MACHINE-LEARNING-BASED IDENTIFICATION OF FACTORS THAT INFLUENCE MOLECULAR VIRUS-HOST INTERACTIONS

HAITING CHAI

SUBMITTED IN FULFILMENT OF THE REQUIREMENT OF THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

INSTITUTE OF INFECTION, IMMUNITY AND INFLAMMATION
COLLEGE OF MEDICAL, VETERINARY AND LIFE SCIENCES
UNIVERSITY OF GLASGOW

May 2022

… the night is long that never finds the day.

William Shakespeare, *Macbeth* (1623)

ABSTRACT

Viruses are the cause of many infectious diseases such as the pandemic viruses: acquired immune deficiency syndrome (AIDS) and coronavirus disease 2019 (COVID-19). During the infection cycle, viruses invade host cells and trigger a series of virus-host interactions with different directionality. Some of these interactions disrupt host immune responses or promote the expression of viral proteins and exploitation of the host system thus are considered ‘pro-viral’. Some interactions display ‘pro-host’ traits, principally the immune response, to control or inhibit viral replication. Concomitant pro-viral and pro-host molecular interactions on the same host molecule suggests more complex virus-host conflicts and genetic signatures that are crucial to host immunity. In this work, machine-learning-based prediction of virus-host interaction directionality was examined by using data from Human immunodeficiency virus type 1 (HIV-1) infection. Host immune responses to viral infections are mediated by interferons (IFNs) in the initial stage of the immune response to infection. IFNs induce the expression of many IFN-stimulated genes (ISGs), which make the host cell refractory to further infection. We propose that there are many features associated with the up-regulation of human genes in the context of IFN-α stimulation. They make ISGs predictable using machine-learning models. In order to overcome the interference of host immune responses for successful replication, viruses adopt multiple strategies to avoid being detected by cellular sensors in order to hijack the machinery of host transcription or translation. Here, the strategy of mimicry of host-like short linear motifs (SLiMs) by the virus was investigated by using the example of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The integration of in silico experiments and analyses in this thesis demonstrates an interactive and intimate relationship between viruses and their hosts. Findings here contribute to the identification of host dependency and anti-viral factors. They are of great importance not only to the ongoing COVID-19 pandemic but also to the understanding of future disease outbreaks.

ACKNOWLEDGEMENTS

First of all, I would like to give special thanks to my supervisors, Drs David Robertson, Joseph Hughes, and Quan Gu for their constant support, encouragement, and inspiration. I felt so privileged to be your student. Thanks also go to Drs Andrew Davison, Suzannah Rihn, and Ke Yuan for their insightful comments in annual assessments.

I am grateful to the Robertson Lab, Bioinformatics Hub, and CVR community for giving me a perfect working environment. I wish to thank Vandana Ravindran, Francesca Young, Sejal Modha, Spyros Lytras, Joseph Busby, and in particular, Ben Stamp for giving interesting ideas in lab meetings.

To friends I made in Glasgow, Belfast, and Bristol, thanks for the time we spent, always.

Finally, I would like to thank my partner, Miss Jia, for her trust and company along this journey. I am greatly indebted to my parents for their continuous supports to undertake this endeavour.

This work was funded by China Scholarship Council under Grant 201706620069.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1. INTRODUCTION TO THE THESIS

This thesis is presented in the format of self-contained chapters. The research presented in the third, fourth and fifth chapters form the main Results chapters with different rationales and independent results. Supplementary files for this work are available at the web server of HIVPRE (http://hivpre.cvr.gla.ac.uk/), ISGPRE (http://isgpre.cvr.gla.ac.uk/) and GitHub archive (https://github.com/HChai01/SARS-COV-2).

1.1 Contributions

This thesis makes the following key research contributions:

(1) Novel perspective of virus-host interaction prediction by relating human proteins to pro-viral or pro-host phenotypes;

(2) Systematic analyses of hosts' innate properties;

(3) Novel feature selection schemes to optimise machine learning models in the context of virus-host interactions;

(4) Web server for predicting human immunodeficiency virus type 1 (HIV-1) interacting proteins and their directionality: pro-viral versus pro-host;

(5) Systematic analyses of innate properties related to the interferon response to virus infection;

(6) Web server for predicting interferon-stimulated human genes;

(7) Elucidation of the mimicry mechanism adopted by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

1.2 Publications and Authorship

This thesis unifies and expands work in the following four publications:

- **Haiting Chai**[1], Quan Gu[1], Joseph Hughes[1] and David L. Robertson[1]. In silico prediction of HIV-1-host molecular interactions and their directionality. *PLOS Computational Biology*. 2022; 18(2): e1009720. https://doi.org/10.1371/journal.pcbi.1009720 PMID: 35134057

Author contributions: DLR, JH, QG and HC conceptualization; HC data curation; HC formal analysis; HC and DLR funding acquisition; HC investigation; DLR, JH, QG and HC methodology; DLR, JH and QG resources; HC software; DLR, JH and QG supervision; DLR, JH, QG and HC validation; HC visualization; HC writing-original draft; DLR, JH, QG and HC writing-review and editing; DLR project administration.

- **Haiting Chai**[1], Quan Gu[1], Joseph Hughes[1] and David L. Robertson[1]. Defining the Characteristics of Interferon-alpha-stimulated Genes: Insight from Data Expression and Machine Learning. Under review.

Author contributions: DLR, JH, QG and HC conceptualization; HC data curation; HC formal analysis; HC and DLR funding acquisition; HC investigation; DLR, JH, QG and HC methodology; DLR, JH and QG resources; HC software; DLR, JH and QG supervision; DLR, JH, QG and HC validation; HC visualization; HC writing-original draft; DLR, JH, QG and HC writing-review and editing; DLR and JH project administration.

- **Haiting Chai**[1], Quan Gu[1], Lukasz Kurgan[2], Joseph Hughes[1] and David L. Robertson[1]. SARS-CoV-2 mimicry of host protein interactions mediated via short linear motifs. To be submitted.

Author contributions: DLR, JH, QG, LK and HC conceptualization; HC data curation; HC formal analysis; HC and DLR funding acquisition; HC investigation; DLR, JH, QG and HC methodology; DLR, JH and QG resources; DLR, JH and QG supervision; DLR, JH, QG and HC validation; HC visualization; HC writing-original draft; DLR, JH, QG, LK and HC writing-review and editing; DLR and JH project administration.

- **Haiting Chai**[1], Joseph Hughes[1] Quan Gu[1] and David L. Robertson[1]. A review of supervised machine learning methods for predicting virus-host molecular interactions. In preparation.

Author contributions: DLR, JH, QG and HC conceptualization; HC data curation; HC formal analysis; HC and DLR funding acquisition; HC investigation; DLR, JH, QG and HC methodology; DLR, JH and QG resources; DLR, JH and QG supervision; DLR, JH, QG and

HC validation; HC visualization; HC writing-original draft; DLR, JH, QG and HC writing-review and editing; DLR and JH project administration.

Affiliation: [1]MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom; [2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA.

1.3 Thesis outline

This first chapter provides an overview of the thesis organisation and contributions. The second chapter provides background information to the thesis with a focus on virus-host molecular interactions. It introduces the virus life cycle with a focus on the entry, replication and release stages. It gives a brief introduction on the interplay between viruses and their hosts including an introduction to the role of host innate immune responses and corresponding strategies viruses developed for antagonizing the immune response. This chapter also introduces the usage of machine-leaning in virus research with a particular focus on typical supervised learning algorithms used for classification.

Chapter Three is a research section of the thesis. It investigates molecular interactions between HIV-1 and human host proteins. It elucidates that some features of HIV-1 interacting proteins (VIPs) make human proteins more likely to be targeted by HIV-1 proteins. The results demonstrate that the direction of HIV-1-host molecular interactions is predictable due to different characteristics of pro-viral and pro-host human proteins. It proposes a machine learning framework developed for the prediction of putative VIPs and their directionality in HIV-1-host molecular interactions.

Chapter Four is another research section. It provides detailed analyses to characterize interferon-α-stimulated human genes (ISG), which are potentially anti-viral to protect the host cells against viral infections. It summarizes factors that may enhance or suppress the stimulation of human genes with the presence of interferon-α (IFN-α). It illustrates that interferon-α-repressed human genes (IRGs), which are down-regulated in the IFN-α system, can have similar properties to ISGs. In this chapter, machine-learning approach is applied in the form of classification.

Chapter Five is the last research section. It investigates the molecular interactions between SARS-CoV-2 and human host proteins. It proposes that SARS-CoV-2 mimics host-like short linear motifs (SLiMs) from the human protein-protein interaction network to facilitate its interaction affinity with host proteins. It elucidates the discovery and evolutionary conservation of 18 SLiMs mimicked by the envelope, nucleocapsid, open reading frame 7a (ORF7a), and non-structural protein 8 (NSP8).

The final chapter summarises the main discoveries of this work. It discusses the limitations to the conducted research and solutions to overcome them. It also discusses some interesting ideas for future research.

CHAPTER TWO

2. INTRODUCTION TO VIRUS-HOST MOLECULAR INTERACTIONS

2.1 Abstract

Viruses, especially human-infecting viruses pose a serious threat to public health. To replicate and persist viruses make intimate relationships with their hosts exploiting host systems through virus-host molecular interactions. Investigating the interplay between virus and the human host contributes to a better understanding of the pathogenesis underlying viral infections and can contribute to the development of novel antiviral drugs or therapeutics. The development of DNA microarray, polymerase chain reaction (PCR), genome and transcriptome sequencing and assay technologies permits the creation of biological data for *in silico* analyses, which in turn facilitates *in vivo* and *in vitro* experiments. Machine learning provides an efficient methodology to decipher key factors embedded in complex biological data and its context. It also gives new insights for potential virus infections, which are of great importance, for example, to the ongoing Coronavirus disease 2019 (COVID-19) pandemic and future disease outbreaks.

2.2 Viral infection and interplay with the host

A virus is an infectious agent that replicates inside living cells. It can infiltrate every cell-based life form in nature [1]. A recent report from UniProt [2] indicates a compilation of 17,039 manually reviewed and more than 5,000,000 unreviewed viral proteins. Viruses can be catalogued into seven Baltimore classes based on the genetic materials present in their virions (**Fig 2.1**). They are known to infect almost all living organisms ranging from animals to microorganisms [3]. Due to a dependency on living entities, viruses have to rapidly exploit host systems for their replication and persistence. This promotes viruses to establish intimate molecular relationships with their hosts, a coevolutionary relationship that constantly shapes their genome through the need to maintain virus-host molecular interactions. [4].

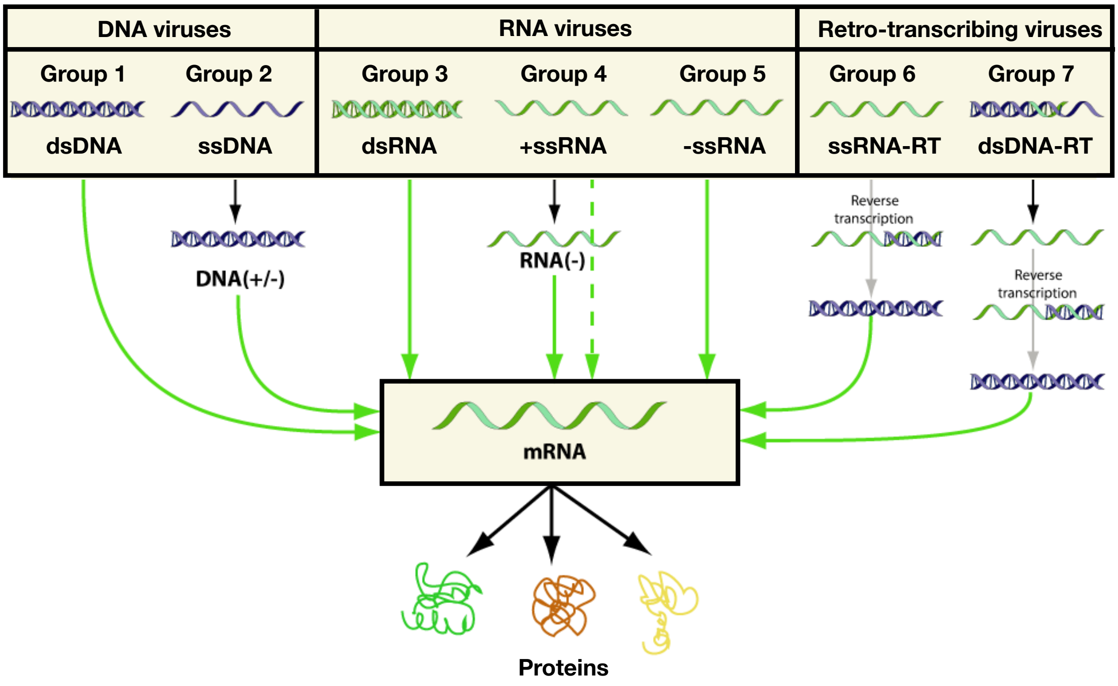


Fig 2.1 The Baltimore classification of virus. To produce mRNA for translation, virus genes in Group 1/2 are transcribed from DNA; those in Group 3/5 are transcribed from RNA; those in Group 4 act as mRNA themselves; those in Group 6 have the path of RNA-DNA-mRNA; those in the last group have the path of DNA-RNA-DNA-mRNA. Figure is reproduced from ViralZone [5] under the Creative Commons licence 4.0. Abbreviations: dsDNA, double-stranded DNA; ssDNA, single-stranded DNA; dsRNA, double-stranded RNA; +ssRNA, positive sense single-stranded RNA; -ssRNA, negative sense single-stranded RNA; ssRNA-RT, single-stranded RNA with a DNA intermediate; dsDNA-RT, double-stranded DNA with a RNA intermediate.

