same SARS-CoV-2 lineage. We show that even with this strict definition, we can extract significant co-occurring mutations across the viral genome. We breakdown the mutational contexts of these mutations, show how they are distributed across the SARS-CoV-2 phylogeny, and use protein language modelling to investigate how these mutations may interact with each other. This type of analysis could be used in future to identify important potentially interacting mutation pairs which may help understand future variants of the virus.

4.2 Introduction

Across datasets like The Cancer Genome Atlas (TCGA), there are numerous genes that appear implicated in cancer progression, yet many are in fact "passenger" events[242]. These genes do not contribute to the cancerous phenotype and as such they make it difficult to identify true driver events. Driver genes are often characterised by their presence across different occurrences of cancer and their links to pathways such as DNA damage repair or the cell cycle[243]. Cancer genes are typically split into proto-oncogenes which when mutated promote cell growth and division, and tumour suppressor genes which when mutated typically lose their function of cell growth and proliferation suppression[244]. The co-occurrence or mutual exclusivity of mutated genes can be used to help identify so-called "driver" genes which in their pathway context can induce cancer development[245–247]. Due to genes having similar phenotypes upon mutation or being part of the same pathway, it is possible to use the patterns of their occurrence to help determine their potential as cancer drivers. The mutual exclusivity of genes may indicate a carcinogenic function derived from disturbing a shared pathway, meaning only one is required to produce a tumour progression phenotype[248]. Co-occurrence of mutations on the other hand may indicate synergistic epistatic effects are require for cancer progression[243,245,247]. With co-occurrence, these synergistic effects can stem from the alteration of complementary pathways, effects in a single pathway, or direct epistasis where mutations interact at a physical level. Viruses also appear to use co-occurrence particularly within antigenic epitopes to evade host immunity[249]. Single antigenic mutations in the Haemagglutinin (HA) protein of Influenza are often rare, with subsequent or

tandem mutations more often required to make meaningful change[250]. Mutations can alter the phenotype or function of a protein by perturbing epistatic interactions between sites or even between other molecules. For example, the K65R and M184V mutations found in the reverse transcriptase of HIV-1 often co-occur as a mechanism for resistance to the use of nucleotide reverse transcriptase inhibitors[251]. These interactions can have synergistic effects, where both mutations are required to alter the phenotype in an advantageous manner for the virus, or antagonistic effects where the presence of both mutants result in a deleterious phenotype. The paired 69 and 70 deletion in SARS-CoV-2 is thought to compensate for immune evasive mutations in the RBD since it improves virus infectivity[252,253]. This could be viewed as synergistic since the virus balances out the infectivity decrease likely gained from the immune evasive mutations. In the former case, the complementary phenotype can be selected for and may emerge via multiple independent acquisitions of the same mutations. Continuing with the SARS-CoV-2 example, this may explain why the 69/70 deletion appears to cycle with new variants[252], particularly after acquisition of RBD changes.

This is similar to cancer where co-occurring genes would be detected across multiple patients or cancers[247]. Where co-occurrence produces a deleterious pair, this may be observed via mutual exclusion i.e. where both mutations independently occur but almost never together[247]. Given the evidence for co-occurrence of mutations within viruses, and the usefulness of co-occurrence and mutual exclusivity-based analysis in cancer, we decided to identify significantly co-occurring mutations within the SARS-CoV-2 phylogeny and investigate their impact on the virus. Due to the nature of co-occurrence and

mutual exclusivity, many samples are required to detect whether mutations are significantly exclusive or co-occurring. With the number of unique sequences and lineages, the SARS-CoV-2 pandemic has produced a dataset of unprecedented size that should allow the detection of significant co-occurrence and mutual exclusivity among mutations. Using the Weighted Sampling Based Mutual Exclusivity (WeSME)[245] pipeline that has been used to identify co-occurring and mutually exclusive genes in cancer, we us it to identify a number of interesting mutational pairs both within the same ORF and across the SARS-CoV-2 genome. WeSME identifies mutually exclusive and co-occurring genes using a weighted sampling approach in order to navigate around the high computational cost needed to comprehensively identify these relationships i.e. checking every sample and checking every co-occurrence or mutually exclusive partner in each sample and between samples. We modified the pipelines input to work with SARS-CoV-2 mutations and lineages rather than cancer driver genes and samples. Using the identified co-occurrences, we further investigate how these pairs are located across the viral phylogeny, determine their mutational contexts, and use large protein language models to determine learn more about the possible epistatic interactions between these mutation pairs. The results demonstrate how this combination of tools can identify co-occurring mutations of interest in a virus and how we can investigate them further to determine their effect.

4.3 Methods

4.3.1 Tree-based Reference Generation

Tree based reference generation is described in Chapter 2. Briefly, an alignment of ~13 million SARS-CoV-2 sequences was extracted and aligned

using data from GISAID. For each Pango lineage, the frequency of each nucleotide is recorded, with nucleotides reaching >75% being called as the reference sequence. If no reference nucleotide is called, the closest parental lineage base is used. This produces a full sequence for each Pango lineage. These reference sequences are then used for the remainder of the analysis.

4.3.2 EXTRACTING MUTATIONS

Mutations can co-occur either via emergence at the same time in the same virus or via acquisition over subsequent generations. Those that occur across generations can either be functionally unrelated or indicate convergence to a functional change. Early mutations co-occur with more mutations over time, yet these new cooccurrence are less likely to have a functional signal the more separated over generations the mutations are. Mutations that repeatedly emerge together might indicate a functional requirement for co-occurrence where absence of either mutation may be deleterious but presence of both is beneficial. For this study, when looking at co-occurring mutations, we use the normalised mutations i.e. mutations that have occurred in a lineage and not present in the most recent parental lineage. This ensures that mutations that co-occur do not co-occur via inheritance. The jaccard index was used to measure the overlap between pairs of mutations in the dataset. A jaccard index of 1 means that the mutations co-occur with each other in every lineage they are observed in. A jaccard index of 0 indicates that the mutations are never observed together within the same lineage. The p-value threshold was set to 0.01, with a jaccard index of 1 for co-occurrence (i.e. always cooccurring). This was repeated for both the nucleotides and the amino acid levels.

4.3.3 Calculating Co-occurrence

We calculated the co-occurrence of mutations using the WeSME python package[254] which was initially used for estimating co-occurrence and mutual exclusivity of driver genes within cancer samples. WeSME works by using gene annotations from different cancer samples, however, we modified the input so that mutations were used instead of gene names, and SARS-CoV-2 lineages were used instead of sample names. Due to the computational inefficiency of checking every combination of co-occurring mutations. WeSME uses a weighted sampling approach derived from mutational frequencies to efficiently check whether co-occurrences are significant, without having to check all combinations in a permutation-based manner. WeSME can calculate both mutual exclusion and co-occurrence using this method, although in this chapter we look at co-occurrence.

4.3.4 Annotation of Mutations

Mutations were annotated at multiple levels to identify patterns of co-occurrence and mutual exclusivity that occur beyond just the nucleotide level. These include the open reading frame (ORF), the domain, a sliding window position within the domain, and the amino acid position at which the mutation occurred in.

4.3.5 Language model Epistasis

Language model scores were derived using the ESM-2 1-billion parameter model. Using the same process as in section 3.4.4, we mutated a site in the protein of interest and measured the change in the likelihoods between the reference and the mutant sequence. The difference in likelihood was then plotted onto a structure of the protein and coloured depending on whether the probability increased or decreased.

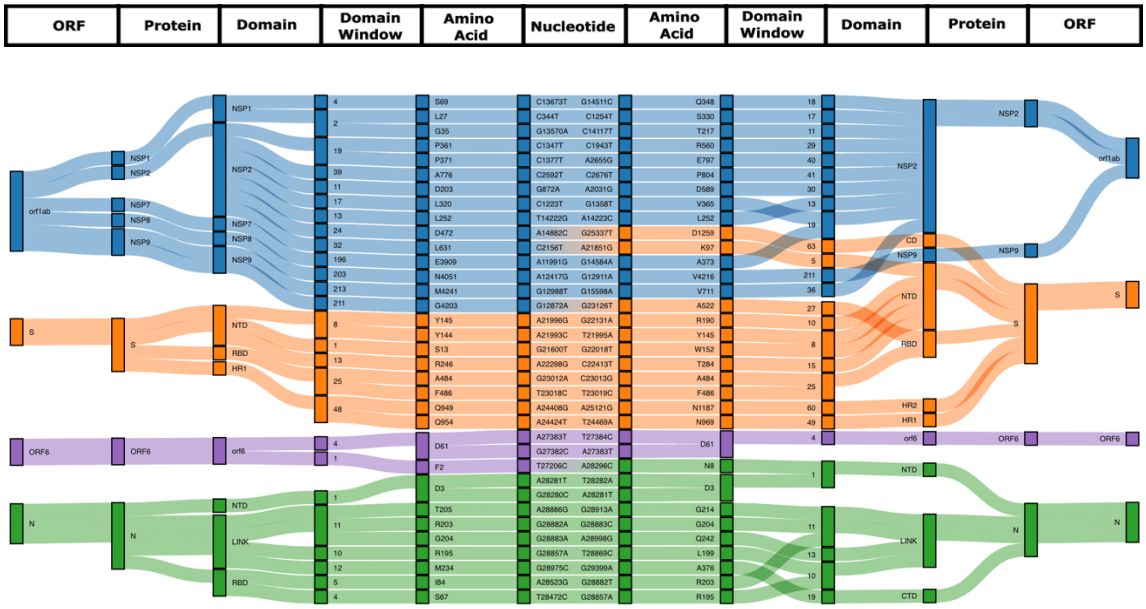
4.4 Results

Using all 13 million SARS-CoV-2 sequences would be computationally intractable. Instead, the representative tree-based lineages sequences of each of the Pango lineages created for the mutational signatures from Chapter 2 were used and their mutations counted. The mutations were normalised i.e. they were counted against a tree-based reference to remove inherited mutations. As such, these mutations always arise together within the same lineage.

4.4.1 Co-Occurrence

For the nucleotide mutations set, 35 pairs of significant co-occurring mutations were identified that have only occurred together within a lineage (Figure 4.1A). Of these, only 6 occurred more than once and these pairs were predominantly in N and ORF6, with 1 in Spike and 1 in NSP2 of ORF1a/b (Figure 4.1B).