2.2.1 Virus life cycle

Despite differences between individual virus types, the life cycle of viruses generally contains three main stages including: (I) viral entry, (II) expression and replication, and (III) viral exit. Exemplified by retro-transcribing (RT) viruses, these stages can be further partitioned as: (1) attachment/binding to the surface of the host cell, (2) fusion of virions into the host cell, (3) reverse transcription to generate viral DNA, (4) integration of viral DNA into the host cell genome, (5) transcription of viral mRNA from the integrated pro-viral DNA, (6) translation of new viral proteins, (7) assembly of virions in preparation for the next round of infection, (8) budding through the host plasma membrane, and (9) extracellular release from the host cell [6,7] (**Fig 2.2**).

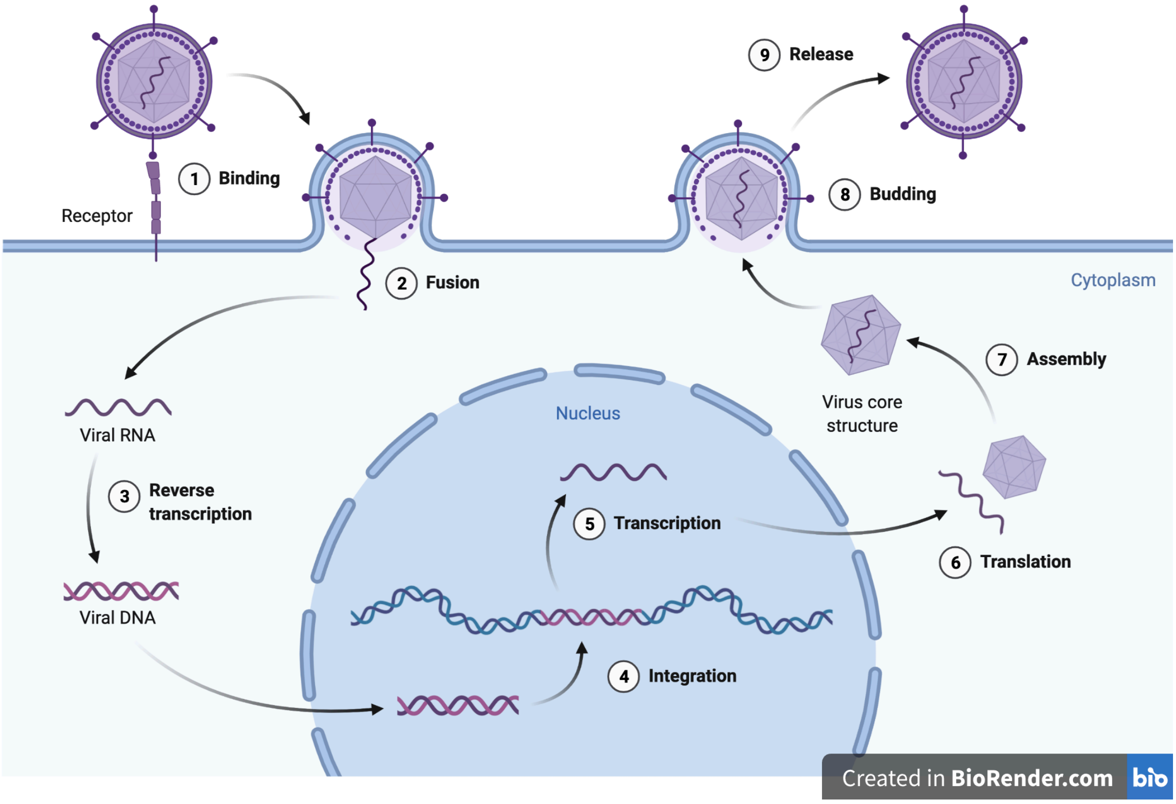


Fig 2.2 The life cycle of retro-transcribing virus indicating the processes from its entry to release. Viral entry includes the first two steps, viral exit includes the last three steps, and the rest steps enable the replication of viral genome in the host cells. Figure is created via the BioRender (https://biorender.com/) under the Creative Commons licence 4.0.

2.2.1.1 Viral entry

Viruses are usually highly specific about the host cells they can infect. This cell tropism is determined in their entry stage, during which the virus particle encounters the host cell and binds to the host entry receptor with their membrane glycoproteins or sites on a viral capsid [8]. The virus particle is metastable to protect the viral genome from immunological recognition and to maximise the release of virions after fusion/endocytosis events [9]. Meanwhile, the host entry receptor also mediates the fusion (for enveloped viruses) or endocytosis (for non-enveloped viruses) events, which allow the virus to access the interior of the host cell for replication [6]. **Table 2.1** lists examples of a number of host entry receptors exploited by human-infecting viruses. Their function in the host system, e.g., regulation or in a signalling pathway is subverted after virus binding, e.g., human immunodeficiency virus type 1 (HIV-1) downregulates expression of coreceptor C-C motif chemokine receptor 5 (CCR5) and C-X-C motif chemokine receptor 4 (CXCR4) [10]. In fact, there are more host molecules (attachment factors) recruited in the entry stage to enhance the affinity between viral particle and host entry receptor, exemplified by C-type lectin domain family 4 member M (CLEC4M) for hepatitis C virus (HCV) [11], severe acute respiratory syndrome coronavirus (SARS-CoV) [12], Japanese encephalitis virus (JEV) [13], West Nile virus (WNV) [14], etc. While many viruses may only choose one specific receptor to invade the host cell, some viruses display a complex dependency of host entry factors in the entry stage. For example, CD4 is the primary host entry receptor of HIV-1. Molecular interaction between CD4 and envelope protein (*env*) of HIV-1 initiate conformational changes to promote further binding to CCR5 and CXCR4 [15].

Table 2.1 Host entry receptors for human-infecting viruses.

Virus	Class	Receptor	Ref.
Adenovirus	dsDNA	CD80, CD86, CD46	[16]
BK virus	dsDNA	Gangliosides GD1b/GT1b	[17]
EBV	dsDNA	HLA-DRB1, ITGB1	[18,19]
HSV-1	dsDNA	PVRL1, PVRL2, HVEM, ITGAV, ITGB6	[20,21]
HSV-2	dsDNA	PVRL1, PVRL2, HVEM	[21,22]
HCMV	dsDNA	ITGAV, ITGB3	[23]
JCPyV	dsDNA	HTR2A	[24]
HPV	dsDNA	ITGA6	[25]
PRV	dsDNA	PVR, PVRL1, PVRL2	[22,26]
VZV	dsDNA	IDE	[27]
HHV-6	dsDNA	CD46, TNFRSF4	[28,29]
HHV-8	dsDNA	ITGAV, ITGB3	[30]
AAV	ssDNA	Heparan sulfate, RPSA	[31,32]
B19V	ssDNA	ITGA5, ITGB1	[33]
Rotavirus	dsRNA	ITGA2, ITGB1	[34]
MRV	dsRNA	JAM-A, ITGB1, NgR1	[35-37]
Coxsackievirus	+ssRNA	ICAM-1, ITGAV, ITGB3, ITGB6, CAR	[38-41]
Dengue virus	+ssRNA	Tyro3, AXL, CLDN1, RPSA	[42-44]
Echovirus	+ssRNA	ITGA2, ITGB1, DAF,	[45,46]
EV71	+ssRNA	Sialic acids, SCARB2, SELPLG	[47-49]
HAV	+ssRNA	HAVCR1	[50]
HCV	+ssRNA	LDLR, CLDN1, CD81, SCARB1	[51-53]
HCoV-229E	+ssRNA	ANPEP	[54]
HCoV-OC43	+ssRNA	Sialic acids	[55]
MERS-CoV	+ssRNA	DPP4	[56]
SARS-CoV	+ssRNA	ACE2	[57]
SARS-CoV-2	+ssRNA	ACE2	[58]
HPEV-1	+ssRNA	ITGAV, ITGB3	[59]
JEV	+ssRNA	DC-SIGN, LSECtin	[13]
HRV	+ssRNA	ICAM-1, LDLR, CDHR3	[60-62]

Norwalk virus	+ssRNA	HBGA	[63]
Poliovirus	+ssRNA	PVR	[64]
Rubella virus	+ssRNA	MOG	[65]
Sindbis virus	+ssRNA	RPSA	[66]
VEE	+ssRNA	RPSA	[67]
WNV	+ssRNA	ITGAV, ITGB3	[68]
Zika virus	+ssRNA	Tyro3, AXL	[69]
EBOV	-ssRNA	HAVCR1, Typo3, AXL, NPC1,	[70-72]
GTOV	-ssRNA	TFRC	[73]
HTNV	-ssRNA	ITGB3	[74]
HeV	-ssRNA	EFNB2, EFNB3	[75]
Influenza A	-ssRNA	CACNA1C, CD207, Sialic acids	[76-78]
Influenza B	-ssRNA	Sialic acids	[79]
Influenza C	-ssRNA	Sialic acids	[80]
hMPV	-ssRNA	ITGA5, ITGB3	[78]
HPIV	-ssRNA	Sialic acids	[81]
JUNV	-ssRNA	TFRC	[73]
Lassa virus	-ssRNA	α-dystroglycan	[82]
LCMV	-ssRNA	α-dystroglycan, CLEC4G, Tyro3, AXL	[82,83]
MACV	-ssRNA	TFRC	[84]
Marburg virus	-ssRNA	HAVCR1	[72]
MeV	-ssRNA	SLAMF1, PVRL4	[85,86]
NiV	-ssRNA	EFNB2, EFNB3	[87,88]
Rabies virus	-ssRNA	nAChR, NCAM1	[89]
RVFV	-ssRNA	DC-SIGN	[90]
UUKV	-ssRNA	DC-SIGN	[90]
VSV	-ssRNA	LDLR	[91]
HIV-1	ssRNA-RT	CD4, CCR5, CXCR4	[92]

Abbreviations: dsDNA, double-stranded DNA; ssDNA, single-stranded DNA; dsRNA, double-stranded RNA; +ssRNA, positive sense single-stranded RNA; -ssRNA, negative sense single-stranded RNA; ssRNA-RT, single-stranded RNA with a DNA intermediate; EBV, Epstein-Barr virus; HSV-1, herpes simplex virus type 1; HSV-2, herpes simplex virus type 2; HCMV, human cytomegalovirus; JCPyV, human JC polyomavirus; HPV, human papillomavirus; PRV, pseudorabies virus; VACV, vaccinia virus; HHV-6, human herpesvirus 6; HHV-8, human herpesvirus 8; VZV, Varicella-zoster virus; AAV, adeno-associated virus; B19V, human parvovirus B19; MRV, mammalian reovirus; EV71, enterovirus 71, HAV, hepatitis A virus; HCV, hepatitis C virus; HCoV-229E, human coronavirus 229E; HCoV-OC43, human coronavirus OC43; MERS-CoV, middle east respiratory syndrome coronavirus; SARS-CoV, severe acute respiratory syndrome coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; HPEV-1, human parechovirus 1; JEV, Japanese encephalitis virus; HRV, human rhinovirus; VEE, Venezuelan equine encephalitis; WNV, West Nile virus; EBOV, Ebola virus; GTOV, Guanarito virus; HTNV, Hantaan virus; HeV, Hendra virus; hMPV, human metapneumovirus; HPIV, Human parainfluenza virus; JUNV, Junin virus; LCMV, lymphocytic choriomeningitis virus; MACV, Machupo virus; MeV, measles virus; NiV, Nipah virus; RVFV, Rift fever valley virus; UUKV, Uukuniemi virus; VSV, vesicular stomatitis virus; HIV-1, human immunodeficiency virus type 1.

2.2.1.2 Viral replication stage

Replication is the key process in the virus life cycle required for persistence and the next round of infections. Viruses from different families usually adopt different strategies for replication but are still under the same constraint of using host translation machinery for the synthesis of viral proteins (**Fig 2.3**) [7]. Exemplified by HIV-1, its replication is initiated with the disintegration of capsid (encoded by *gag*). Its reverse transcriptase (encoded by *pol*) then transcribes the viral RNA into DNA (Step 3 in **Fig 2.2**). This procedure involves frequent recombination (combining of genetic material from different virus RNAs) and is usually error-prone (generating mutations) due to the lack of proofreading activity [93]. Consequently, it causes a rapid evolution of HIV-1 and may promote the emergency of new virus variants, for example, selection for drug resistance [94]. In the host nucleus, its integrase (encoded by *pol*) integrates the viral DNA into the host cell genome [95] (Step 4 in **Fig 2.2**). Finally, multiple copies of new HIV-1 RNAs are generated by the host transcription machinery, some of which are translated into new HIV proteins (Step 5~6 in **Fig 2.2**).