A



B

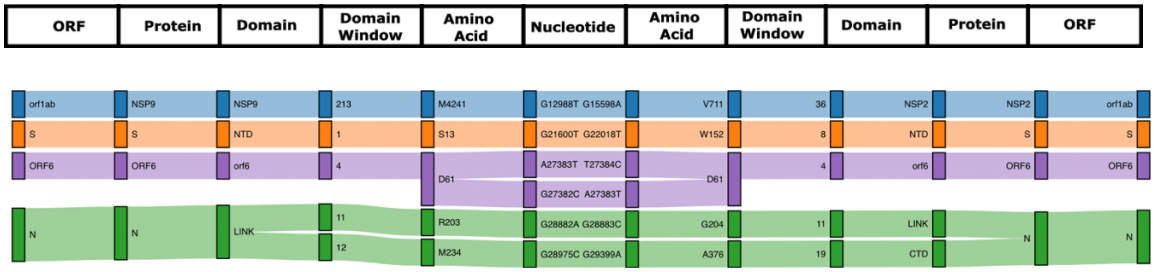


Figure 4.1: (A) Sankey plot showing the co-occurrence between nucleotides in SARS-CoV-2 excluding inherited mutations. (A) shows the nucleotide level co-occurrences. (B) shows those nucleotide co-occurrences that occurred more than once. Each column is an annotation of the nucleotide mutation. Where annotations do not exist for a column, the above annotation is inherited (the S ORF is also the S protein). Domain windows are defined as 20 amino acid blocks of a given domain.

Between ORF co-occurrences were less common, with only 3 co-occurrences at the nucleotide level which did not happen more than once in the tree-based lineages. These occurrences were predominantly between ORF1a/b and Spike, with 1 other between ORF6 and N (Figure 4.1A and B).

4.4.2 Nucleotide contexts of Mutations

After determining which nucleotide mutations co-occur across the SARS-CoV-2 genome, we investigated their nucleotide contexts for evidence of mutational processes preferentially producing co-occurring mutations. A→G mutations were most prominent at CAA, GAA, and AAT contexts (Figure 4.2A). The T→C mutations (which on the negative strand are equivalent to A→G) showed biases at TTT and ATT contexts. Interestingly, this ATT context is equivalent to the AAT context on the positive strand. C→T mutations and their negative strand equivalent G→A mutations were also abundant in the normalised set. GCT and TCT mutations were the most abundant C→T mutations, while GGG, TGG, TGT and AGA were all popular contexts for G→A mutations. G→T mutations were also present although no preferred context was observed including in the for negative-sense C→A substitutions. Other substitutions were less frequent.

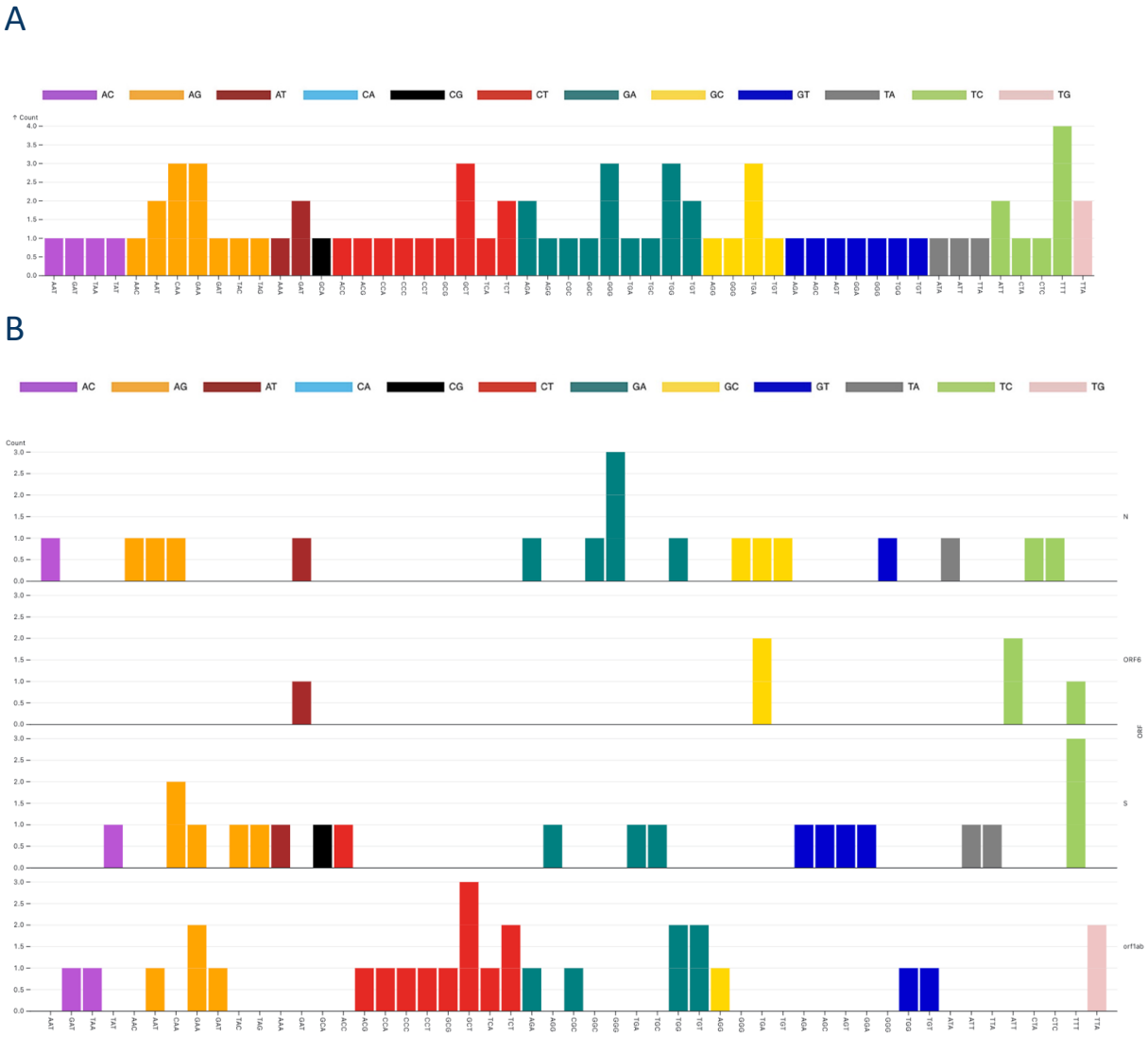


Figure 4.2: (A) Mutational composition of co-occurring nucleotides. (B) Mutational composition of co-occurring nucleotides split by ORF.

Stratifying substitutions by ORF (Figure 4.2B), we see different distributions of substitution types and contexts. There appears to be little observable preference of substitution or context to a given ORF, with ORF1a/b and Spike having the most substitutions as expected due to their size. The Nucleocapsid protein (N) does have a number of mutations despite being smaller. Interestingly, C→T substitutions are seemingly rare in ORFs outside of ORF1a/b.

4.4.3 THE LINEAGES OF CO-OCCURRING MUTATIONS

Using the co-occurring mutations annotated onto the SARS-CoV-2 tree, we can see how these mutations are distributed through the phylogeny.

The ORF1a/b co-occurrences only appears in sub-lineages of B.1.258. It's possible therefore that this co-occurrence is an artifact rather than a true co-occurrence (Figure 4.3). Using Outbreak.info[255] to check using up-to-date GISAID sequences, B.1.258 contains both mutations, suggesting that the tree-reference may not have had enough sequences to make the correct base call for this position. The Spike mutation appears in 2 sub-lineages of B.1 although the B.1 lineage does not contain these nucleotide substitutions either in the tree reference or in Outbreak.info (Figure 4.3). As such, these mutations do appear to be significant co-occurrences. The N mutations have two different lineage distributions (Figure 4.3). The first co-occurring mutations (G28882A and G28883C) are beside each other but are within 2 different codons of the N protein. This change appears in 6 distinct lineages, again suggesting this is a real significant mutation. One of these lineages B.1.1 is a precursor to all the current VOC lineages. The second N co-occurring pair is found in two lineages B.1.160 and B.1.1.445, although neither lineage preceded nor arose from a VOC lineage. The ORF 6 co-occurrence pairs are all within the same codon and both co-occur with the A27283T mutation and result in the D61L. Both lineages containing the co-occurrences are major Omicron variants BA.2 and BA.4.

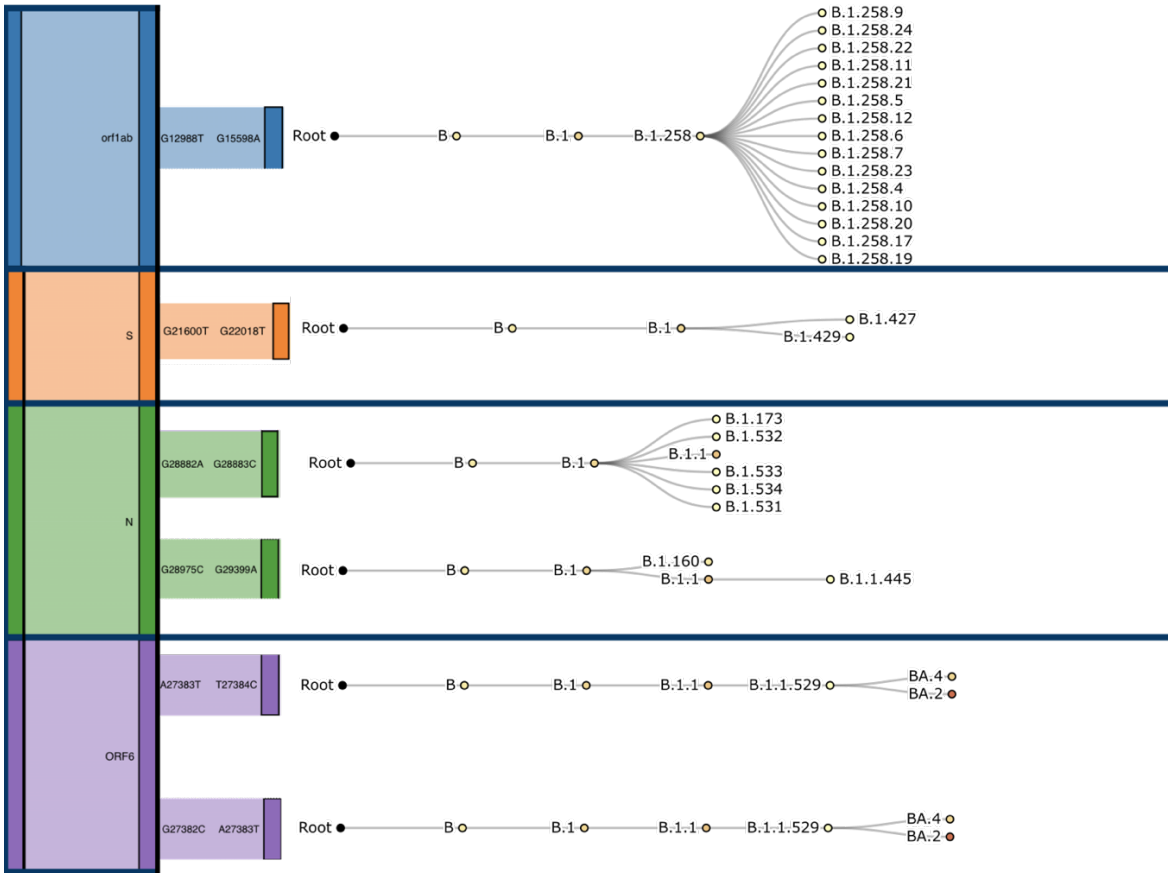


Figure 4.3: Lineages of SARS-CoV-2 that contain co-occurring mutations. Co-occurring mutations filtered by at least 2 co-occurrences. End node lineages represent those where the co-occurrences appear.

4.4.4 Language Model Epistasis of Co-occurrences

Delving into the effects of co-occurrence on the proteins they occur in, we looked at the amino acid substitutions that resulted from co-occurring nucleotide mutations. We selected the N:G28882A and the N:G28883C mutations since they occurred within the same protein and independently arose in a number of lineages. We measured the effect that both substitutions individually had as well as the effect that both substitutions combined had on the reference probabilities of the other positions and filtered positions that had greater than a 1% difference.

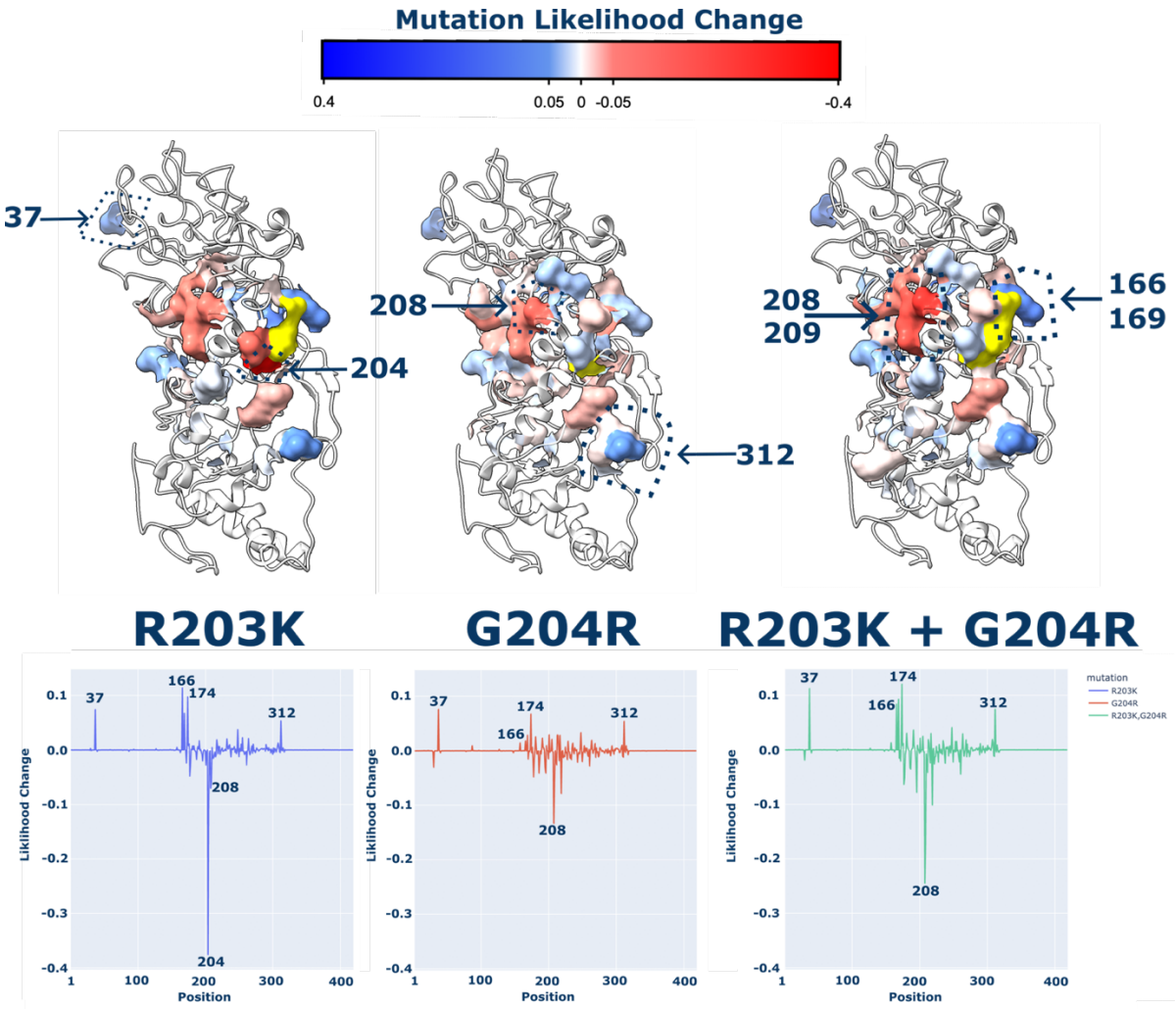


Figure 4.4: Structures of the N protein coloured by the change in likelihood cause by co-occurring substitutions. Mutated positions are coloured yellow, positive likelihood changes are coloured blue while negative likelihoods are coloured red.

The R203K mutation on its own causes a large negative drop in the likelihood of G204 by almost 40%, suggesting a massive decrease in the compatibility of these 2 positions (Figure 4.4). This suggests that the G204R mutation is compensatory due to a large incompatibility between the K and G amino acids. Due to the linker regions flexibility[256], it is hard to know from a structural perspective how these mutations may directly affect one another, since this flexibility could lead to incompatibilities not observed in the predicted

structure. However, the region appears to be arginine rich[256], which might be why the loss of arginine at 203 is accompanied by a gain in arginine at 204, and why the position becomes less favourable to remain as a glycine following R203K. The arginine density also make the region more alkaline than the other domains of the protein, and its basic nature make it a likely RNA binding region[256], which may again explain the co-occurrence of the loss and gain of the arginine. The other position that is hampered by all substitutions is 208, which decreases in probability after both substitutions but is particularly impacted by the combination of R203K and G204R. The decrease at 208 and 209 is less severe than at 204 and is at a different surface of the protein which may be why it is accommodated. A number of positions including 37, 166, 169, 174 and 312 all increase their probabilities following any of the mutation combinations.

4.5 Discussion

Despite looking at a very stringent category of co-occurrence (i.e. co-occurrence without inheritance) we have identified several co-occurring mutations across the SARS-CoV-2 phylogeny and sequence. The dominance of ORF1a/b and Spike mutations is expected given their sizes (Figure 4.1A), however the lack of multiple co-occurrences of these mutations is interesting, particularly since ORF6 and the N protein have multiple mutations of this type (Figure 4.1B). The ORF1a/b co-occurrences seemed to be condensed within NSP2 which appears to have a variety of functions including repressing interferon by binding to interferon mRNA[257] to prevent translation and promotes translation of viral replication proteins by through interactions with host translational mechanisms[258]. The mutations appear to be distributed throughout NSP2,

which is interesting given NSP2's variety of possible functions since this might hint at multiple functional domains. However, only one of these NSP2 involved mutations appears in the sequence set multiple times although this is most likely linked to an artifact of processing. Spike co-occurrences were primarily located in the NTD which likely suggests an immune evasion role given the NTD is home to an antigenic "supersite"[223]. 5 of the co-occurring mutations lie within this supersite at positions 13, 144, 145, 152, 246 and the only remaining co-occurrence with multiple instances is between 2 mutations that both lie within the supersite (between S13 and W152). The multiple co-occurrences re-affirm the importance of this domain and suggests that a number of different mutations in this site can be beneficial. Another interesting co-occurrence is within the HR1 domain and involves the Q954 and N969 positions. As mentioned in Chapter 2, the N969K mutation is an impactful substitution that appears to have wide ranging effects on the protein according to ESM-2 likelihood changes (Figure 3.6). Its co-occurrence with the Q954H mutation at the basal Omicron lineage B.1.1.529 and presence within the fusogenic region of the Spike made these 2 mutations possible candidates for Omicrons change in entry phenotype[227]. Its presence in ~99% of lineages (checked using CoV-Spectrum[259]) following 2022 and Omicrons emergence suggest that these mutations may be important for function of the variant. The Nucleocapsid protein has several co-occurring mutations predominantly within the linker domain of the protein. The N protein is primarily used to package the viral RNA into the virion[260] and has 2 main structural domains at the NTD and CTD of the protein. The NTD section binds to RNA while the CTD primarily allows for dimerization since it binds to other N proteins. These

functional domains are connected by a disordered region known as the linker which contains a domain for post-translational modification to alter function and also keeps the domains separated from each other[260]. Mutations in this region (particularly within the 199 and 205 amino acid stretch) have been show to produce remarkably large differences to the number of viruses produced in an infected cell, increasing by as much as 50 fold in the presence of the R203M mutation[261]. A number of the co-occurring sites detected occur within this section of the linker region, in particular positions 203, 204 and 205.

ORF 6 has 2 amino acid mutations and 5 nucleotide mutations that co-occur. This is somewhat surprising give the proteins small size of only 61 amino acids. The protein is thought to function as an interferon antagonist[191,262], so mutations in the protein may help with viral fitness in improving this antagonism.

Co-occurrence contexts show a similar picture to what was observed in Chapter 2 with most mutations resulting from Signature 1, 2 or 3 substitution types (Figure 4.2 and Figure 2.8). Breaking this down by ORFs did not produce any strong associations between substitution category and ORF, albeit ORF1a/b did contain the majority of C→T mutations but this is more likely due to its size (Figure 4.2B).