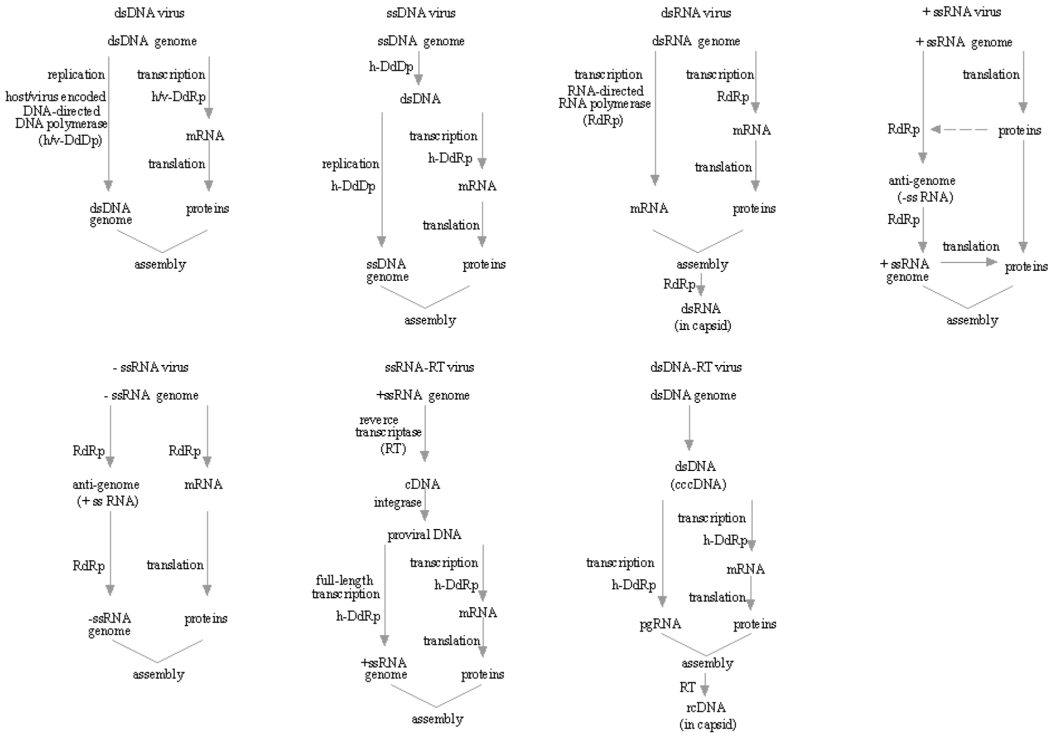


Fig 2.3 Illustration of the replication process for different viruses. The mRNA used for translation are transcribed from DNA in dsDNA/ssDNA viruses and from RNA in dsRNA/-

ssRNA viruses. In ssRNA-RT viruses, they are produced through a path of RNA-DNA-mRNA while in dsDNA-RT, they are produced through a path of DNA-RNA-DNA-mRNA. +ssRNA viruses use their mRNA for translation directly. Figure is reproduced from the KEGG MEDICUS [96] under the Creative Commons licence 4.0. Abbreviations: dsDNA, double-stranded DNA; ssDNA, single-stranded DNA; dsRNA, double-stranded RNA; +ssRNA, positive sense single-stranded RNA; -ssRNA, negative sense single-stranded RNA; ssRNA-RT, single-stranded RNA with a DNA intermediate; dsDNA-RT, double-stranded DNA with a RNA intermediate.

2.2.1.3 Viral release stage

Release is the last stage of the viral life cycle. After the production of new viral protein through host translation machinery (Step 6 in **Fig 2.2**), the assembly process integrates viral proteins and essential components to form virions at the host plasma membrane, internal membrane or in the cytoplasm [97] (Step 7 in **Fig 2.2**). Particles of non-enveloped viruses (e.g. human parvovirus B19) then can be released through the mode of cell lysis [98]. By contrast, enveloped viruses (e.g., HIV-1) use a different mode for the release. They acquire their external envelopes through the envelopment and budding process, during which the host membrane is gradually constricted until being severed to release the virus particle [97] (Step 8~9 in **Fig 2.2**). This process is mediated by some machineries such as endosomal sorting complex required for transport (ESCRT) [99]. However, successful release does not imply a mature status of the progeny virus. They still have to undergo many maturation steps to become an infectious virus for the next round of infections. Following the example given in **Section 2.2.1.2**, the assembly, envelopment and budding procedures produce immature HIV-1. Shortly after the release from the host cell surface, its protease (encoded by *gag-pol*) induces the proteolytic processing of Gag precursor to form a mature virion [100,101].

2.2.2 Interplay between viruses and their hosts

Viruses establish an intimate molecular-level exploitation of their hosts during their life cycles from entry to exit. Pro-viral molecular interactions between viruses and their host dependency factors (HDFs) permit their frequent exploitation of diverse regulation and signalling pathways in hosts [4]. However, when infection happens, the host immune system quickly responds and mounts the first defensive line to prevent viral invasion and to inhibit viral replication [102]. These initial immune responses are mediated by innate anti-viral cytokines secreted from host cells after the infection, which are known as interferons (IFNs)

[103]. IFN signals trigger the expression of IFN-stimulated genes (ISGs), which makes the host cell refractory to be further infected [104]. Viruses in turn have evolved complex strategies to antagonise host immune responses (**Table 2.2**). As exemplified in the **Fig 2.4**, toll-like receptor (TLR) signalling pathway can recognise virus-associated molecular patterns and activate cellular factors to regulate the expression of type I interferons (IFNs) [105]. Some of these factors such as nuclear factor kappa B subunit 1 (NFKB1, alias symbol: NF-kB) [106] and mitogen-activated protein kinase 1 (MAPK1, alias symbol: ERK) [107,108] are anti-viral and capable to perform pro-host molecular interactions with the virus. In KEGG database [96], HIV-1 was recorded to commandeer TLR signalling pathway by inflicting pro-viral inhibition effects on NFKB1 [109], RELA proto-oncogene (RELA, previous symbol: NFKB3) [110], MAPK1 [111], which furthers disrupt IFN signal transduction and host immune responses [112]. Interestingly, NFKB1, RELA and MAPK1 are also anti-viral [106,107,113-116], which means their interaction with HIV-1 occurs in both directions and so can be characterised as 'bidirectional'.

Table 2.2 Some strategies used by viruses to escape from host immune response.

Strategy	Virus	Ref.
Hiding their viral genome to escape from IFN detection.	Poxviruses	[117]
Avoiding binding to cellular sensors of viral infection.	KSHV	[118]
Mimicking of host-like short linear motifs.	HPV	[119]
Mimicking of host-like structural motifs.	SARS-CoV-2	[120]
Inactivating cellular factors involved in IFN signal transduction.	HCMV	[121]
Regulating phosphorylation to disrupt IFN signaling.	HIV-1	[122]
Regulating ubiquitinylation that controls IFN induction	Influenza B	[123]
Cleavage of factors essential for IFN responses.	Dengue virus	[124]
Degradation of factors essential for IFN responses.	Zika virus	[125]
Inhibiting host transcription of anti-viral genes.	Adenovirus	[126]
Interfering RNA processing of anti-viral genes.	VSV	[127]
Preventing synthesis of anti-viral genes.	HCV	[128]
Sequestering IFNs for their receptors.	EBOV	[129]

Abbreviations: IFN, interferon; KSHV, Kaposi's sarcoma-associated herpesvirus; HPV, human papillomavirus; SARS-CoV-2; HCMV, human cytomegalovirus; HIV-1, human immunodeficiency virus type 1; VSV, vesicular stomatitis virus; HCV, hepatitis C virus; EBOV, Ebola virus.

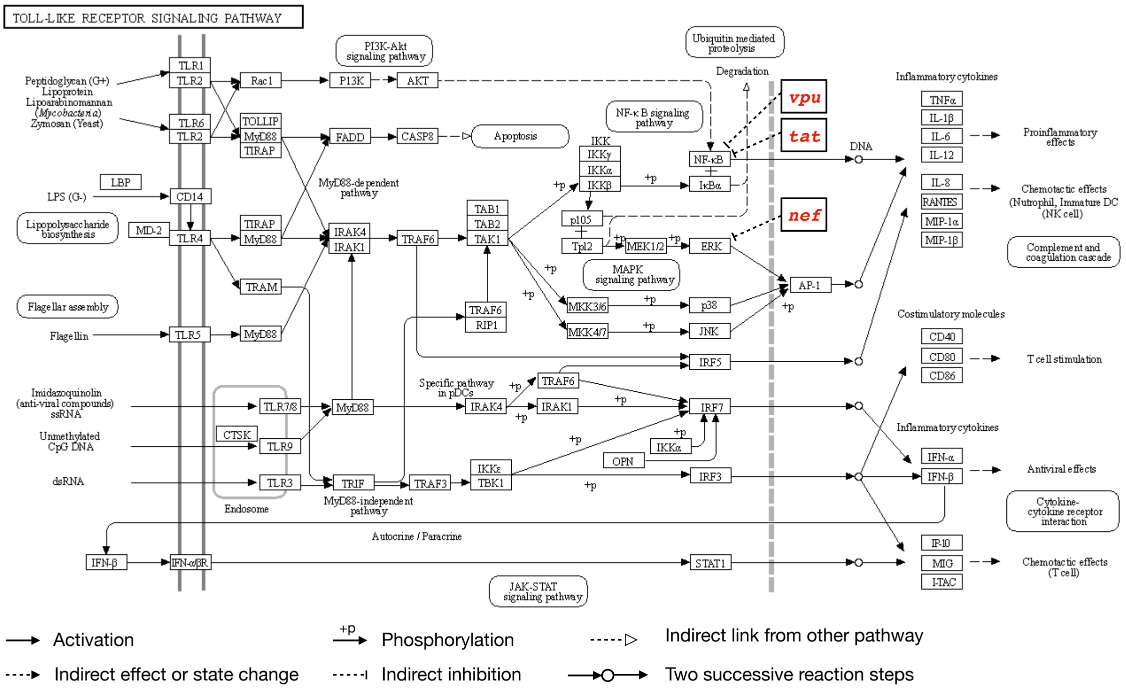


Fig 2.4 Example of viral hijacking of host cell machinery. Here, *vpu*, *tat* and *nef* of HIV-1 (highlighted in red) commandeer TLR signalling pathway by indirectly inhibiting NF-kB and ERK during the infection. Figure is reproduced from the KEGG MEDICUS [96] under the Creative Commons licence 4.0. Abbreviations: HIV-1, human immunodeficiency virus type 1; TLR, toll-like receptor; NF-kB, nuclear factor kappa B subunit 1; ERK, mitogen-activated protein kinase 1.

Reliable data about virus-host molecular interactions are crucial to precise understanding of the mechanisms underlying viral infection and associated pathogenesis. There are many *in vitro* methods to detect protein-protein interactions (PPIs) between viruses and their hosts, e.g., co-immunoprecipitation (co-IP) [130], and pull-down assays [131]. Exemplified by Co-IP, it is conducted based on the concept of immunoprecipitation (IP). It aims to enrich a protein complex consisting of one known antigen protein (bait, e.g., cloned viral protein) and unknown proteins (prey, e.g., suspected virus-interacting proteins) that are bound directly or indirectly to the bait [130]. After pulling out the protein complex with the immobilized antibody of 'bait' protein, the unknown 'prey' proteins can be identified with many technologies such as western blot [132]. Pull-down assay is analogous with co-IP in methodology but uses the 'bait' protein to purify any interacting proteins in the lysate [131]. Additionally, crosslinking bait and prey proteins with a covalent bond provides another effective solution to identify virus-host interactions especially when such interactions are transient [133]. The successful detection of PPIs contributes to the establishment of an integrative knowledgebase that can be used by computational approaches for the virus-host research. Virus-PPI data can be retrieved from many databases including UniProt [2], IntAct [134], BioGRID [135],VirHostNet [136], VirusMentha [137], and viruses.STRING [138].

2.3 Application of supervised machine learning in virus research

Machine learning (ML) researchers are developing better automatic algorithms to solve recognition, classification and prediction problems based on available data [139]. ML provides a high-throughput way to improve the understanding of biological activities and address specific tasks in virology [140]. Supervised learning is an important branch of machine learning. It uses known sample data to train models for a desired classification or prediction of new samples. Typically, these 'known sample data' contain two main parts: labels indicating real values corresponding to samples (for regression problems) or indicating classes (for classification problems), and feature vectors quantifying biological observations. The rapid development of DNA microarray [141], polymerase chain reaction (PCR) [142], RNA sequencing [143] and assay [130,131,133] technologies permits a diverse label source and sufficient feature profiles for supervised learning (**Table 2.3**).

Table 2.3 Common sources for data curation.

Source	Description	Ref.
GenBank	Database compiling publicly available DNA sequences.	[144]
Ensembl	Database compiling genomic data for vertebrate species.	[145]
RefSeq	Database compiling nucleotide sequences and their protein products.	[146]
UniProt	Database compiling protein sequence and function information.	[2]
PDB	Database compiling protein tertiary structures.	[147]
GISAID	Database compiling virus-related genetic, clinical and epidemiological data.	[148]
NCBI Virus	Database compiling viral sequences.	[149]
KEGG	Database integrating systems, genomic, chemical and health information.	[96]
Reactome	Database compiling molecular details of cellular processes.	[150]
SIGNOR	Database compiling the information of signaling network.	[151]
GTEx	Project compiling tissue-specific gene expression and regulation data.	[152]
ELM	Database compiling experimentally validated sequence motifs.	[153]
DisProt	Database compiling intrinsically disordered proteins.	[154]
SCOP	Database classifying protein and domains.	[155]
Pfam	Database classifying protein sequences into families and domains.	[156]
CDD	Database compiling conserved domains.	[157]
3DID	Database compiling domain-peptide and domain-domain interactions.	[158]
IntAct	Database compiling molecular interactions.	[134]
BioGRID	Database compiling biological interactions.	[135]
MINT	Database compiling molecular interactions.	[159]
STRING	Database compiling known and predicted associations between proteins.	[160]
HIPPIE	Database compiling human PPIs.	[161]
viruses.STRING	Database compiling virus-host PPIs.	[138]

VirHostNet	Database compiling virus-host PPIs.	[136]
VirusMentha	Database compiling virus-host PPIs.	[137]

Abbreviations: PDB, RCSB protein data bank; SIGNOR, signalling networking open resource; GTEx, genotype-tissue expression; ELM, eukaryotic linear motif; SCOP, structural classification of proteins database; CDD, conserved domain database; 3DID, three-dimensional interacting domains; BioGRID, biological general repository for interaction datasets; HIPPIE, human integrated protein-protein interaction reference; PPI, protein-protein interaction.

2.3.1 Supervised learning in regression problems

In regression problems, supervised learning aims to find a fitting curve that best fits the given label-feature sets [162]. Since the outcome of a biological event (represented by the label) is usually dependent on more than one factor (quantified by the feature), the applications of simple regression models are limited in biological research [163]. Alternatively, multiple regression models are used, for example, in some recent research to predict the epidemic trend of coronavirus disease 2019 (COVID-19) [164,165]. A linear multiple regression can be formulated as:

$$\widehat{y_i}(\mathrm{X}) = \beta_0 + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \cdots + \beta_n \times x_{in} \qquad \mathbf{(2.1)}$$

where $\widehat{y_i}$ is the predicted dependent variable (label) of sample X, n is the dimension of the feature vector; β_i is the slope coefficient for explanatory variable x_i (feature), and β_0 is the vertical intercept. Its performance can be evaluated by the coefficient of determination, which is also known as R-squared (R^2) [166]. This criterion measures how strong the linear relationship is between the label and feature vectors:

$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{y_i})^2}{\sum_i (y_i - \overline{y_i})^2} \qquad \mathbf{(2.2)}$$

where y_i, $\overline{y_i}$, and $\widehat{y_i}$ are the real-value, mean value and predicted value of the dependent variable (label).