Digging into where the co-occurrences appeared throughout the SARS-CoV-2 phylogeny shows that 2 of the ORF6 and 1 of the N mutations that appeared twice occurred either in VOC sequences or in sequence that preceded them (Figure 4.3). The ORF 6 mutations are all within a single codon causing the D61L substitution and emerged in the BA.2 and BA.4 Omicron lineages. BA.2

derivative lineages make up most of the currently circulating sequences with 97% of sequences in 2024 containing the D61L mutation due to the presences of the JN lineages which are a BA.2 derivative[259]. Despite this, it appear the D61L mutation actually severely impairs the functioning of ORF6 when binding to the nuclear pore proteins Nup98-Rae1[263]. This inhibition of binding subsequently prevents reduces ORF6s ability to inhibit interferon signalling via the JAK-STAT pathways and as such contributes to an increased interferon response relative to non-mutated ORF6. BA.2 and BA.4 are highly antigenically distinct from pre-omicron SARS-CoV-2 variants, so it is possible that while the D61L mutation should increase the interferon response (and likely decrease viral fitness), the advantages gained elsewhere in the variants may allow D61L to be accommodated. Alternatively, D61L may have some other selectively beneficial effect that has not yet been discovered.

The G28882A and G28883C co-occurrence in the Nucleocapsid produce the R203K and G204R mutations in the linker region of the nucleocapsid and this occurs in the B.1.1 lineage as well as a few other smaller lineages (Figure 4.3). B.1.1 is an important lineage since Alpha, Gamma and Omicron are all descendant from it. The co-occurrence within the nucleocapsid mutational hotspot[261], as well as its inclusion in most of the VOCs make this co-occurrence particularly interesting. The combined mutations appear to improve viral fitness and increase infectivity, with the cause associated with the viruses increased replicative efficiency[256]. While Delta and Beta do not share these mutations, both contain similar substitutions (R203M in Delta and T205I in Beta) which adds further evidence for a strong selective pressure on the sites there.

Language modelling of the mutational epistasis between the R203K and G204R mutations appears to show that the R203K mutation is very poorly accommodated by the neighbouring G204 which shows a nearly 40% drop in likelihood (Figure 4.4). This might indicate why these mutations occur together, with the G204R mutation being required to accommodate the more impactful 203 substitution. A number of other sites do increase in their likelihoods including a number of positions in the NTD (37, 166, 169, 174) and the CTD (312). Most minor likelihood fluctuations occurred within the linker region, however increases in the NTD likelihoods could indicate why these mutations in this region appear to improve RNA packing[261] since the NTD is involved in binding the RNA[260]. The 208 and 209 positions of the N protein take a larger decrease in likelihood as a result of the mutations (particularly R203K) however this lies outside of this highly mutable region and as such may not be as necessary for the improved functional phenotype.

Using co-occurrence to investigate important SARS-CoV-2 mutations is clearly a useful methodology. Combined with well annotated trees and powerful language model metrics, co-occurrence analysis can uncover interesting new details about how and why SARS-CoV-2 evolves in the way it does. While we only looked at the most stringent of co-occurrence categories in this analysis, expanding the analysis to incorporate and investigate convergence of mutations across viral generations may give an even better insight into the mutational pressures and phenotypes of co-occurring mutations. Much more work can be continued here in the future to make this analysis more viable, practical and useful.

Firstly, expanding the analysis to include mutually exclusive relationships is important. There is a rich history in cancer showing the importance of mutual exclusion in driver genes. Given the similarities between cancer and viruses on co-occurring mutations, it seems possible that a similar pattern will be observed with mutually exclusive viral mutations. Evidence already exists for mutually exclusive mutations in SARS-CoV-2, with the R346 and N450 sites remaining exclusive in 6 million different Omicron sequences[264].

Next, while we have annotated each mutation with its structural context (amino acid change, domain, ORF etc), we have not investigated further whether co-occurrence (or mutual exclusivity) resides at levels above the nucleotide. While nucleotide mutations are the basic molecular unit, co-occurrence might only provide a functional consequence at the amino acid level or higher. We know that convergent sites have mutated to different amino acids (R203K/M in the Nucleocapsid[256], E484A/K/Q in the RBD[225]), yet by looking at co-occurrence at the nucleotide level we may miss identifying co-occurrences that occur at these higher levels. By expanding to these higher levels, we can try to find the "meaningful" level of co-occurrence and mutual exclusivity i.e. do mutations simply have to co-occur at the same sites, same domains, or even the same ORFs to have an effect.

Another critical area of interest is looking at co-occurrence that includes inheritance. Our analysis aims to identify "interesting" co-occurrence, i.e. co-occurrence that results because of the selection associated with a gained fitness advantage or new phenotype. Most mutations in the SARS-CoV-2 genome that co-occur are likely to be functionally independent, so identifying

co-occurring pairs that are not is a challenge. However, doing so will uncover co-occurrences that the current pipeline ignores.

Lastly, making the analysis more general purpose so that it can be used to analyse other viruses is an obvious next step. SARS-CoV-2 benefits from a wealth of sequencing and research, however there is likely enough data for viruses like HIV and Influenza that expanding the analysis is warranted.

5 Conclusion

5.1 Discussion

This thesis is a demonstration of how inter-disciplinary thinking and collaboration can provide novel insights across scientific fields. It began as an experiment to see whether techniques from cancer genomics and machine learning could be translated over into virology despite the many significant differences between these three areas of research. It is apparent that there are several areas where these domains overlap and that many useful insights can come from combining ideas across disciplines.

Like many of the best laid plans, many of these approaches expanded to become much larger parts of this thesis than expected. The initial idea for Chapter 2 to uncover signatures of mutational processes in SARS-CoV-2 started out as a master's project in 2020 where I applied a pre-existing tool used to find signatures in cancer genomes to the newly emerged viral sequences from SARS-CoV-2. It quickly became evident that because of viral specific factors such as transmission pre-existing tools would be unlikely to extract meaningful mutational signatures, hence began a project that continued in my PhD. We contributed to evidence that there are three predominant mutational processes that are producing mutations in SARS-CoV-2, characterised these processes as mutational signatures (Figure 2.8 and Figure 2.9) and demonstrated that these processes are not static but have dynamic activities throughout the pandemic (Figure 2.10). We see that different lineages appear to have different exposures of each signature, which may be due to phenotypic differences (Figure 2.12). We related these

signatures to the amino acid substitutions that were likely produced by them and show how these mutational processes are likely to be the primary source of SARS-CoV-2's evolutionary capacity (Figure 2.13).

Language modelling is another domain which has quickly expanded outwith its initial domain of natural language processing and into biological data. Initially a subsection of Chapter 2, this expanded into its own chapter once we appreciated the possibility's language models offer. We followed the study by Hie et al[193] and generated semantic and grammatical scores using the larger amino acid based protein language model ESM[100,118]. We did this for every possible amino acid in the spike protein of SARS-CoV-2 and showed that language model scores capture important properties of the protein such as conservation (Figure 3.4). We took this further showing that model embeddings appear to "understand" the differences between the major SARS-CoV-2 VOCs (Figure 3.2) and can be used to estimate an evolutionary pseudotime that correlated with real lineage emergence dates (Figure 3.2B and Figure 3.3). We then demonstrated that the embeddings and logits could be used to train effective predictors for a variety of biologically meaningful variables, indicating that these representations are meaningful and predictive (Figure 3.10). We show that using language model likelihood changes following inserted mutations, we can measure the possible epistatic effects the mutation has on surrounding amino acids (Figure 3.6). The opportunity to model epistasis computationally based in single sequences is interesting, particularly given much of the change we observe in the virus over time is stepwise. We can check these mutations to determine possible future mutations, and supplemented with prior DMS scores can evaluate the potential impact of

current and future variants. Using the whole protein DMS, we also observed sites that were frequently affected by mutations at multiple sites across the protein (Figure 3.8) which we hypothesise to be positions of structural or functional constraint within the protein. These positions appear to be linked to amino acids with important structural features such as cysteine and proline (Figure 3.9), again indicating at these positions are structurally constrained. With the models understanding of the protein having been established, we presented the idea of using the language models in a surveillance setting, processing each new haplotype and generating language model scores for each (Figure 3.11). We showed that averaged embeddings could be used to represent a circulating virus background to measure against, supplementing the original Wuhan-Hu-1 reference which has long since become extinct (Figure 3.11). This chapter aimed to show how language models even in this early state can be used to answer interesting biological questions. Language model scores represent a new way of measuring differences between sequences, embeddings and likelihoods make for powerful predictors and likelihood changes offer interpretable insights into epistatic interactions and their effects on mutations across the protein sequence.

Finally, we investigated the phenomenon of mutational co-occurrence and how this manifested across the SARS-CoV-2 phylogeny during the pandemic. We narrowed in on a particular type of co-occurrence where mutations always co-occur together. We discovered many co-occurring mutations across the phylogeny and gathered information on their locations at various levels such as nucleotide, amino acid, domain and ORF (Figure 4.1). We then investigated whether these mutations shared similar contextual patterns, and if this could

be explained by regions of the genome such as ORF (Figure 4.2). We further investigated the co-occurrences that emerged multiple times in the SARS-CoV-2 phylogeny, and found that the mutations in ORF6 and N are both present in VOC sequences or VOC precursor lineages (Figure 4.3). We then utilised language models to predict how the N protein mutations might interact with each other and other positions in the protein and discovered that the G204R mutation may be important for accommodating the R203K substitution (Figure 4.4). We showed in this chapter that by investigating co-occurring mutations, we can discover interesting mutational interactions that hint at possible functional consequences for viral proteins. We think that in future, this pipeline could be a viable way of investigating what co-occurring mutations mean for the lineages that contain them.

5.2 Future Work

The current state of machine learning for use in virology has never been more exciting. During my PhD, a number of transformative resources and tools have been released which have now become almost as ubiquitous in use as BLAST. First, the release of AlphaFold[99] transformed structural biology allowing for de-novo predictions of structures. The ability to fold proteins is powerful since the 3D structure is often what drives protein function and is a basic requirement for computational docking experiments that are used to help discover new drugs. AlphaFold has since improved to predict not only monomer structures, but protein-protein complexes and missense mutation. The ability to predict structures computationally is of critical importance especially in understudied areas of biology. Structural biology is expensive and

time consuming, while AlphaFold can generate structures with little cost and much quicker, making structural biology accessible to many more researchers. A further bonus of larger structural databases is that structural homology can be better understood. The release of Foldseek[88] allows these questions to be answered at scales that simply were not possible prior to its release. Foldseek allows users to search for structural similarities in much the same way as we traditionally do with sequence similarity. For virology, this is important since many virus proteins differ enormously at the sequence level but have strong structural conservation that allows for similar function. Viruses are known to mimic host proteins in order to hijack host cellular machinery, meaning that structural homology of viral proteins can also be used to help determine their protein interactions and possibly functions.