2.3.2 Supervised learning in classification problems

In classification problems, supervised learning aims to find the optimal decision boundary or surface that partitions the feature space into two or more parts. Unlike its usage in regression problems, here, supervised learning produces qualitative mapping indicating the class to which a testing example belongs to [167]. In a biological context, such mapping can be exemplified as protein sequence to, e.g. its intrinsic disorder status (binary classification)

[168], protein structure to function (multi-class classification) [169], etc. There are many learning algorithms designed to solve the classification problems in virus research (**Table 2.4**) [170,171].

Table 2.4 Examples of supervised learning algorithms used in virus research.

Research focus	Learning algorithms	Ref.
Predicting HIV-1-human interactions	SVM, RF	[172-174]
Identifying SARS-CoV-2 infections	SVM, RF, NN	[175-179]
Forecasting dengue epidemics	SVM, DT, LR	[180]
Predicting EBOV-human interactions	SVM, KNN, DT, NB	[140]
Predicting HCV-human interactions	SVM	[181]
Predicting HPV-human interactions	SVM	[181]
Predicting human adaptation of influenza A virus	SVM, RF, DT, NN	[182]
Mapping the transmission risk of Zika virus	RF, NN	[183]
Predicting human-infecting viruses	SVM, RF, KNN, NB	[3,184]

Abbreviations: HIV-1, human immunodeficiency virus type 1; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; EBOV, Ebola virus; HCV, hepatitis C virus; HPV, human papillomavirus; LR, logistic regression; NB, naïve Bayes; KNN, k-nearest neighbors; DT, decision tree; RF, random forest; NN, neural network; SVM, support vector machine.

Logistic regression (LR) is the simplest learning algorithm to model the probability of an occurrence of a biological event such as phenotypes and diseases [185,186]. Its output can be formulated as:

$$P(\mathrm{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_n \times x_n)}} \tag{2.3}$$

where β_i is the slope coefficient for explanatory variable x_i (feature), n is the dimension of the feature vector, and $\beta_0$is the vertical intercept.

Similar to LR, Naïve Bayes (NB) algorithm is also probability-based. It considers the contribution of each feature to the biological event conditionally independent [187,188]. Its output can be formulated as:

$$P(c_j|x_1, x_2, \ldots x_n) = \frac{P(x_1|c_j) \times P(x_2|c_j) \times \ldots \times P(x_n|c_j) \times P(c_j)}{P(x_1) \times P(x_2) \times \ldots \times P(x_n)} \tag{2.4}$$

where c_j is the j^{th} class of the given data and n is the dimension of the feature vector.

K-nearest neighbors (KNN) algorithm is designed based on the principle that similar samples belong to the same class (**Fig 2.6**) [189]. Discrete features in the input data are transformed by the algorithm into a multi-dimensional space, and then assigned different contribution weights to their neighbors based on distance metrics such as Euclidean distance, Minkowski distance, and Mahalanobis distance [190,191]. Its output is determined by class labels of neighbors surrounding the testing sample:

$$\widehat{y_i}(\mathrm{X}) = \frac{\sum_{x_i \in N_k} y_i}{k} \tag{2.5}$$

where $\widehat{y_i}$ and y_i is the real-value and predicted value of the dependent variable (label), and N_k contains k closest neighbors of sample X in the feature space. The k-value is crucial to the performance of a KNN classifier (**Fig 2.5**) but is usually determined empirically, e.g., chosen as the number close to the square root of the data size [192,193].

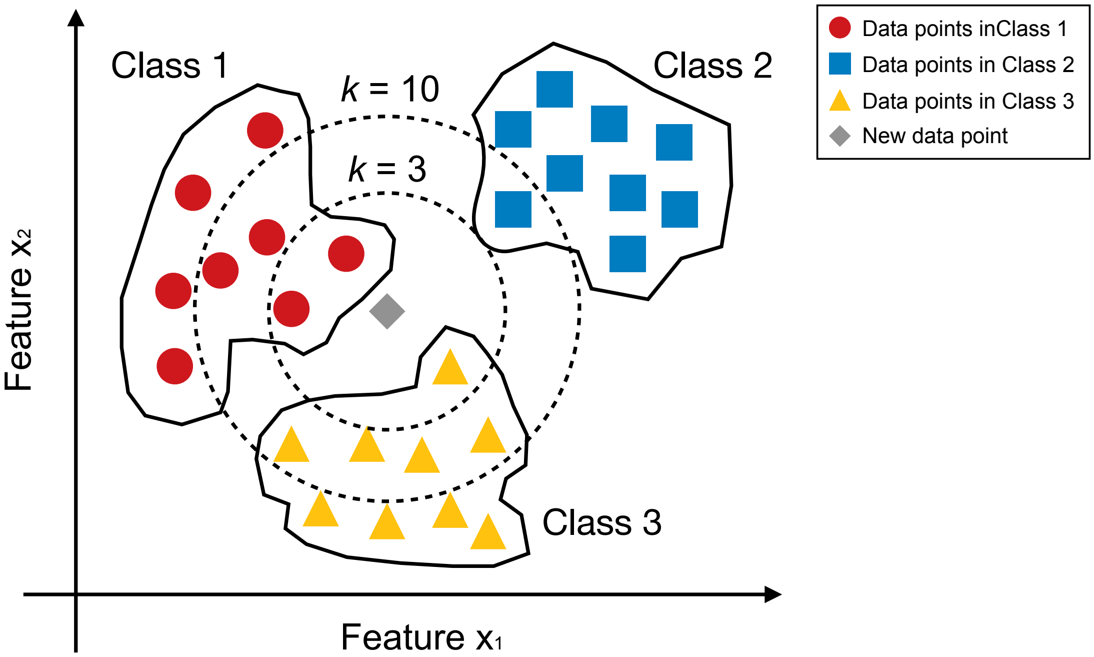


Fig 2.5 Example of KNN classifier in a two-dimensional feature space. The tested sample (grey diamond) is classified as Class 1 when using a k-value of three and is classified as Class 2 when using a k-value of ten. Abbreviations: KNN, k-nearest neighbors.

Decision tree (DT) [194] is a flowchart-like structure where each chance node represents an estimation on the feature and each leaf node determines the class of the data point. They can be used in multi-classification problems [195]. Entropy theory [196] contributes to the construction of a good DT model. The best split of DT is achieved by maximising the entropy gain [196]. DT usually performed well but lacks stability as even a slight change in the training data may alter the structure of the whole model.

Random forest (RF) is a collection of multiple random DTs. RF algorithm is initialised by randomly selecting samples from the original data with replacement. It builds several bootstrapped datasets with the same size as the original one. This procedure is known as bootstrapping [194]. A DT is then trained for each bootstrapped dataset independently with randomly selected features. These two random procedures make the algorithm less sensitive to the original training data. The final classification result is determined by the results of all DTs through a weighted mechanism such as majority voting [197].

Last but not the least, support vector machine (SVM) models are one of the most popular machine learning methods in use in many fields. It has been frequently used in virus research, e.g., to identify infection and predict virus-host molecular interactions (**Table 2.1**). It uses nonlinear kernel function (e.g., radial basis function and polynomial [198]) to transform the data that are non-linearly separable to high-dimensional space for better linear classification [199]. The main purpose of SVM modeling is to find a maximum-margin hyperplane to separate two classes in the feature space (**Fig 2.6**). Position and orientation of the hyperplane are determined by the surrounding data points (support vectors) (**Fig 2.6**) [200]. In the transformed high-dimension space for a binary classification problem, such maximum-margin hyperplane can be formulated as:

$$W^T \cdot \mathrm{X}_i - b = 0 \qquad \textbf{(2.6)}$$

where W is the normal vector of the hyperplane, X_i is the real vector representing features, and b is the bias term. New data points are then classified into the positive or negative class based the output produced by the signum function of $(W^T \cdot \mathrm{X}_i - b)$.

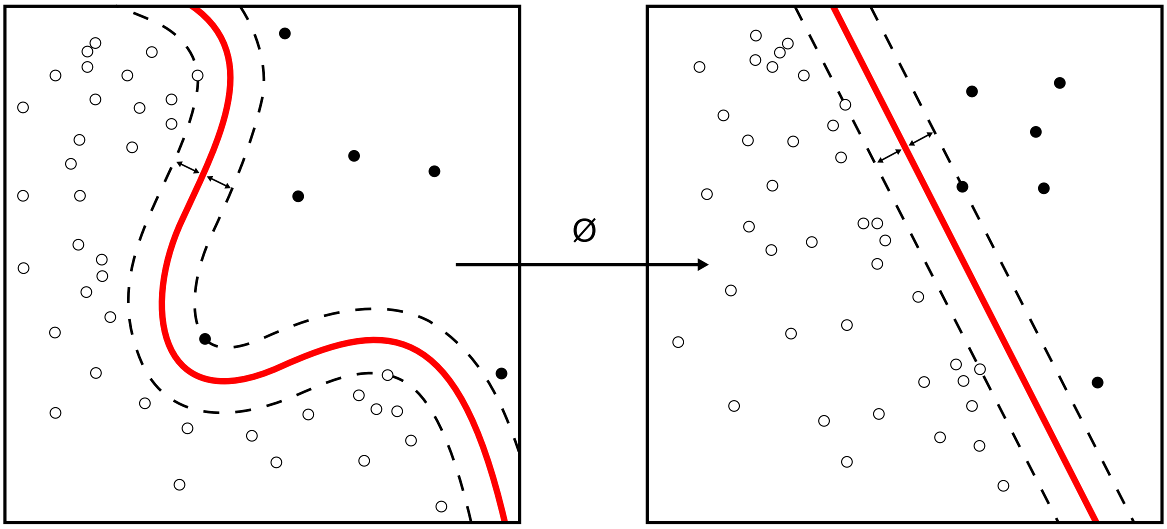


Fig 2.6 Key rationale of SVM. Nonlinear kernel function ϕ maps the data from a low-dimensional space (left) to a high-dimensional one (right). Red lines partition the two classes (black and white data points) in both spaces. Distance between two dotted lines is the

maximum margin between two classes. Figure is taken from Wikimedia Commons under the Creative Commons licence 4.0. Abbreviations: SVM, support vector machine.

2.3.3 Evaluation for classification problems

In the classification problem, samples in the dataset are required to differ from each other to an appropriate degree as similar or duplicated samples are expected to be well-identified when some of them are used to train the machine learning model. Sequence similarity or identity is widely used to control it in machine learning-based biological projects [201].

There are two main evaluation procedures in classification problems. The first one is processed in the training stage through N-fold cross-validation [202]. Exemplified by five-fold cross-validation, first, the training dataset is divided into five subsets. Each subset is then tested on the machine learning model generated from the rest four subsets. The overall performance of the machine learning method in the training stage is evaluated by combining the testing results on each subset. The second evaluation procedure is processed in the testing stage. Testing samples are usually required to be independent of the samples used for training. They can either be separated from the main dataset before the training stage or be collected from other related projects.

In two-class classification problems, the testing samples are labelled either positive or negative by the machine learning model. It results in two successfully predicted outcomes including true positives (TP) and true negatives (TN). Incorrectly predicted positive and negative samples are called false negatives (FN) and false positives (FP), respectively. These four outcomes are used to assess the prediction performance of the aforementioned classifiers. Sensitivity, recall, or true positive rate (TPR) is used to assess the performance in predicting positive samples, which can be formulated as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.7}$$

Specificity, selectivity, or true negative rate (TPR) is used to assess the performance in predicting negative samples, which can be formulated as:

$$Specificity = \frac{TN}{TN + FP} \tag{2.8}$$

Precision is used to measure the quality of the predicted positives and in tension with the criterion of sensitivity/recall/TPR. It can be formulated as:

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

Accuracy, F_1 score, and the Matthews correlation coefficient (MCC) provide comprehensive assessments for the prediction performance. They can be formulated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{2.10}$$

$$F_1\ score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \tag{2.11}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{2.12}$$

In addition, some other criteria such as the receiver operating characteristic (ROC) curve and precision-recall curve are also widely used in machine learning-related projects [201,203].

CHAPTER THREE

3. PREDICTION OF HIV-1-HOST MOLECULAR INTERACTIONS

3.1 Abstract

Human immunodeficiency virus type 1 (HIV-1) continues to be a major cause of disease and premature death. As with any virus, HIV-1 exploits a host cell to replicate. Improving our understanding of these molecular interactions between virus and human host proteins is crucial for a mechanistic understanding of virus biology, infection and host antiviral activities. This knowledge will permit the identification of host molecules for targeting by drugs with potential antiviral properties. Here, we propose a data-driven approach for the analysis and prediction of HIV-1-host interacting proteins with a focus on the directionality of the interaction (host-dependency versus antiviral factors). Using support vector machine learning models and features encompassing genetic, proteomic and network properties, our results reveal some significant differences between virus interacting human proteins (VIPs) and non-virus interacting human proteins (non-VIPs). As assessed by comparison with the HIV-1 infection pathway data in the Reactome database (sensitivity > 90%, threshold = 0.5), these models have good generalization properties. We demonstrate that the direction of HIV-1-host molecular interactions is also predictable due to different characteristics of 'forward'/pro-viral versus 'backward'/pro-host proteins. Additionally, we infer the previously unknown direction of interaction between HIV-1 and 1351 human host proteins. A web server for performing predictions is available at http://hivpre.cvr.gla.ac.uk/.

3.2 Introduction

Human immunodeficiency virus type 1 (HIV-1) is the cause of acquired immunodeficiency syndrome (AIDS) and constitutes a major cause of human disease and associated comorbidities. Virus infection involves viral molecules exploiting the host cell in order to replicate. The engagement of the HIV-1 envelope glycoprotein and cell-surface receptors, CD4 and either the membrane-spanning C-C motif chemokine receptor 5 (CCR5) or C-X-C motif chemokine receptor 4 (CXCR4), initiates virus attachment and entry into the cell [204-206]. Virus molecules including the HIV-1 regulatory factors (*tat* and *rev*) and accessory proteins (*vpr*, *vif*, *nef*, and *vpu*) ensure viral persistence, replication, dissemination, and transmission by modulating the surface and intracellular environment of the infected cell [207-211]. The production of HIV-1 *gag*/*pol* polyproteins is essential for assembly, release and maturation of new virions [101]. Protein-protein interactions (PPIs) between virus and host molecules enable the virus to infect and exploit the host system so as to use specific cell sub-systems to replicate and persist despite the host immune response [101,204-206,209-213]. Conversely, there are many human host proteins that function as antiviral factors and are part of the immune response capable of controlling virus numbers by counteracting the infection [214-216]. Improving our understanding of these HIV-1-host PPIs can provide insights into the molecular mechanisms underlying pathogenesis. Determining the nature of virus-host interactions [217] is thus of importance for the discovery of potential host inhibitors or targets [218] for antiviral therapeutics exemplified by the CCR5 antagonist maraviroc [219]. Intuitively, there are many more possible drug-targets in the host compared to HIV's compact genome, which codes for relatively few proteins. To efficiently direct laboratory experiments and make use of rapidly accumulating data in the post-genomic era, the development of efficient *in silico* approaches has become an important area of research focus.

Over the past few years, several computational studies on HIV-1 have characterised attributes of HIV-1 interacting human proteins based on various data, e.g., gene ontology (GO) annotations [220], interaction network profiles [221], disease pathways [222], and post-transcriptional modification profiles [223]. A hierarchical biclustering system was constructed by MacPherson *et al* [217] to designate HIV-1-host PPIs directionality, polarity, and control properties. This research demonstrated how VIPs can be grouped by related virus-associated perturbations that can be distinguished from non-VIPs. Curation of the extensive experimental literature has permitted an HIV-1-host PPI dataset to be compiled [220]. This can be used for the purpose of modeling and predictions via machine learning algorithms. For example, in [173] a random forest (RF) model was constructed by including 35 features for the prediction of HIV-1-host interaction pairs. Further work integrated semi-supervised learning, multi-task learning and neural networks [224]. Subsequently, biclustering-based approaches [225,226] were applied along with an association rule mining technique. Supervised machine learning methods [172,174] have also been implemented using the support vector machine (SVM) and datasets with different positive-to-negative ratios. Based on the assumption that proteins with similar sequence or structural patterns tend to share common interaction partners, studies have also predicted possible HIV-1-host interaction pairs by integrating protein short linear motifs (SLiMs) or protein structure [227-229].

Owing to the contribution made by computational approaches [140,170,230,231], it is possible to obtain a list of potential VIPs with high confidence. However, there are still many improvements that can be made. First, the majority of the published methods [172-174,224-229] are highly dependent on the type of interacting HIV-1 molecules. The defined non-VIPs could be false negatives relative to different HIV-1 proteins [232], for example, non-*env*-interacting protein cyclin T1 (CCNT1) interacts with HIV-1 during infection as it is targeted by *gag* and *tat* proteins [207,233,234]. Second, the nature of the molecular interaction is important for understanding pro-viral interactions versus host antiviral activities. Crudely this can be broken down to the 'directionality' of the interaction [217]: 'forward'/pro-viral versus 'backward'/pro-host proteins; a prediction task addressed for the first time in this study. Additionally, there also exists a group of human host proteins having both pro-viral and pro-host properties, i.e., are 'bidirectional' in nature, for example, CD4 [235,236]. Third, although the expansion of feature coverage provides a clearer picture for the classification problem, it also induces a series of problems such as feature redundancy [237] and overfitting [238-240].

To address these points, we proposed a computational approach for the analysis and prediction of HIV-1-host molecular interactions (presented diagrammatically in **Fig 3.1**). Contrary to previous prediction-based studies [172-174,224-229], we introduced a broader definition for HIV-1 interacting proteins. Human proteins having interactions with one or multiple HIV-1 proteins were all referred to as VIPs. Non-VIPs represented those human proteins without any record of being directly involved in an HIV-1 interaction. We designed three procedures to maximise the set of non-VIPs and to reduce their chance of being false negatives. Three tags: 'forward' (pro-viral), 'backward' (pro-host) and 'bidirectional' (pro-viral and pro-host) were assigned to VIPs to capture the direction of the virus-host interaction during the HIV-1 life cycle [217,241]. In total, we encoded 671 features based on the data retrieved from multiple databases [161,241-246] to characterise the human proteins at genetic, transcriptomic, proteomic and network levels. We also measured the contribution of individual features and different feature combinations. We constructed different feature sets via two feature selection schemes to generate prediction models on the training datasets with SVM [200]. Performance on the testing datasets demonstrated good prediction quality and generalization capability of our VIP prediction models. A web server for HIV-1-host molecule prediction is available at http://hivpre.cvr.gla.ac.uk/.

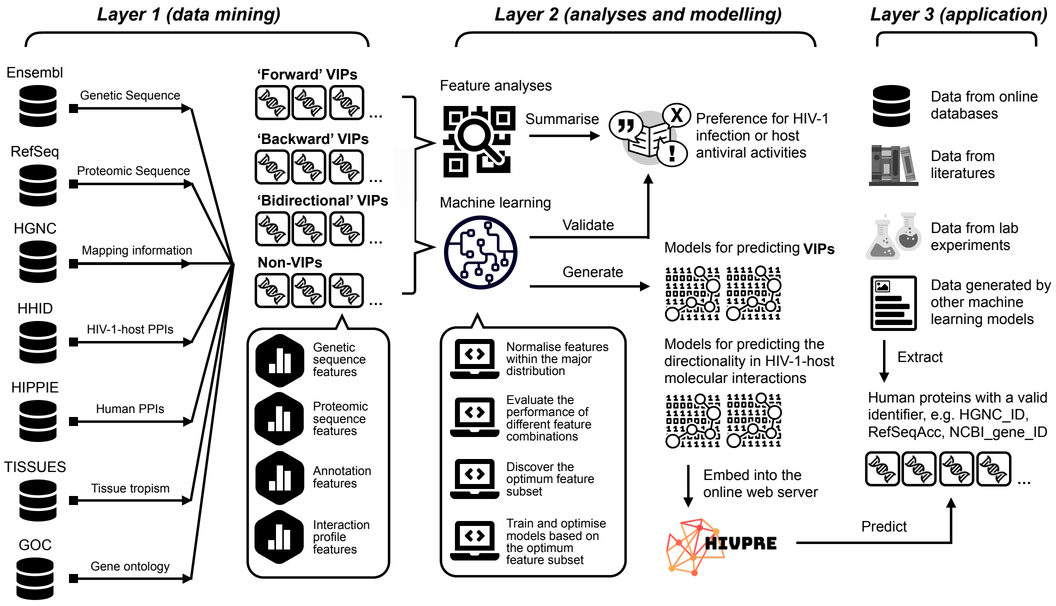

Layer 1 (data mining)
Layer 2 (analyses and modelling)
Layer 3 (application)
Ensembl
Genetic Sequence
RefSeq
Proteomic Sequence
HGNC
Mapping information
HHID
HIV-1-host PPIs
HIPPIE
Human PPIs
TISSUES
Tissue tropism
GOC
Gene ontology
'Forward' VIPs
'Backward' VIPs
'Bidirectional' VIPs
Non-VIPs
Genetic sequence features
Proteomic sequence features
Annotation features
Interaction profile features
Feature analyses
Summarise
Preference for HIV-1 infection or host antiviral activities
Machine learning
Validate
Generate
Models for predicting VIPs
Models for predicting the directionality in HIV-1-host molecular interactions
Normalise features within the major distribution
Evaluate the performance of different feature combinations
Discover the optimum feature subset
Train and optimise models based on the optimum feature subset
Embed into the online web server
HIVPRE
Predict
Data from online databases
Data from literatures
Data from lab experiments
Data generated by other machine learning models
Extract
Human proteins with a valid identifier, e.g. HGNC_ID, RefSeqAcc, NCBI_gene_ID

Fig 3.1 Diagrammatic representation of the project pipeline separated into three procedural layers. The components are from Wikimedia Commons under the Creative Commons licence 4.0. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, virus interacting human proteins; non-VIPs, non-virus interacting human proteins; HGNC, HUGO Gene Nomenclature Committee; HHID, the HIV-1 Human Interaction database; HIPPIE, the Human Integrated Protein-Protein Interaction rEference database; GOC, Gene Ontology Consortium.

3.3 Methods

3.3.1 Dataset curation

We retrieved 16215 HIV-1-host PPI records from the HIV-1 Human Interaction Database (HHID) (https://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/) (ver. September 2017) [241], involving 7120 HIV-1-host interaction pairs and 3854 distinct VIPs (**S3.1 Data**). Protein sequences for the VIPs were collected from the NCBI's RefSeq database (ver. GRCh38) [243]. To avoid over-representation of similar protein sequences in the dataset, we grouped them into 2881 clusters using the CD-HIT [247,248] with a threshold of 40% sequence similarity [247,248] and picked the longest sequence in each cluster as representative. This was to prevent producing feature vectors with high similarity, biasing the prediction performance. These 2881 representative VIPs formed our positives in dataset S1 (**Table 3.1**).

Following the methods of MacPherson *et al.* [217], we assigned the VIPs direction tags: 'forward', 'backward' or both/'bidirectional' (**Fig 3.2**). These directions are deduced from the outcome of HIV-1-host molecular interactions. Keywords describing the virus-to-host interaction such as 'regulates', 'modulates' and 'stimulates' hint a 'forward' direction while some like 'regulated by', 'modulated by' and 'stimulated by' indicate a 'backward' direction. The VIPs obtained both 'forward' and 'backward' tags are reassigned a 'bidirectional' tag. However, some HIV-1-host interactions were generally described without obvious preference indicating their directionality, for example, using keywords of 'interact with' and 'binds'. The VIPs with such interactions are classified as 'undefined' if they do not have a 'bidirectional' tag. Detailed designation about the interaction keywords are available online at https://doi.org/10.1371/journal.pcbi.1000863.s004.

'Forward' VIPs (e.g., C-X-C motif chemokine ligand 10, CXCL10) are pro-viral proteins. These host molecules are targeted by HIV-1, so-called host-dependency factors [249], and have no recorded antiviral response to the infection during the virus life cycle [250]. 'Backward' VIPs (e.g., apolipoprotein L1, APOL1) are pro-host proteins, they are associated with control or inhibition of the viral infection [251]. 'Bidirectional' VIPs (e.g., CXCR4) are targeted by HIV-1 (forward direction) and can produce pro-host responses (backward direction) by influencing the same or different HIV-1 proteins during the viral infection. Some of these are potential therapeutic targets to inhibit virus replication by making host molecules unavailable to the virus [15,252]. Since the "direction" of some HIV-1-host molecular interactions have not been clearly defined, these VIPs were not included in our analysis. Collectively, we obtained 188 (~6.5%) 'backward', 1007 (~35.0%) 'forward', 335 'bidirectional' (~11.6%), and 1351 (~46.9%) 'undefined' VIPs from dataset S1 to construct another three datasets, i.e., S2, S3 and S4 for direction-related predictions (**Table 3.1**).

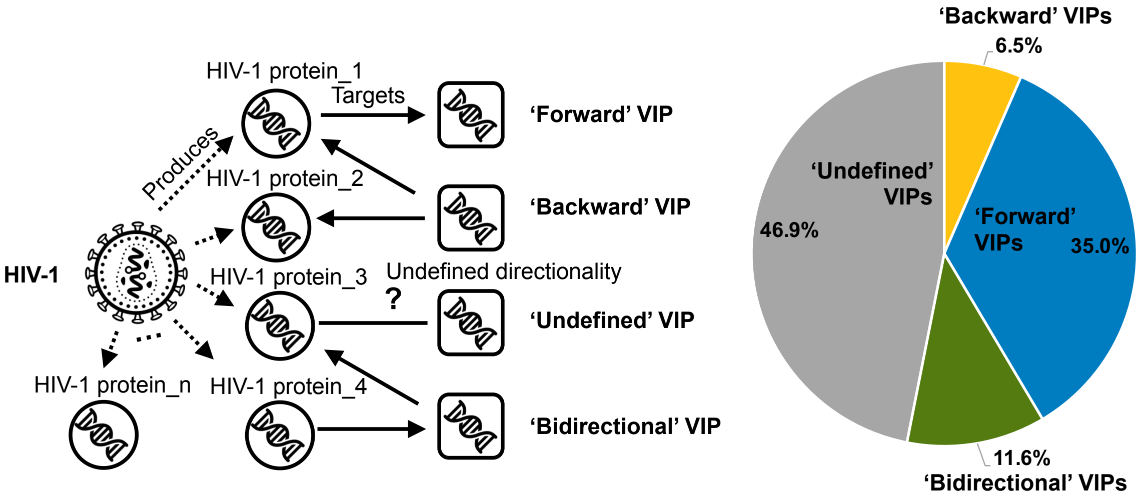


Fig 3.2 Representation of the characterisation of the types of virus interacting proteins (VIPs). VIPs were tagged as 'forward', 'backward' or 'bidirectional' based on the key words describing their interaction(s) with HIV-1 proteins [241] and directionality designated in MacPherson *et al.* (https://doi.org/10.1371/journal.pcbi.1000863.s004) [217]. The direction was classed as 'undefined' if this information is not available. The direction tag for each VIP is provided in **S1 Data**. Figure created using the BioRender (https://biorender.com/) under the Creative Commons licence 4.0. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; PPI, protein-protein interactions.