Biological language models have achieved widespread adoption over the last few years. Initially these were predominantly protein language models, however with increasing model parameter counts, context lengths and expanding datasets, there are now language models for DNA, RNA and even single cell datasets[265–267]. These models are in their infancy, yet the rapidly developing capabilities in the natural language field suggests that there is a lot more potential for these models to improve in coming years. Within the protein language domain these models are now being picked apart to determine what they learn, how they do it, and what they can be best used for. Work from the Li et al.[268] has shown the models appear to scale efficiently for learning structure based tasks but appear to plateau on more functional prediction tasks. This might suggest that the masked language modelling objective used to train these models could be reaching its limit for learning

these properties. However, it presents an opportunity to investigate other training objectives such as next sentence prediction or using an expanded set of mask tokens that may help the model learn these interesting properties of sequence. The ProteinBERT model from Brandes et al.[269] for example add annotations to the training input so that their model learns to link sequence to functional annotations. Work from Zhang et al.[270] has shown that performing operations directly on model logits produces interpretable contact maps of the sequence, showing that the model probabilities themselves encode structural information about sequence without the need to train an additional regression head. This also complements the work from Li et al.[268] since it shows that the predominant features that are learnt by the model weights appear to be structurally important motifs or co-evolving positions rather than more abstract concepts such as the energy functions or folds of the protein. Insights such as these expose the current limitations of these models, which in future will be areas in which improvements can be made. ESMFold[100] still produces on average less accurate structures than AlphaFold, but critically can make structures without alignments a major limitation of AlphaFold. Where known sequence homology is limited, ESM-like folding tools offer impressive folding capabilities. The ESM Metagenomic Atlas built using ESMFold produced structures for over 700 million protein sequence, vastly expanding the number of total protein structures and offering a new structural insight into vast quantities of metagenomic data. Chapter 3 covered many of the applications of language models even before their use to produce protein structures, yet there are still many things that can be investigated in future to improve them. Understanding more about how finetuning based on specific viral species (i.e.

Influenza, Coronaviruses etc) or specific proteins like glycoproteins might change the model is an open question. Determining the extent to which language models understand protein-protein interactions when embedded together or individually is another important area that needs to be explored, particularly since AlphaFold 3 now appears to predict complex protein-molecular interactions. New model architectures such as Mamba[271] and Hyena[272] permit large extensions to the context window of language models, yet it is not clear that the efficiencies afforded by there architectures are worth the trade-off to interpretability that are found with traditional attention-based mechanisms. Evo[123] is a nucleotide model based on a striped-hyena architecture can reach whole genome level contexts, yet despite this enormous context increase in many tasks is comparable with models orders of magnitude smaller. However, its ability to predict gene essentiality in *E. coli* is something that previously could not be done in the same way. I think it is highly likely that over the coming years, we will see remarkable advances in the abilities of these models, and hopefully a sweet spot between context length, performance, and ease of use.

The work of this thesis lays the groundwork for a multitude of different future directions. The tools developed as part of Chapter 2 to uncover mutational signatures can be readily implemented in the event of a future pandemic. Given the usefulness of genomic sequencing during the SARS-CoV-2 pandemic and the increased accessibility of sequencing machines across the world, it seems highly likely that future pathogen outbreaks will be extensively sequenced. With ever larger sequencing databases, the viability of observing mutational signature patterns increases and so too does its usefulness as a

technique to identify mutational processes that enable viral evolution.

Language modelling continues to progress at breakneck speed, leaving several avenues of investigation open including fine-tuning new viral specific models, experiential validation of model predicted epistatic interactions, and expanding beyond protein models to look at RNA and DNA effects.

Investigating putative evolutionary trajectories is another possibility since language models allow for quick inferences on the probability of different sequence outcomes. With tools like Evo-velocity, and even using the logits or embeddings straight from the model, it may be possible to map out future trajectories that the virus may take. Unintuitively, we know this is not always the most likely mutations the language models suggest since these models are unaware of the immediate selective environment the proteins are in from the host immune system. As such, developing ways to supplement these models with this information (potentially via fine-tuning) could allow for predictive fitness landscapes. Looking at co-occurrence and mutually exclusive sets of mutations may help here, as these mutations often appear to require the other mutation due to intense selective pressure or exclude it from ever occurring. Uncovering the mechanisms behind these patterns can give us a better understanding of the selective landscape as well as the protein constraints. Expanding the analysis to incorporate co-occurrence across different generations of the virus will also be important, since convergence of mutations clearly indicates useful adaptations that should be investigated further.

To conclude, the work completed during this PhD will continue to be developed further in the coming years. Incorporating techniques and methods from machine learning has allowed me to answer interesting questions about the

evolution of SARS-CoV-2 throughout the pandemic. I believe that the work here forms a basis for several projects in the future, which I hope to pursue. These include further questions about SARS-CoV-2, but also how these methods can be applied to other viruses, bacteria or even cancer.

Protein language models are easily transferable to different domains and are often trained in a manner that allows this. ESM-2[100] is not a virus specific PLM, yet we used it to accomplish a number of virus specific tasks. While fine-tuning PLMs may help them to better perform in different domain areas, many of the base models remain somewhat agnostic to the proteins they are given. This means that we can begin to perform in-silico PLM experiments on cancer genes like BRCA1 or TP53 which might help us to reason with their mutations often being drivers for cancer progression[273]. Antimicrobial resistance (AMR) is a huge emerging issue since it drastically harms our main line of defence against bacteria (antibiotics). PLMs are already being used to help identify the resistance causing bacterial genes[274], however there are opportunities to also perform in-silico DMS to help identify genes that can best be targeted by new or existing antibiotics. In-vitro DMS has already proved useful in identifying key bacterial proteins for antibiotic targeting[275], so scaling up this type of analysis to scan larger numbers of bacterial proteins could prove increasingly useful as AMR increases in prevalence. Again, the pairing of this type of DMS analysis with co-occurrence and mutual exclusivity investigations as we demonstrated in Chapter 3 with the nucleocapsid could be used to identify mechanistic reasons for the acquisition of AMR by bacterial genes where single mutations are not an obvious cause.

Looking back to viruses, I find there is an exciting opportunity with PLMs and generative modelling to ask whether we can make actionable predictions of evolutionary trajectories. Evo-velocity is not currently used to predict forward trajectories, yet thinking of PLMs as a sequence context aware version of transition matrices like the BLOcks SUbstitution Matrix (BLOSUM) could allow us to simulate possible forward steps in an evolutionary process. Combining this with the work on mutational processes in Chapter 2 could allow us to estimate likely trajectories not just at the nucleotide level where mutational processes generate possible viable genomes, but at the amino acid level where selection on the functional protein may dictate whether those genomes are viable or fit for function. Again, while SARS-CoV-2 is a blueprint for what virus datasets should look like, we should push this work into other viral datasets to show whether these approaches offer a generalised way to looking at evolution from a machine learning perspective.

There is much more work to be done, and I look forward to continuing with it.

BIBLIOGRAPHY

1. Freeth, T. *et al.* A Model of the Cosmos in the ancient Greek Antikythera Mechanism. *Sci Rep* **11**, 5821 (2021).

2. Turing, A. M. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **s2-42**, 230–265 (1937).

3. Al-Hashimi, H. M. Turing, von Neumann, and the computational architecture of biological machines. *Proceedings of the National Academy of Sciences* **120**, e2220022120 (2023).

4. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115–133 (1943).

5. Koonin, E. V., Dolja, V. V. & Krupovic, M. The logic of virus evolution. *Cell Host & Microbe* **30**, 917–929 (2022).

6. Stern, A. & Andino, R. Viral Evolution. *Viral Pathogenesis* 233–240 (2016) doi:10.1016/B978-0-12-800964-2.00017-3.

7. Murray, S. M. *et al.* The impact of pre-existing cross-reactive immunity on SARS-CoV-2 infection and vaccine responses. *Nat Rev Immunol* **23**, 304–316 (2023).

8. Yang, X. *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine* **8**, 475–481 (2020).

9. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine* **382**, 1199–1207 (2020).

10. Gelderblom, H. R. Structure and Classification of Viruses. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, Galveston (TX), 1996).
11. Taylor, M. W. What Is a Virus? in *Viruses and Man: A History of Interactions* (ed. Taylor, M. W.) 23–40 (Springer International Publishing, Cham, 2014). doi:10.1007/978-3-319-07758-1_2.
12. Creager, A. N. H., Scholthof, K.-B. G., Citovsky, V. & Scholthof, H. B. Tobacco Mosaic Virus: Pioneering Research for a Century. *The Plant Cell* **11**, 301–308 (1999).
13. Beijerinck, M. W. *Über Ein Contagium Vivum Fluidum Als Ursache Der Fleckenkrankheit Der Tabaksblätter*. (Müller, 1898).
14. Telenti, A. *et al.* After the pandemic: perspectives on the future trajectory of COVID-19. *Nature* **596**, 495–504 (2021).
15. Horimoto, T. & Kawaoka, Y. Influenza: lessons from past pandemics, warnings from current incidents. *Nat Rev Microbiol* **3**, 591–600 (2005).
16. Landovitz, R. J., Scott, H. & Deeks, S. G. Prevention, treatment and cure of HIV infection. *Nat Rev Microbiol* **21**, 657–670 (2023).
17. Plowright, R. K. *et al.* Pathways to zoonotic spillover. *Nat Rev Microbiol* **15**, 502–510 (2017).
18. Harvey, E. & Holmes, E. C. Diversity and evolution of the animal virome. *Nat Rev Microbiol* **20**, 321–334 (2022).
19. Jones, R. A. C. Global Plant Virus Disease Pandemics and Epidemics. *Plants* **10**, 233 (2021).

20. Koonin, E. V., Krupovic, M. & Agol, V. I. The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? *Microbiol Mol Biol Rev* **85**, e00053-21.