We performed three procedures to improve the quality of the non-VIP dataset and reduce potential false negatives. First, we chose human proteins produced by the canonical transcript since these proteins expressed the main function of the gene [253]. Second, human proteins sharing more than 40% sequence similarity with any of the reported 3854 VIPs in HHID were excluded to prevent sequences similar to the known VIPs overly influencing the modelling and predictions. Third, we controlled for the sequence similarity of non-VIPs at a 40% level as we did for VIPs. This would prevent predictions being influenced by similar combinations of feature vectors. As a result, we obtained 7261 nonredundant non-VIPs to form the negatives in dataset S1 (**Table 3.1**).

As we applied different criteria compared to others to classify our VIPs/non-VIP datasets, it was hard to make comparisons between our predictions and previous ones [172-174,224-229]. We introduced two testing datasets consisting of VIPs with high experimental confidence from Reactome [150] and Gordon *et al.* [254] in order to assess the generalization capability of our machine learning models. A breakdown of the VIPs and non-VIPs used in this study is listed in **Table 3.1** and more detailed information is provided in **S3.2 Data**.

Table 3.1 Breakdown of VIP and non-VIP datasets used.

Dataset[a]	Positives	Negatives
Main dataset S1	2881 VIPs	7261 non-VIPs
Training S1'	2304 VIPs	2304 non-VIPs
Independent testing S1"	577 VIPs	4957 non-VIPs
Main dataset S2	188 'Backward' VIPs	1007 'Forward' VIPs
Training S2'	150 'Backward' VIPs	150 'Forward' VIPs
Independent testing S2"	38 'Backward' VIPs	857 'Forward' VIPs
Reference dataset S3	335 'Bidirectional' VIPs	
Blind testing dataset S4	1351 'Undefined' VIPs	
Testing dataset S5	234 VIPs	
Testing dataset S6	356 VIPs	

[a]*Dataset S1 and S2 were constructed for the prediction of VIPs and their directionality in HIV-1-host PPIs. 80% of positives and an equal number of negatives were randomly selected for training while the remaining 20% of proteins were used for testing. Dataset S3 was constructed for prediction of 'bidirectional' VIPs while S4 was constructed for the prediction of putative 'forward', 'backward' or 'bidirectional' VIPs. Testing datasets S5 and S6 were retrieved from two resources with high experimental confidence: the HIV-1 infection pathway in the Reactome [150], https://reactome.org/PathwayBrowser/#/R-HSA-162906, and viral host-dependency epistasis map linked to HIV function [254]. The lists of proteins sampled for training and independent testing are provided in* ***S3.2 Data****.* Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, HIV-1 non-interacting human proteins.

3.3.2 Feature generation

In this study, we encoded 671 different features mainly from six online databases: Ensembl 98 [244], RefSeq (ver. GRCh38) [243], TISSUES 2.0 [245], Human Integrated Protein-Protein Interaction rEference database (HIPPIE) 2.2 [161], HHID (ver. September 2017) [241], and Gene Ontology Consortium (GOC) (ver. November 2019) [246]. Among them, 537 features were used to distinguish VIPs from non-VIPs while 584 features were used to investigate the directionality of HIV-1-host molecular interactions (**Table 3.2**). Based on the data sources, our encoded features could be divided into four groups: (1) genome-based, (2) proteome-based, (3) annotation-based and (4) interaction profile-based features. Source code for generating features are available at: https://github.com/HChai01/HIVPRE.

Table 3.2 Breakdown of encoded features.

Task	Catalogue	No. of features	Feature source
Predicting VIP	Genome-based	107	Genetic sequences and alignments
	Proteome-based	128	Protein sequences and predictors
	Annotation-based	292	GO mapping and tissue expression
	Interaction profile-based	10	Human PPIs
Predicting the directionality	Genome-based	107	Genetic sequences and alignments
	Proteome-based	164	Protein sequences and predictors

Annotation-based	292	GO mapping and tissue expression
Interaction profile-based	21	Human PPIs and HIV-1-host PPIs

Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; GO, gene ontology; PPI, protein-protein interaction.

3.3.2.1 Genome-based sequence features

We compiled 107 genome-based sequence features for each human protein which included alternative splicing, nucleotide composition, codon usage and a measure of evolutionary conservation. Information in the alternative splicing data was encoded into four features to represent the evolution of phenotypic complexity in human genes [255,256]: the number of transcripts, protein-coding transcripts or open reading frames (ORFs), exons and unit exon in transcripts (UET). Nucleotide composition represented the distribution of four basic nucleobases and their phosphodiester bonds-combinations, e.g., CpG, in the coding region of genetic sequences [257]. The usage of the existing 64 codons was calculated in each nucleotide sequence to reflect the balance between mutational biases and natural selection for translational optimization in different classes [258]. For evolutionary conservation, we collected the data from BioMart (ver. September 2019) [244] and calculated the number of paralogues, synonymous substitutions (ds), non-synonymous substitutions (dN) and the ratio of dN to dS within human paralogues and orthologues in four homininae genomes: chimpanzee, gorilla, orangutan and gibbon. These features were used to assess the selection on protein sequences and estimate mutational processes affecting the molecules [259].

3.3.2.2 Proteome-based sequence features

We encoded 251 features from proteome-based sequence data for the prediction of VIPs and their directionality. Discrete sequence information was calculated as amino acid compositions, while linear information was analysed from the perspective of SLiMs and intrinsic disorder. We generated 37 types of amino acid composition based on the differences in individual amino acids or their physiochemical attributes [260]. Ambiguous or other types of amino acids, e.g., selenocysteine, pyrrolysine etc. were masked as ‘X’ and ignored in this study. We used the MERCI program [261] to detect conserved sequence patterns as a result of strong purifying selection [262], obtaining 206 SLiMs overrepresented in the group of VIPs and ‘backward’ VIPs (Pearson’s Chi-squared test, $P<0.05$). We chose to retain as many ‘useful’ features as possible at this stage thus further correction procedure such as Benjamini-Hochberg correction was not applied to avoid type I error in multiple hypothesis testing [263,264]. The occurrence of these SLiMs was then split and encoded into 206 categorical features. Four features measuring the overall representation of VIP- or 'backward' VIP-enriched SLiMs were added as hedges against random error caused by data imbalance [265]. The disordered regions in human protein sequences were identified using the Espritz 1.3 [266] and IUPred 2A [267] as such regions have been linked to VIPs [268].

3.3.2.3 Annotation-based features

We encoded 292 annotation-based features with a binary system from the collected tissue and gene expression data. Among these, 66 features were generated by mapping the GO terms to the child term of three main GO root terms: molecular function (GO:0003674), cellular component (GO:0005575) and biological process (GO:0008150) [246]. They characterised the domain in which human proteins might be involved such as binding (child term of molecular function, GO:0005488), intracellular (child term of cellular component, GO:0005622) and metabolic process (child term of biological process, GO:0008152) when interacting or not interacting with HIV-1 molecules [220]. The remaining 226 features were encoded for 226 tissues based on the experimentally verified data in TISSUES [245]. They were used to reflect the association between tissue tropism and HIV-1 infections at the molecular level [269].

3.3.2.4 Interaction profile-based features

Interaction profile-based features were generated from HIV-1-host PPIs [241] and the human interactome [161]. We used 11 features to represent the degree to which a confirmed VIP was central to the life cycle of HIV-1 [204-206]. Specifically, one feature was encoded to count the number of HIV-1 gene-products interacting with human host molecules and the remaining ten were binary-encoded to capture the interaction relationships between the host molecule and the corresponding HIV-1 gene-product, e.g. *gag*, *tat* or the antisense protein gene *asp* [270]. We retrieved 332,701 experimentally verified human-human PPIs with at least medium confidence (HIPPIE score > 0.63) involving 17,607 human proteins from HIPPIE [161] to pinpoint proteins with potential pathological or therapeutic relevance [271,272]. The NetworkAnalyzer [273] was used to calculate ten different network features: the average shortest distance, degree, neighbourhood connectivity, betweenness, stress, closeness, eccentricity, radiality, topological coefficient, and clustering coefficient. Human proteins not involved in the human-human PPI network were assigned zero values on all of the aforementioned network features.

3.3.3 Supervised machine learning and feature selection

We applied a supervised machine learning method for the prediction tasks. We chose the SVM with the radial basis function [200] as the core classifier after comparing it with k-nearest neighbors (KNN), decision tree (DT) and random forest (RF) [170]. The SVM algorithm aims to find an appropriate hyperplane in the feature space for classifying the majority of positive and negative samples. It can tolerate the existence of some noisy or incorrect data but may be biased by different feature scales or imbalanced positive-to-negative ratios as it was designed to calculate the margin of the data [274]. Additionally, although the SVM algorithm can map the current feature space to a higher dimensional one for better classification [200], it is a sub-optimal strategy for including too many features for modelling even if they are all instructive. This can result in overfitting of the machine learning model [239] leading to a loss of robustness [275]. To address these points, we first used an undersampling strategy [265] to randomly construct balanced training datasets (**S3.2 Data**). Secondly, discrete features were normalised according to their majority distribution in order to share an equal range with categorical/binary features:

$$Norm(v) = \begin{cases} 1, v > UB(v) \\ \dfrac{v - LB(v)}{UB(v) - LB(v)}, LB(v) < v < UB(v) \\ 0, v < LB(v) \end{cases} \tag{3.1}$$

where $LB(v)$ and $UB(v)$ are the lower and upper bound representing the 5th and 95th percentile within the target feature values. Next, we used a SVM-based forward selection scheme with the evaluation of area under the receiver operating characteristic curve (AUC) to optimise the feature set for the general case (**Fig 3.3**). In this scheme, we first introduced the Fisher-Markov Selector [276] to measure the importance of individual features. The iteration of this scheme then started with the most important feature. The rest features were added one by one to the feature set in the following iterations based on their importance until the prediction model achieved its best AUC value. We assumed that the usage of better performing features are less likely to influence or damage the complementarity of features in the set, which is crucial to training and modelling [237]. This feature selection scheme produced two outcomes: the optimum feature set and the lowest number of features.

```
BEGIN
Initialisation: Balanced dataset $S_0 = \{(L_1, v_1^0), .., (L_1, v_n^0), (L_2, v_{n+1}^0), ..., (L_2, v_{2n}^0)\}$, original feature set $F_m = (f_1, f_2, ..., f_m)$, machine learning classifier $C$, feature evaluation algorithm $A$, prediction evaluation criterion $E$, loop pointer $i = 2$.
  (1) Evaluate the importance of individual feature $a = A(S_o)$.
  (2) Create descending rank list based on the feature importance in $a$, $L = (f_1', f_2', ..., f_m')$.
  (3) Use the most important feature to create feature set $F_1 = f_1'$.
  (4) Update feature vector $v_x^1$, dataset $S_1$ and evaluate the prediction, $P_1 = C(S_1)$, $e_1 = E(P_1)$.
  While $i \leq m$:
    (5) Update feature set to include one well-performed feature based on $L$, $F_i = (f_1', ..., f_i')$;
    (6) Update feature vector $v_x^i$, dataset $S_i$ and evaluate the prediction, $P_i = C(S_i)$, $e_i = E(P_i)$;
    (7) Calculate the improvement after including the new feature, $I_i = (e_i - e_{i-1})/e_{i-1}$;
    (8) Update loop pointer $i = i + 1$.
  End
  Output: $F_i$ achieving the best $e_i$ and $F_i$ achieving the best $I_i$.
END
```

Fig 3.3 The pseudo-code of the feature selection Scheme 1. We chose the SVM [200] as the base machine learning classifier and the Fisher-Markov Selector [276] to calculate the importance of an individual feature. AUC was chosen as the prime criterion to evaluate the prediction performance on datasets with multiple labels. Abbreviations: SVM, support vector machine; AUC, area under the receiver operating characteristic curve.

The complementarity of different features implies information synergies, which can be measured by calculating the change of system entropy after the introduction of additional features [196,237,277]. However, it is hard to discriminate if the combination of several random features can achieve better complementarity compared to using an equal number of well-performing features. The selection strategy requires reconsideration if the impact of feature synergy has overwhelmed the usage of 'important' features on the prediction performance. Here, we propose a second feature selection scheme by referring forward feature selection, backward feature elimination and exhaustive feature selection strategies [278]. Our second feature selection scheme took into account both feature importance and complementarity (**Fig 3.4**). As opposed to the first feature selection scheme (**Fig 3.3**), this scheme started with a set of features with good complementarity. It then contained two main branches: the first expands the coverage of features by introducing well-performing features, while the second reduces the dimension of the feature sets by removing badly-performing features.

```
BEGIN
Initialisation: Feature sets with good complementarity $F_0' = (f_1, f_2, \dots, f_s)$, the rest feature list $F_0 = (f_{s+1}, f_{s+2}, \dots, f_m)$, balanced dataset $S_0 = \{(L_1, v_1^0), .., (L_1, v_n^0), (L_2, v_{n+1}^0), \dots, (L_2, v_{2n}^0)\}$, machine learning classifier $C$, feature evaluation algorithm $A$, prediction evaluation criterion $E$, loop pointer $i = j = 1$.
  (1) Evaluate the importance of individual feature $a = A(S_o)$.
  (2) Create descending rank list for $F_0$ based on the feature importance in $a$, $L_1 = (f_1', f_2', \dots, f_{m-s}')$.
  (3) Create ascending rank list for $F_0'$ based on the feature importance in $a$, $L_2 = (f_1'', f_2'', \dots, f_s'')$.
  While $i \leq m - s$:
    (4) Update feature set to include one well-performed features based on $L_1$, $F_i = (f_1, \dots, f_s, f_1', \dots, f_i')$;
    (5) Update feature vector $v_x^i$, dataset $S_i$ and evaluate the prediction, $P_i = C(S_i)$, $e_i = E(P_i)$;
    (6) Update loop pointer $i = i + 1$.
  End
  (7) Determine $F_i$ achieving the best $e_i$.
  While $j < s$:
    (8) Update feature set to remove one badly-performed feature based on $L_2$, $F_j' = F_i - (f_1'', \dots, f_j'')$;
    (9) Update feature vector $v_x^j$, dataset $S_j$ and evaluate the prediction, $P_j = C(S_j)$, $e_j = E(P_j)$;
    (10) Update loop pointer $j = j + 1$.
  End
  Output: $F_j'$ achieving the best $e_j$.
END
```

Fig 3.4 The pseudo-code of the feature selection Scheme 2. We used an SVM [200] as the base machine learning classifier and the Fisher-Markov Selector [276] to calculate the importance of an individual feature. AUC was chosen as the prime criterion to evaluate the prediction performance on datasets with multiple labels. Abbreviations: SVM, support vector machine; AUC, area under the receiver operating characteristic curve.

3.3.4 Performance evaluation

In order to assess the performance of different feature subsets, we adopted five-fold cross-validation on training datasets (dataset S1' and S2'), in which human proteins were randomly divided into five nearly equal parts and further generated five different testing (one portion) and training (the remaining four portions) sets. The overall quality of prediction models constructed from the feature subset was then evaluated based on the produced prediction scores via six criteria including sensitivity, specificity, accuracy, precision, Matthews Correlation Coefficient (MCC) [279] and AUC on the combination of five separate testing results. The evaluation of other independent testing datasets was also processed with the aforementioned six criteria except in the case of the reference dataset S3, testing dataset S5 and S6, which only used sensitivity controlled by the threshold.

3.4 Results

3.4.1 HIV-1-host interaction pairs

Compared with available benchmark datasets [173,174,224,225,228], our main dataset S1 included more HIV-1-host PPI data than previous studies (**Fig 3.5A**). Initial statistical results on HIV-1-host interaction pairs indicated that the majority of HIV-1-host molecular interactions emerged during *env*-mediated membrane fusion [204-206] and *tat*-mediated transcellular transport [207,208,280]. In our main dataset S1, there were 996 (~35%) VIPs with interactions with multiple HIV-1 proteins. Some VIPs such as nuclear factor kappa B subunit 1 (NFKB1), interferon gamma (IFNG) and interferon beta 1 (IFNB1) are reported to interact with products produced by almost all HIV-1 genes [110,281-286]. We illustrate the preference of co-occurring HIV-1-host PPIs interfering or being induced by the same VIP in **Fig 3.5B**. This figure can be divided into ten clades according to the target HIV-1 gene product. It reveals a picture of host targets shared among HIV-1 gene products of *tat*, *env*, *gag*, *nef*, *gag-pol* and *vpr*, and the interaction preference underlying HIV-1 invasion, replication and assembly [101,204,210]. Despite the rank ordered by the number of HIV-1-host interacting pairs (**Fig 3.5A**), HIV-1 *tat*-interacting proteins gave higher interaction priority to *vpr* than *gag-pol*. On the other hand, *tat* was less involved in the interactions with HIV-1 *gag*-, *nef*- and *gag-pol*-interacting proteins than expected. HIV-1 *env*-interacting proteins showed a preference to interact with *nef*, which is also involved in the early stage of HIV-1 infection [210].

Statistical results indicated an overlap of human host proteins targeted by *env*, *tat* and *nef*. ‘Forward’ VIPs targeted by HIV-1 *gag-pol*, *vif*, and *rev* were less likely to interact with other HIV-1 proteins. An estimated 61% of ‘forward’ VIPs targeted by *vpu* were also influenced by *nef*. After checking data with more detailed directionality information (**S3.1 Data**), we found that ‘forward’ or ‘backward’ VIPs were generally associated with fewer interactions, while that of ‘bidirectional’ VIPs tended to be associated with higher numbers of interactions (**Fig 3.6**).

Compared with other ‘forward’ VIPs, *nef*- or *vpu*-interacting ‘forward’ VIPs tended to be targeted by more HIV-1 proteins (Mann–Whitney U test: P=3.0E-35). Meanwhile, *vpu*-interacting ‘backward’ VIPs were more frequently targeted by multiple HIV-1 proteins than other ‘backward’ VIPs (P=4.2E-05). Collectively, it is common to observe human host proteins interacting with multiple HIV-1 proteins [241].

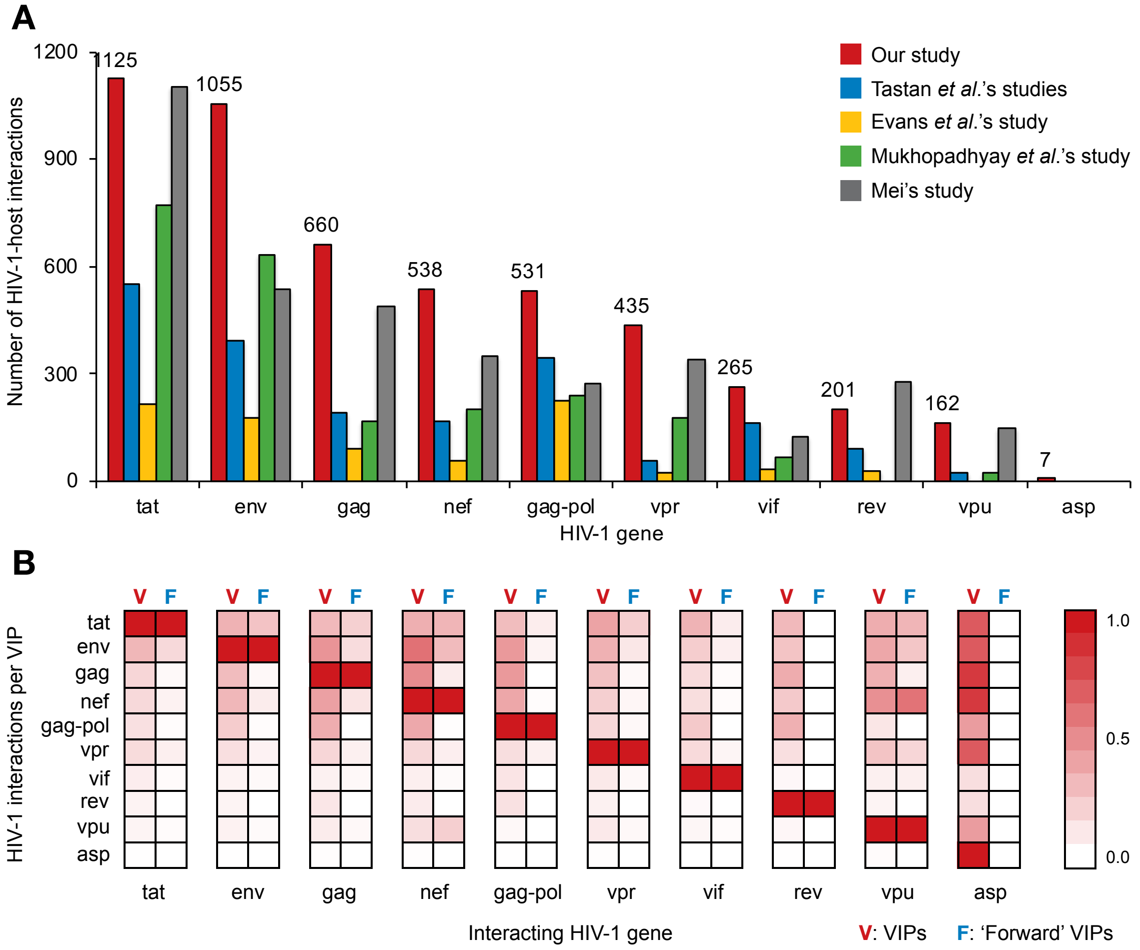


Fig 3.5 Comparison of HIV-1-host interaction datasets. (A) Comparison of data used in previous studies [173,174,224,225,228]. (B) illustration of the preference of co-occurring HIV-1-host interactions for VIPs (**V** in red) and 'forward' VIPs (**F** in blue). Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein.

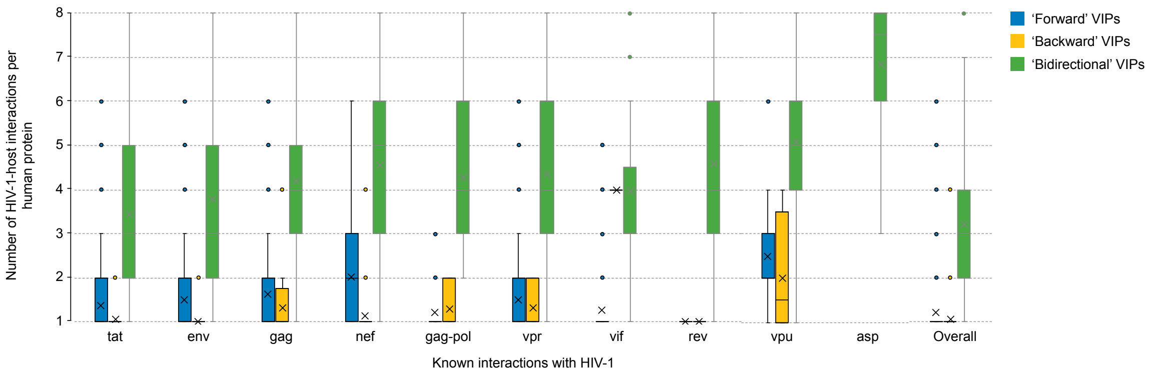


Fig 3.6 The preference of co-occurring HIV-1-host interactions for different VIPs. Boxes in the plot represented the major distribution of values (from the first to the third quartile); outliers were added for values higher than two-fold of the third quartile; cross symbol marked the position of the average value including the outliers; upper and lower

whiskers showed the maximum and minimum values excluding the outliers. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein.

3.4.2 Feature expression in the collected data

In this study, we obtained 2881 nonredundant VIPs from 16215 HIV-1-host PPI records and 7261 high-quality non-VIPs from the human proteome. Based on the outcome of HIV-1-host PPIs [287], only 1530 (~53%) of VIPs showed clear direction: ‘forward’, ‘backward’, or ‘bidirectional’ (**Fig 3.2**). In total 671 features were collected from the genetic sequence, proteomic sequence, annotation, and interaction profile data. To investigate the predictive signals in these features, we analysed their properties in the different VIP and non-VIP datasets.

3.4.2.1 Characterisation of features linked to ORFs, duplication rate and evolution conservation

The alternative splicing allows individual genes to generate multiple messenger RNA (mRNA). It has been widely discovered in an estimated 95% of multi-exon human genes [288]. This posttranscriptional process contributes to transcript variation and can produce protein isoforms with related or distinct functions, thus it has been regarded as an important driver of the evolution of phenotypic complexity in human genes [255,256]. We found that approximately 88% of our collected human genes had more than one transcript but only 76% of them formed multiple ORFs after undergoing alternative splicing. Non-VIP genes were significantly enriched with small numbers of transcripts (<6, Pearson's Chi-squared test: P=9.1E-95) or protein-coding transcripts (<4, Pearson's Chi-squared test: P=1.2E-98) while VIP genes tended to have a large number of transcripts or protein-coding transcripts (**Fig 3.7**). We found approximately 30% of non-VIP genes were non-polymorphic, but for VIPs, the ratio reduced to 12%. This provided a strong signal of inhibition for HIV-1 infection in the proteins of interest (Pearson’s Chi-squared test: P=2.8E-71). On the other hand, human proteins coding from genes with more ORFs were more likely to interact with HIV-1 (>=4, Mann–Whitney U test: P=1.1E-27) and the interacting direction tended to be ‘backward’ or ‘bidirectional’ rather than ‘forward’ if the number of ORFs reached a very high level (Mann–Whitney U test: P=0.056, 0.007, respectively).

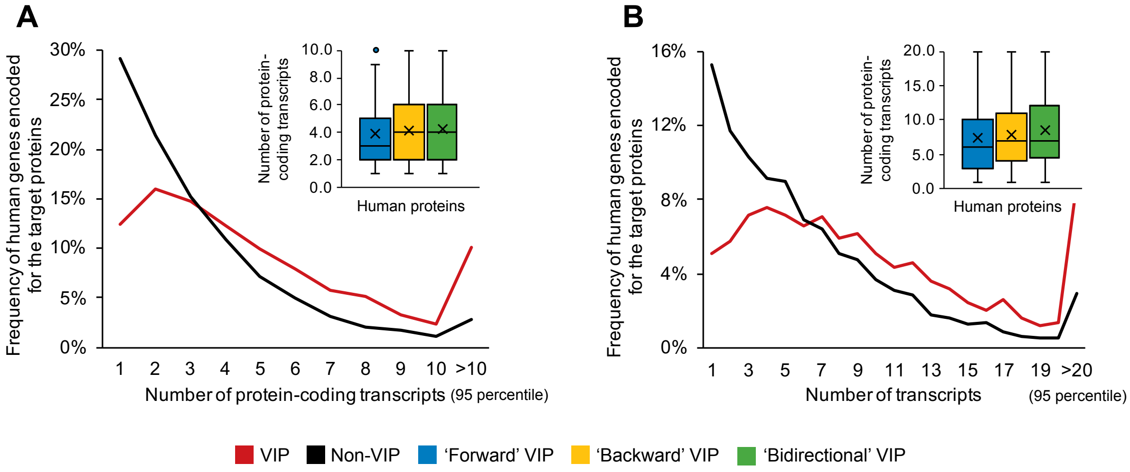


Fig 3.7 Representation of features about alternative splicing. Insets in panels A and B represented the major distribution of expression values (from the first to the third quartile); outliers were defined by expression values higher than two-fold of the third quartile; cross symbol marked the position of average expression value including outliers; upper and lower whiskers showed the maximum and minimum expression values excluding outliers. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; non-VIP, HIV-1 non-interacting human protein.