21. Baltimore, D. Expression of animal virus genomes. *Bacteriol Rev* **35**, 235–241 (1971).

22. Payne, S. Introduction to RNA Viruses. *Viruses* 97–105 (2017) doi:10.1016/B978-0-12-803109-4.00010-6.

23. Weber, F., Wagner, V., Rasmussen, S. B., Hartmann, R. & Paludan, S. R. Double-Stranded RNA Is Produced by Positive-Strand RNA Viruses and DNA Viruses but Not in Detectable Amounts by Negative-Strand RNA Viruses. *J Virol* **80**, 5059–5064 (2006).

24. Rampersad, S. & Tennant, P. Replication and Expression Strategies of Viruses. *Viruses* 55–82 (2018) doi:10.1016/B978-0-12-811257-1.00003-6.

25. Balvay, L., Lastra, M. L., Sargueil, B., Darlix, J.-L. & Ohlmann, T. Translational control of retroviruses. *Nat Rev Microbiol* **5**, 128–140 (2007).

26. Siddell, S. G. *et al.* Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV). *Journal of General Virology* **104**, 001840 (2023).

27. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5**, 536–544 (2020).

28. Zerbini, F. M. *et al.* Changes to virus taxonomy and the ICTV Statutes ratified by the International Committee on Taxonomy of Viruses (2023). *Arch Virol* **168**, 175 (2023).

29. Gregory, T. R. Understanding Evolutionary Trees. *Evo Edu Outreach* **1**, 121–137 (2008).

30. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303–314 (2012).

31. Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nat Rev Genet* **21**, 428–444 (2020).

32. Sato, A. *et al.* Phylogeny of Darwin's finches as revealed by mtDNA sequences. *Proceedings of the National Academy of Sciences* **96**, 5101–5106 (1999).

33. Pybus, O. G. *et al.* Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences* **109**, 15066–15071 (2012).

34. Biek, R., Henderson, J. C., Waller, L. A., Rupprecht, C. E. & Real, L. A. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences* **104**, 7993–7998 (2007).

35. Tegally, H. *et al.* Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern. *Cell* **186**, 3277-3290.e16 (2023).

36. Fitch, W. M. Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *Journal of Molecular Biology* **26**, 499–507 (1967).

37. Wakeley, J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution* **11**, 158–162 (1996).

38. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425 (1987).

39. Susko, E. & Roger, A. J. Long Branch Attraction Biases in Phylogenetics. *Systematic Biology* **70**, 838–843 (2021).

40. Felsenstein, J. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Biology* **27**, 401–410 (1978).

41. Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* **5**, 50 (2005).

42. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507 (1976).

43. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

44. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).

45. Wolf, Y. I. *et al.* Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* **5**, 1262–1270 (2020).

46. Simmonds, P. *et al.* Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* **15**, 161–168 (2017).

47. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

48. Pekar, J. E. *et al.* The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science* **377**, 960–966 (2022).

49. Lytras, S., Xia, W., Hughes, J., Jiang, X. & Robertson, D. L. The animal origin of SARS-CoV-2. *Science* **373**, 968–970 (2021).

50. Virology: Coronaviruses. *Nature* **220**, 650 (1968).

51. Almeida, J. D. & Tyrrell, D. A. J. The Morphology of Three Previously Uncharacterized Human Respiratory Viruses that Grow in Organ Culture. *Journal of General Virology* **1**, 175–178 (1967).

52. Temmam, S. *et al.* Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* **604**, 330–336 (2022).

53. Steiner, S. *et al.* SARS-CoV-2 biology and host interactions. *Nat Rev Microbiol* 1–20 (2024) doi:10.1038/s41579-023-01003-z.

54. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nat Rev Mol Cell Biol* **23**, 3–20 (2022).

55. Peacock, T. P. *et al.* The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nat Microbiol* **6**, 899–909 (2021).

56. Bayati, A., Kumar, R., Francis, V. & McPherson, P. S. SARS-CoV-2 infects cells after viral entry via clathrin-mediated endocytosis. *Journal of Biological Chemistry* **296**, 100306 (2021).

57. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**, 14–25 (2018).

58. Woo, H. *et al.* Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *J Phys Chem B* **124**, 7128–7137 (2020).

59. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat Med* **26**, 450–452 (2020).

60. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell–cell fusion but does not affect virion entry. *Virology* **350**, 358–369 (2006).

61. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Molecular Cell* **78**, 779-784.e5 (2020).

62. Takeda, M. Proteolytic activation of SARS-CoV-2 spike protein. *Microbiology and Immunology* **66**, 15–23 (2022).

63. Bhatt, P. R. *et al.* Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome. *Science* **372**, 1306–1313 (2021).

64. Finkel, Y. *et al.* The coding capacity of SARS-CoV-2. *Nature* **589**, 125–130 (2021).

65. Sun, Y. *et al.* Restriction of SARS-CoV-2 replication by targeting programmed −1 ribosomal frameshifting. *Proceedings of the National Academy of Sciences* **118**, e2023051118 (2021).

66. Hillen, H. S. *et al.* Structure of replicating SARS-CoV-2 polymerase. *Nature* **584**, 154–156 (2020).

67. Wallace, L. E., Liu, M., Kuppeveld, F. J. M. van, Vries, E. de & Haan, C. A. M. de. Respiratory mucus as a virus-host range determinant. *Trends in Microbiology* **29**, 983–992 (2021).

68. Minkoff, J. M. & tenOever, B. Innate immune evasion strategies of SARS-CoV-2. *Nat Rev Microbiol* **21**, 178–194 (2023).

69. Kawai, T. & Akira, S. Innate immune recognition of viral infection. *Nat Immunol* **7**, 131–137 (2006).

70. Luo, X. *et al.* Molecular Mechanism of RNA Recognition by Zinc-Finger Antiviral Protein. *Cell Reports* **30**, 46-52.e4 (2020).
71. Chemudupati, M. *et al.* From APOBEC to ZAP: Diverse mechanisms used by cellular restriction factors to inhibit virus infections. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1866**, 382–394 (2019).
72. Hewitt, E. W. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology* **110**, 163–169 (2003).
73. Vyas, J. M., Van der Veen, A. G. & Ploegh, H. L. The known unknowns of antigen processing and presentation. *Nat Rev Immunol* **8**, 607–618 (2008).
74. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* **11**, 823–836 (2011).
75. Shah, K., Al-Haidari, A., Sun, J. & Kazi, J. U. T cell receptor (TCR) signaling in health and disease. *Sig Transduct Target Ther* **6**, 1–26 (2021).
76. Alberts, B. *et al.* T Cells and MHC Proteins. in *Molecular Biology of the Cell. 4th edition* (Garland Science, 2002).
77. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat Rev Microbiol* **21**, 361–379 (2023).
78. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116–121 (2021).
79. Carabelli, A. M. *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat Rev Microbiol* **21**, 162–177 (2023).
80. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* **19**, 409–424 (2021).

81. Combe, M. & Sanjuán, R. Variation in RNA Virus Mutation Rates across Host Cells. *PLOS Pathogens* **10**, e1003855 (2014).

82. Gorbalenya, A. E., Enjuanes, L., Ziebuhr, J. & Snijder, E. J. Nidovirales: Evolving the Largest Rna Virus Genome. *Virus Research* **117**, 17–37 (2006).

83. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol* **19**, 155–170 (2021).

84. Holmes, E. C. Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol* **11**, 543–546 (2003).

85. Kimura, M. The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet* **66**, 367–386 (1991).

86. Perry, A. J., Hulett, J. M., Likić, V. A., Lithgow, T. & Gooley, P. R. Convergent Evolution of Receptors for Protein Import into Mitochondria. *Current Biology* **16**, 221–229 (2006).

87. Nomburg, J. *et al.* Birth of protein folds and functions in the virome. *Nature* 1–8 (2024) doi:10.1038/s41586-024-07809-y.

88. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246 (2024).

89. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

90. Ito, J. *et al.* Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant. *Nat Commun* **14**, 2671 (2023).

91. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189-5200.e7 (2021).

92. Gregory, D. A. *et al.* Genetic diversity and evolutionary convergence of cryptic SARS- CoV-2 lineages detected via wastewater sequencing. *PLOS Pathogens* **18**, e1010636 (2022).

93. da Silva Filipe, A. *et al.* Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. *Nat Microbiol* **6**, 112–122 (2021).

94. Vöhringer, H. S. *et al.* Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600**, 506–511 (2021).

95. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403–1407 (2020).

96. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution* **7**, veab064 (2021).

97. Tracking SARS-CoV-2 variants. https://www.who.int/activities/tracking-SARS-CoV-2-variants.

98. WHO announces simple, easy-to-say labels for SARS-CoV-2 Variants of Interest and Concern. https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern.

99. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

100. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

101. Nchioua, R. *et al.* SARS-CoV-2 Is Restricted by Zinc Finger Antiviral Protein despite Preadaptation to the Low-CpG Environment in Humans. *mBio* **11**, 10.1128/mbio.01930-20 (2020).

102. Cooper, D. N., Mort, M., Stenson, P. D., Ball, E. V. & Chuzhanova, N. A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Human Genomics* **4**, 406 (2010).

103. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacol* **38**, 23–38 (2013).

104. Simmonds, P., Xia, W., Baillie, J. K. & McKinnon, K. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla – selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics* **14**, 610 (2013).

105. Young, F., Rogers, S. & Robertson, D. L. Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLOS Computational Biology* **16**, e1007894 (2020).

106. Brierley, L. & Fowler, A. Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. *PLOS Pathogens* **17**, e1009149 (2021).

107. Aun, E., Brauer, A., Kisand, V., Tenson, T. & Remm, M. A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLOS Computational Biology* **14**, e1006434 (2018).

108. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).

109. Cam, L. M. L. & Neyman, J. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. (University of California Press, 1967).

110. Tokuda, E. K., Comin, C. H. & Costa, L. da F. Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and its Applications* **585**, 126433 (2022).

111. Sarker, I. H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* **2**, 420 (2021).

112. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

113. Oxford English Dictionary. grammar, n., sense 5.b. (2024).

114. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. Preprint at https://doi.org/10.48550/arXiv.1301.3781 (2013).

115. Vaswani, A. *et al.* Attention Is All You Need. Preprint at http://arxiv.org/abs/1706.03762 (2023).

116. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).

117. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.1810.04805 (2019).

118. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).

119. Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *cels* **13**, 274-285.e6 (2022).

120. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems* **12**, 969-982.e6 (2021).

121. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* **41**, 1099–1106 (2023).

122. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* **13**, 4348 (2022).

123. Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with Evo. 2024.02.27.582234 Preprint at https://doi.org/10.1101/2024.02.27.582234 (2024).

124. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nature Methods* **15**, 399–400 (2018).

125. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* **374**, 20150202 (2016).

126. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).

127. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

128. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

129. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3**, 246–259 (2013).

130. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

131. Petersen, E. *et al.* Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *The Lancet Infectious Diseases* **20**, e238–e244 (2020).

132. Dewi, A. *et al.* Global policy responses to the COVID-19 pandemic: proportionate adaptation and policy experimentation: a study of country policy response variation to the COVID-19 pandemic. *Health Promot Perspect* **10**, 359–365 (2020).

133. Kirby, T. New variant of SARS-CoV-2 in UK causes surge of COVID-19. *The Lancet Respiratory Medicine* **9**, e20–e21 (2021).

134. Lauring, A. S. & Hodcroft, E. B. Genetic Variants of SARS-CoV-2—What Do They Mean? *JAMA* **325**, 529–531 (2021).

135. COVID-19 epidemiological update – 12 April 2024. https://www.who.int/publications/m/item/covid-19-epidemiological-update-edition-166.

136. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).

137. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119 (2021).

138. Bugembe, D. L. *et al.* Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1) with altered spike protein in Uganda. *Nat Microbiol* **6**, 1094–1101 (2021).

139. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics & Development* **24**, 52–60 (2014).

140. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research* **45**, D777–D783 (2017).

141. Simmonds, P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5**, 10.1128/msphere.00408-20 (2020).

142. Ratcliff, J. & Simmonds, P. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology* **556**, 62–72 (2021).

143. Sanjuán, R. & Domingo-Calap, P. Mechanisms of viral mutation. *Cell. Mol. Life Sci.* **73**, 4433–4448 (2016).

144. Mathieu, E. *et al.* Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).

145. Dong, E. *et al.* The Johns Hopkins University Center for Systems Science and Engineering COVID-19 Dashboard: data collection process, challenges faced, and lessons learned. *The Lancet Infectious Diseases* **22**, e370–e376 (2022).

146. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**, 533–534 (2020).

147. Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat Hum Behav* **5**, 529–538 (2021).

148. Harari, S. *et al.* Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nat Med* **28**, 1501–1508 (2022).

149. Wei, J. *et al.* Risk of SARS-CoV-2 reinfection during multiple Omicron variant waves in the UK general population. *Nat Commun* **15**, 1008 (2024).

150. Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).

151. Heo, M.-H., Kwon, Y. D., Cheon, J., Kim, K.-B. & Noh, J.-W. Association between the Human Development Index and Confirmed COVID-19 Cases by Country. *Healthcare* **10**, 1417 (2022).

152. Azgari, C., Kilinc, Z., Turhan, B., Circi, D. & Adebali, O. The Mutation Profile of SARS-CoV-2 Is Primarily Shaped by the Host Antiviral Defense. *Viruses* **13**, 394 (2021).

153. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2**, (2022).

154. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

155. Picardi, E., Mansi, L. & Pesole, G. Detection of A-to-I RNA Editing in SARS-COV-2. *Genes* **13**, 41 (2022).

156. Ringlander, J. *et al.* Impact of ADAR-induced editing of minor viral RNA populations on replication and transmission of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **119**, e2112663119 (2022).

157. Bloom, J. D., Beichman, A. C., Neher, R. A. & Harris, K. Evolution of the SARS-CoV-2 Mutational Spectrum. *Molecular Biology and Evolution* **40**, msad085 (2023).

158. Ruis, C. *et al.* Mutational spectra distinguish SARS-CoV-2 replication niches. 2022.09.27.509649 Preprint at https://doi.org/10.1101/2022.09.27.509649 (2022).

159. Wang, P. *et al.* Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021).

160. Wang, Z. *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature* **592**, 616–622 (2021).

161. Ou, J. *et al.* Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. *Sig Transduct Target Ther* **7**, 1–9 (2022).

162. Shafer, M. M. *et al.* Human origin ascertained for SARS-CoV-2 Omicron-like spike sequences detected in wastewater: a targeted surveillance study of a cryptic lineage in an urban sewershed. 2022.10.28.22281553 Preprint at https://doi.org/10.1101/2022.10.28.22281553 (2023).

163. Chaguza, C. *et al.* Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. *CR Med* **4**, (2023).

164. Bamgboye, E. L. *et al.* COVID-19 Pandemic: Is Africa Different? *Journal of the National Medical Association* **113**, 324–335 (2021).

165. Herng, L. C. *et al.* The effects of super spreading events and movement control measures on the COVID-19 pandemic in Malaysia. *Sci Rep* **12**, 2197 (2022).

166. Graudenzi, A., Maspero, D., Angaroni, F., Piazza, R. & Ramazzotti, D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *iScience* **24**, (2021).

167. Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Science Advances* **6**, eabb5813 (2020).

168. Yi, K. *et al.* Mutational spectrum of SARS-CoV-2 during the global pandemic. *Exp Mol Med* **53**, 1229–1237 (2021).

169. Aroldi, A. *et al.* Characterization of SARS-CoV-2 Mutational Signatures from 1.5+ Million Raw Sequencing Samples. *Viruses* **15**, 7 (2023).

170. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979–993 (2012).

171. Langenbucher, A. *et al.* An extended APOBEC3A mutation signature in cancer. *Nat Commun* **12**, 1602 (2021).

172. Kim, K. *et al.* The roles of APOBEC-mediated RNA editing in SARS-CoV-2 mutations, replication and fitness. *Sci Rep* **12**, 14972 (2022).

173. Trypsteen, W., Cleemput, J. V., Snippenberg, W. van, Gerlo, S. & Vandekerckhove, L. On the whereabouts of SARS-CoV-2 in the human body: A systematic review. *PLOS Pathogens* **16**, e1009037 (2020).

174. Wanner, N. *et al.* Molecular consequences of SARS-CoV-2 liver tropism. *Nat Metab* **4**, 310–319 (2022).

175. Lamers, M. M. *et al.* SARS-CoV-2 productively infects human gut enterocytes. *Science* **369**, 50–54 (2020).

176. Qian, Q. *et al.* Direct Evidence of Active SARS-CoV-2 Replication in the Intestine. *Clinical Infectious Diseases* **73**, 361–366 (2021).

177. Guo, M., Tao, W., Flavell, R. A. & Zhu, S. Potential intestinal infection and faecal–oral transmission of SARS-CoV-2. *Nat Rev Gastroenterol Hepatol* **18**, 269–283 (2021).

178. Zhang, Y. *et al.* Isolation of 2019-nCoV from a Stool Specimen of a Laboratory-Confirmed Case of the Coronavirus Disease 2019 (COVID-19). *CCDCW* **2**, 123–124 (2020).

179. Xiao, F. *et al.* Infectious SARS-CoV-2 in Feces of Patient with Severe COVID-19 - Volume 26, Number 8—August 2020 - Emerging Infectious Diseases journal - CDC. doi:10.3201/eid2608.200681.

180. Chan, K.-H. *et al.* Factors affecting stability and infectivity of SARS-CoV-2. *Journal of Hospital Infection* **106**, 226–231 (2020).

181. Takata, M. A. *et al.* CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550**, 124–127 (2017).

182. Zimmer, M. M. *et al.* The short isoform of the host antiviral protein ZAP acts as an inhibitor of SARS-CoV-2 programmed ribosomal frameshifting. *Nat Commun* **12**, 7193 (2021).

183. Mourier, T. *et al.* Host-directed editing of the SARS-CoV-2 genome. *Biochemical and Biophysical Research Communications* **538**, 35–39 (2021).

184. Li, Z., Wu, J. & DeLeo, C. J. RNA damage and surveillance under oxidative stress. *IUBMB Life* **58**, 581–588 (2006).

185. Simmonds, P. & Ansari, M. A. Extensive C->U transition biases in the genomes of a wide range of mammalian RNA viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral RNA. *PLOS Pathogens* **17**, e1009596 (2021).

186. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16 (2019).

187. Thorne, L. G. *et al.* Evolution of enhanced innate immune evasion by SARS-CoV-2. *Nature* **602**, 487–495 (2022).

188. Liu, G. & Gack, M. U. SARS-CoV-2 learned the 'Alpha'bet of immune evasion. *Nat Immunol* **23**, 351–353 (2022).

189. Chen, K. *et al.* SARS-CoV-2 Nucleocapsid Protein Interacts with RIG-I and Represses RIG-Mediated IFN-ß Production. *Viruses* **13**, 47 (2021).

190. Catanzaro, M. *et al.* Immune response in COVID-19: addressing a pharmacological challenge by targeting pathways triggered by SARS-CoV-2. *Sig Transduct Target Ther* **5**, 1–10 (2020).

191. Miorin, L. *et al.* SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling. *Proceedings of the National Academy of Sciences* **117**, 28344–28354 (2020).

192. Willett, B. J. *et al.* SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nat Microbiol* **7**, 1161–1179 (2022).

193. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).

194. Harris, Z. S. Distributional Structure. *WORD* **10**, 146–162 (1954).

195. Elliott, P. *et al.* Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science* **374**, eabl9551 (2021).

196. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).

197. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. & Vapnik, V. Support Vector Regression Machines. in *Advances in Neural Information Processing Systems* vol. 9 (MIT Press, 1996).

198. Lytras, S. *et al.* Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination. *Genome Biology and Evolution* **14**, evac018 (2022).

199. Martin, D. P. *et al.* Selection Analysis Identifies Clusters of Unusual Mutational Changes in Omicron Lineage BA.1 That Likely Impact Spike Function. *Molecular Biology and Evolution* **39**, msac061 (2022).

200. Murrell, B. *et al.* Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLOS Genetics* **8**, e1002764 (2012).

201. Kosakovsky Pond, S. L. & Frost, S. D. W. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and Evolution* **22**, 1208–1222 (2005).

202. Sweredoski, M. J. & Baldi, P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics* **24**, 1459–1460 (2008).

203. Overington, J., Donnelly, D., Johnson, M. S., Sali, A. & Blundell, T. L. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci* **1**, 216–226 (1992).

204. Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S. & Overington, J. P. JOY: protein sequence-structure representation and analysis. *Bioinformatics* **14**, 617–623 (1998).

205. Gong, S. & Blundell, T. L. Discarding Functional Residues from the Substitution Table Improves Predictions of Active Sites within Three-Dimensional Structures. *PLOS Computational Biology* **4**, e1000179 (2008).

206. Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biology Direct* **7**, 12 (2012).

207. Joost Schymkowitz *et al.* The FoldX web server: an online force field. *Nucleic Acids Research* **33**, W382–W388 (2005).

208. Starr, T. N. *et al.* Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377**, 420–424 (2022).

209. Yisimayi, A. *et al.* Repeated Omicron exposures override ancestral SARS-CoV-2 immune imprinting. *Nature* **625**, 148–156 (2024).

210. Dadonaite, B. *et al.* Full-spike deep mutational scanning helps predict the evolutionary success of SARS-CoV-2 clades. 2023.11.13.566961 Preprint at https://doi.org/10.1101/2023.11.13.566961 (2023).

211. Nemet Ital *et al.* Third BNT162b2 Vaccination Neutralization of SARS-CoV-2 Omicron Infection. *New England Journal of Medicine* **386**, 492–494 (2022).

212. Lauring, A. S. *et al.* Clinical severity of, and effectiveness of mRNA vaccines against, covid-19 from omicron, delta, and alpha SARS-CoV-2 variants in the United States: prospective observational study. *BMJ* **376**, e069761 (2022).

213. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).

214. Greaney, A. J. *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat Commun* **12**, 4196 (2021).

215. Cao, Y. *et al.* Dynamic Interactions of Fully Glycosylated SARS-CoV-2 Spike Protein with Various Antibodies. *J. Chem. Theory Comput.* **17**, 6559–6569 (2021).

216. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2020).

217. Zhang, J. *et al.* Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* **372**, 525–530 (2021).

218. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* **55**, 1512–1522 (2023).

219. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).

220. Cerutti, G. *et al.* Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell Host & Microbe* **29**, 819-833.e7 (2021).

221. Cui, Z. *et al.* Structural and functional characterizations of infectivity and immune evasion of SARS-CoV-2 Omicron. *Cell* **185**, 860-871.e13 (2022).

222. Liu, Z. *et al.* Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host & Microbe* **29**, 477-488.e4 (2021).

223. Lok, S.-M. An NTD supersite of attack. *Cell Host & Microbe* **29**, 744–746 (2021).

224. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332-2347.e16 (2021).

225. Schröder, S. *et al.* Characterization of intrinsic and effective fitness changes caused by temporarily fixed mutations in the SARS-CoV-2 spike E484 epitope and identification of an epistatic precondition for the evolution of E484A in variant Omicron. *Virology Journal* **20**, 257 (2023).

226. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. 2021.10.04.463034 Preprint at https://doi.org/10.1101/2021.10.04.463034 (2022).

227. Peacock, T. P. *et al.* The altered entry pathway and antigenic distance of the SARS-CoV-2 Omicron variant map to separate domains of spike protein. 2021.12.31.474653 Preprint at https://doi.org/10.1101/2021.12.31.474653 (2022).

228. McCallum, M. *et al.* Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. *Science* **375**, 864–868 (2022).

229. Yang, K. *et al.* Structure-based design of a SARS-CoV-2 Omicron-specific inhibitor. *Proceedings of the National Academy of Sciences* **120**, e2300360120 (2023).

230. Watanabe, K. & Suzuki, Y. Protein thermostabilization by proline substitutions. *Journal of Molecular Catalysis B: Enzymatic* **4**, 167–180 (1998).

231. Choi, E. J. & Mayo, S. L. Generation and analysis of proline mutants in protein G. *Protein Engineering, Design and Selection* **19**, 285–289 (2006).

232. Wong, J. W. H., Ho, S. Y. W. & Hogg, P. J. Disulfide Bond Acquisition through Eukaryotic Protein Evolution. *Molecular Biology and Evolution* **28**, 327–334 (2011).

233. Livesey, B. J. & Marsh, J. A. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol Syst Biol* **19**, e11474 (2023).

234. Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nat Biotechnol* **42**, 275–283 (2024).

235. Saito, A. *et al.* Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature* **602**, 300–306 (2022).

236. Andrews, N. *et al.* Covid-19 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant. *New England Journal of Medicine* **386**, 1532–1546 (2022).

237. Mykytyn, A. Z. *et al.* Antigenic cartography of SARS-CoV-2 reveals that Omicron BA.1 and BA.2 are antigenically distinct. *Science Immunology* **7**, eabq4450 (2022).

238. Yang, S. *et al.* Fast evolution of SARS-CoV-2 BA.2.86 to JN.1 under heavy immune pressure. *The Lancet Infectious Diseases* **24**, e70–e72 (2024).

239. Lynch, M. Evolutionary diversification of the multimeric states of proteins. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E2821 (2013).

240. Avraham, O., Tsaban, T., Ben-Aharon, Z., Tsaban, L. & Schueler-Furman, O. Protein language models can capture protein quaternary state. *BMC Bioinformatics* **24**, 433 (2023).

241. Ito, J. *et al.* A Protein Language Model for Exploring Viral Fitness Landscapes. 2024.03.15.584819 Preprint at https://doi.org/10.1101/2024.03.15.584819 (2024).

242. Kumar, S. *et al.* Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell* **180**, 915-927.e16 (2020).

243. Raphael, B. J., Dobson, J. R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine* **6**, 5 (2014).

244. Lee, E. Y. H. P. & Muller, W. J. Oncogenes and Tumor Suppressor Genes. *Cold Spring Harb Perspect Biol* **2**, a003236 (2010).

245. Kim, Y.-A., Madan, S. & Przytycka, T. M. WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* **33**, 814–821 (2017).

246. Cisowski, J. & Bergo, M. O. What makes oncogenes mutually exclusive? *Small GTPases* **8**, 187–192 (2016).

247. El Tekle, G. *et al.* Co-occurrence and mutual exclusivity: what cross-cancer mutation patterns can tell us. *Trends in Cancer* **7**, 823–836 (2021).

248. Deng, Y. *et al.* Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Briefings in Bioinformatics* **20**, 254–266 (2019).

249. Chen, H., Zhou, X., Zheng, J. & Kwoh, C.-K. Rules of co-occurring mutations characterize the antigenic evolution of human influenza A/H3N2, A/H1N1 and B viruses. *BMC Medical Genomics* **9**, 69 (2016).

250. Ac, S., Tc, H., Ms, H. & Wh, L. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, (2007).

251. Ly, J. K. *et al.* The Balance between NRTI Discrimination and Excision Drives the Susceptibility of HIV-1 RT Mutants K65R, M184V and K65R+M184V. *Antivir Chem Chemother* **18**, 307–316 (2007).

252. Meng, B. *et al.* Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Reports* **35**, (2021).

253. Wilkinson, S. A. J. *et al.* Recurrent SARS-CoV-2 mutations in immunodeficient patients. *Virus Evolution* **8**, veac050 (2022).

254. Kim, Y.-A., Madan, S. & Przytycka, T. M. WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* **33**, 814–821 (2017).

255. Tsueng, G. *et al.* Outbreak.info Research Library: a standardized, searchable platform to discover and explore COVID-19 resources. *Nat Methods* **20**, 536–540 (2023).

256. Wu, H. *et al.* Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host & Microbe* **29**, 1788-1801.e6 (2021).

257. Xu, Z. *et al.* SARS-CoV-2 impairs interferon production via NSP2-induced repression of mRNA translation. *Proceedings of the National Academy of Sciences* **119**, e2204539119 (2022).

258. Korneeva, N. *et al.* SARS-CoV-2 viral protein Nsp2 stimulates translation under normal and hypoxic conditions. *Virology Journal* **20**, 55 (2023).

259. Chen, C. *et al.* CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics* **38**, 1735–1737 (2022).

260. Morse, M., Sefcikova, J., Rouzina, I., Beuning, P. J. & Williams, M. C. Structural domains of SARS-CoV-2 nucleocapsid protein coordinate to compact long nucleic acid substrates. *Nucleic Acids Research* **51**, 290–303 (2023).

261. Syed, A. M. *et al.* Rapid assessment of SARS-CoV-2–evolved variants using virus-like particles. *Science* **374**, 1626–1632 (2021).

262. Miyamoto, Y. *et al.* SARS-CoV-2 ORF6 disrupts nucleocytoplasmic trafficking to advance viral replication. *Commun Biol* **5**, 1–15 (2022).

263. Kehrer, T. *et al.* Impact of SARS-CoV-2 ORF6 and its variant polymorphisms on host responses and viral pathogenesis. *Cell Host & Microbe* **31**, 1668-1684.e12 (2023).

264. Focosi, D., Quiroga, R., McConnell, S., Johnson, M. C. & Casadevall, A. Convergent Evolution in SARS-CoV-2 Spike Creates a Variant Soup from Which New COVID-19 Waves Emerge. *Int J Mol Sci* **24**, 2264 (2023).

265. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

266. Zhang, Y. *et al.* Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research* **52**, e3 (2024).

267. Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* **4**, 852–866 (2022).

268. Li, F.-Z., Amini, A. P., Yue, Y., Yang, K. K. & Lu, A. X. Feature Reuse and Scaling: Understanding Transfer Learning with Protein Language Models. 2024.02.05.578959 Preprint at https://doi.org/10.1101/2024.02.05.578959 (2024).

269. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).

270. Zhang, Z. *et al.* Protein language models learn evolutionary statistics of interacting sequence motifs. 2024.01.30.577970 Preprint at https://doi.org/10.1101/2024.01.30.577970 (2024).

271. Gu, A. & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Preprint at https://doi.org/10.48550/arXiv.2312.00752 (2023).

272. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. Preprint at https://doi.org/10.48550/arXiv.2306.15794 (2023).

273. Na, B. *et al.* Therapeutic targeting of BRCA1 and TP53 mutant breast cancer through mutant p53 reactivation. *npj Breast Cancer* **5**, 1–10 (2019).

274. Wu, J. *et al.* PLM-ARG: antibiotic resistance gene identification using a pretrained protein language model. *Bioinformatics* **39**, btad690 (2023).

275. Dewachter, L. *et al.* Deep mutational scanning of essential bacterial proteins can guide antibiotic development. *Nat Commun* **14**, 241 (2023).