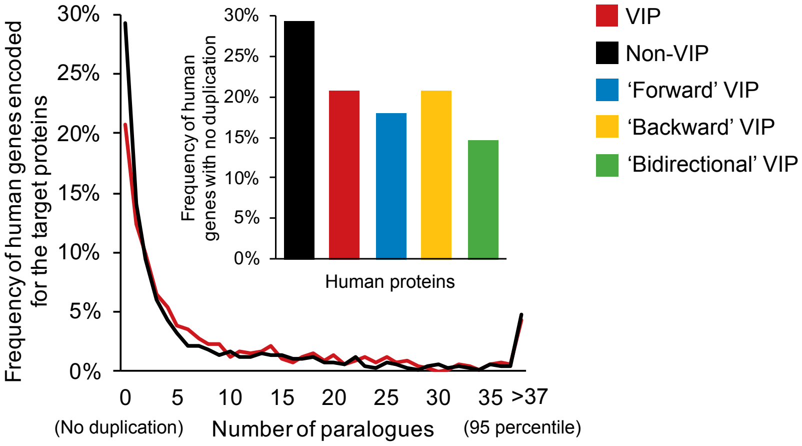


Fig 3.8 A breakdown of paralogues for different human proteins. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; non-VIP, HIV-1 non-interacting human protein.

The duplication events of genes accumulate loss-of-function mutations (degeneration) within paralogues and hence promote the subfunctionalization of the ancestral genes [289,290]. The paralogue data within human species indicated that proteins encoded by singleton genes were less likely to interact with HIV-1 (Pearson's Chi-squared test: P=2.1E-16) (**Fig 3.8**). Furthermore, human proteins produced by genes with more duplications had a higher chance to be targeted by HIV-1 (Mann–Whitney U test: P=9.7E-

4). These results suggested an association between less-degenerative mutations and HIV-1 infections. The orthologue data (**Fig 3.9**) indicated that HIV-1 molecules were more likely to interact with human proteins encoded from conserved genes (measured by dN/dS ratios, Mann–Whitney U test: P=9.5E-7, 2.8E-10, 4.9E=6, 2.9E-6, respectively). Comparing with 'forward' VIP genes, 'backward' VIPs were generally more conserved, which showed significant differences in the dN/dS ratio within human-gorilla, human-orangutan or human-gibbon orthologues (Mann–Whitney U test: P=0.038, 0.004, 0.029, respectively).

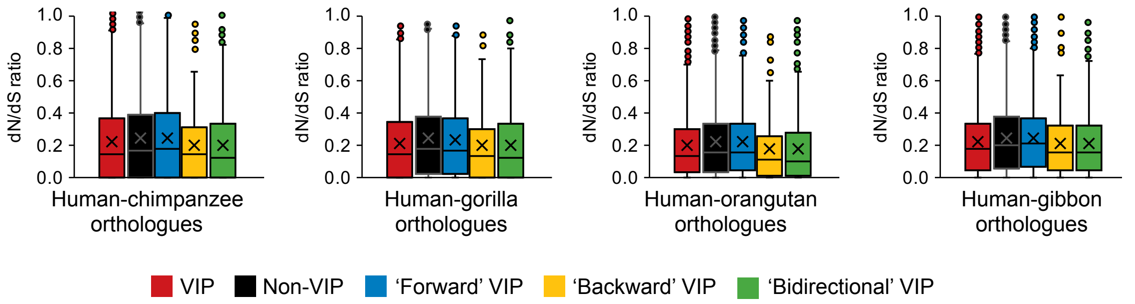


Fig 3.9 Representation of dN/dS ratio among human and four homininae genomes. Insets here represented the major distribution of expression values (from the first to the third quartile); outliers were defined by expression values higher than two-fold of the third quartile; cross symbol marked the position of average expression value including outliers; upper and lower whiskers showed the maximum and minimum expression values excluding outliers. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; non-VIP, HIV-1 non-interacting human protein; dN, non-synonymous substitutions; dS, synonymous substitutions.

In a nutshell, some evolution-related properties of human proteins influenced their interrelation with HIV-1. Large number of ORFs, high duplication rates and evolution conservation were found to have a positive influence on the human proteins, promoting HIV-1-host PPIs. Extremes of these features, huge number of ORFs and singleton cases, for instance, provide clues for the inhibition or antiviral capabilities of genes in the presence of HIV-1.

3.4.2.2 Characterisation of features in different sequence patterns

From the collected data, we found obvious enrichment of adenine in VIP genes (Mann–Whitney U test: P=1.9E-20) (**Fig 3.10**). The majority of adenine-related nucleobase groups are enriched in coding sequences (CDS) of VIP genes (**Fig 3.10**, **Fig 3.11A**). Alternatively, cytosine tended to be depleted in VIP genes (Mann–Whitney U test: P=1.3E-10) (**Fig 3.10**) thus three cytosine-starting nucleobases groups, i.e., CpT, CpC and CpG, all showed significant depletion in the CDS of VIP genes (**Fig 3.10**, **Fig 3.11A**). Although thymine was slightly depleted in the CDS of human genes (**Fig 3.10**), it still made an important contribution in classifying VIPs and non-VIPs from the perspective of codon usage (**Fig 3.11C**). Among the 28 VIP-preferred codons, 20 codons contained at least one thymine.

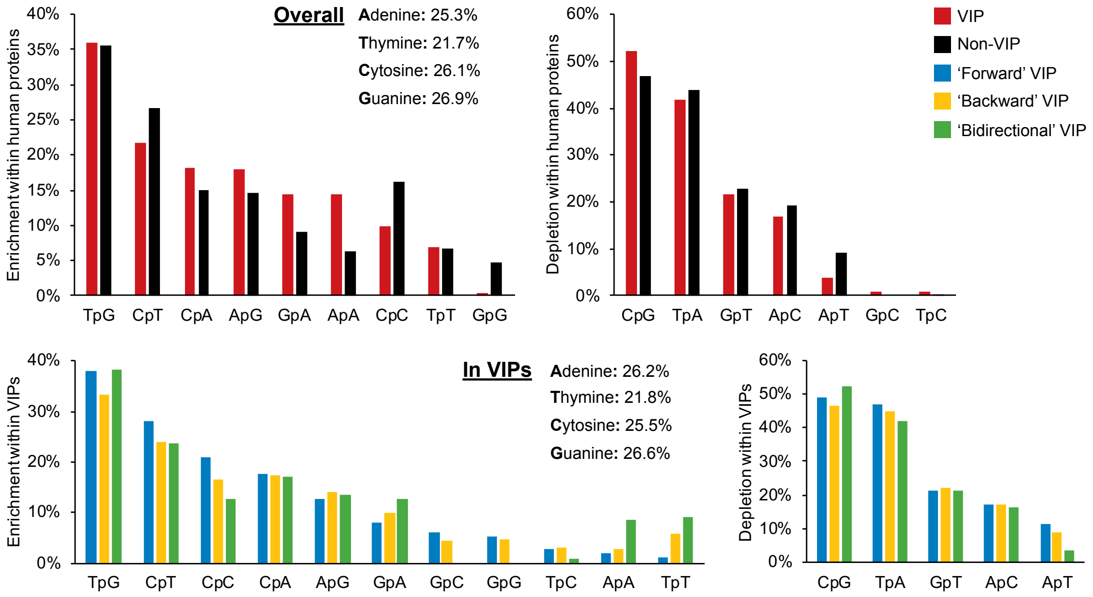


Fig 3.10 Enrichment and depletion of nucleotides linked by phosphodiester bonds in the group of human proteins or VIPs. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, HIV-1 non-interacting human proteins.

Enrichment of adenine, depletion of cytosine, and differential codon usage preferences of VIP genes all influenced the distribution of amino acids in the protein sequence [291], which also contributed to the signal distinguishing VIPs from non-VIPs. As shown in **Fig 3.11B**, we found acidic or negatively charged amino acid: aspartic acid (D) and glutamic acid (E), amide amino acid: asparagine (N) and glutamine (Q) were all significantly enriched in VIPs. Hydrophilicity, polarity, or even the size of amino acids are presumably good features to identify VIPs (**S3.3 Data**). Differences between 'backward' and 'forward' VIPs were generally not obvious from the perspective of nucleotide compositions, codon usages, or amino acid compositions. However, differences between 'bidirectional' and 'forward' VIPs were notable in 56% of nucleotide composition features, 41% of codon usage features, and 51% of amino acid composition features.

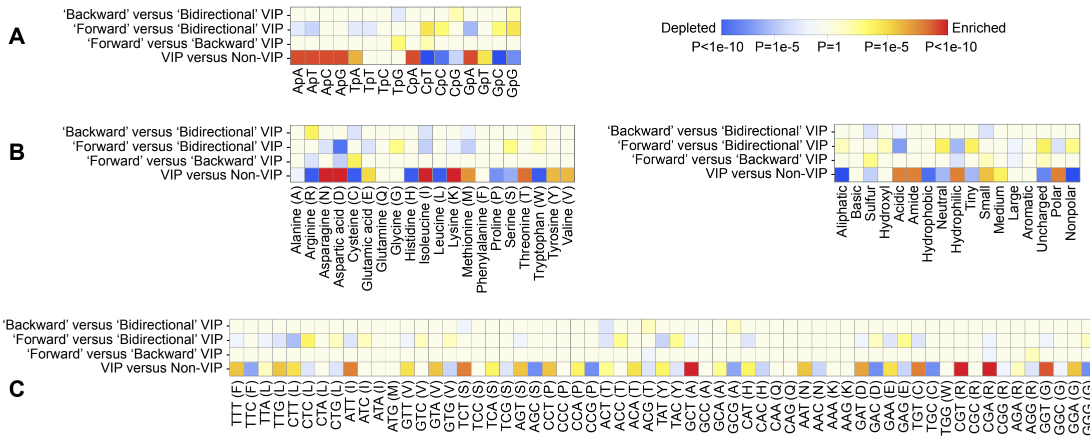
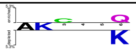
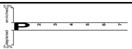
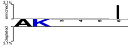
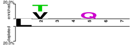
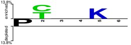
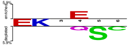
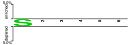
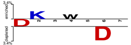
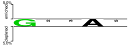
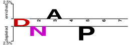
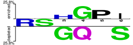
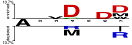
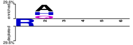
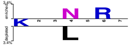
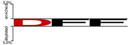
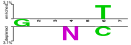
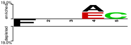
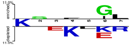
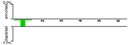
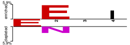
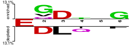
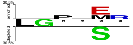
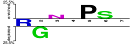
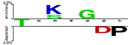
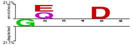
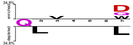
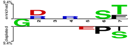
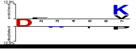
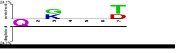


Fig 3.11 The difference of (A) nucleobase groups linked with phosphodiester bonds, (B) amino acid compositions and (C) codon usages in different classes. Detailed data about the heat maps were provided in **S3.3 Data**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; non-VIPs, HIV-1 non-interacting human protein.

We obtained some enriched sequence patterns by using the the MERCI program [261]. We found that some sequence patterns were highly similar to each other, which might result from small deletions, insertions or mutations [261]. Therefore, we integrated these sequence patterns and obtained 206 SLiMs enriched in VIPs and 'backward' VIPs. Some top-ranked SLiMs are listed in **Table 3.3**. The dash symbol in a SLiM represented a position occupied by no or one random amino acid. The results showed that there were many specific SLiMs presenting in VIPs, but they were scarce in non-VIPs. Some elements in the SLiM were conserved and less influenced by random amino acid changes, which might be related to a contributing signal from their evolutionary history [292]. Alanine (A) was found at high frequency (74%) in VIP-SLiMs, followed by lysine (K) and glutamic acid (E) observed in 60% and 55% of VIP-enriched SLiMs. Both K and E were significantly enriched in VIPs (**Fig 3.11B**), but aspartic acid (D) and isoleucine (I) seemed to be irrelevant to the conserved region even if they were highly enriched in VIP sequences (**S3.3 Data**). Differences in SLiMs were also observed between 'backward' and 'forward' VIPs. Amino acid A and leucine (L) seemed to be important to 'backward' VIPs as they were found in 51% and 55% of the enriched SLiMs (**S3.3 Data**). Additionally, we found notable differences in the overall abundance of SLiMs between the compared classes. We found 54 VIP-enriched SLiMs in the sequence of a VIP, namely plectin (PLEC) but the highest co-occurrence frequency of these SLiMs only reached 42 within the group of non-VIPs (found in the spen family transcriptional repressor, SPEN). Around 90% of VIPs contained at least one VIP-enriched

SLiM versus 82% of non-VIPs. The difference of cooccurrence status was also observed in ‘backward’ VIP-enriched SLiMs. The cumulative frequency of ‘backward’ VIP-enriched SLiMs was 97.9% in ‘backward’ VIPs and reduced to 92.5% in ‘forward’ VIPs (**Fig 3.12**).

Table 3.3 Top 20 enriched SLiMs in VIPs and ‘backward’ VIPs.

SLiM	VIP/ non-VIP	P-value[a]	Expression[b]	SLiM	‘Backward’/ ‘Forward’ VIP	P-value	Expression
AK-K-E	200/257	9.4E-14		P-E-R-V	25/37	4.7E-08	
AK-A-E	201/266	7.0E-13		L-D-T-R	27/50	1.5E-06	
E-AK-K	220/310	6.4E-12		L-R-I-G	20/33	6.8E-06	
EKE-K	221/315	1.3E-11		P-D-SS	25/50	1.5E-05	
EK-A-K	238/351	2.8E-11		D-K-Q-E	19/32	1.6E-05	
AA-K-K	184/249	3.1E-11		S-L-IS	22/41	1.7E-05	
D-Q-L-K	217/312	3.9E-11		G-SAA	21/39	2.6E-05	
D-L-K-D	241/359	4.5E-11		RS-G-S	21/39	2.6E-05	
A-D-D-E	204/288	4.6E-11		R-RS-L	23/46	3.5E-05	
K-K-E-P	240/359	6.9E-11		DFF	20/37	3.9E-05	
G-K-K-V	236/352	8.1E-11		F-H-M	20/37	3.9E-05	
K-G-K-G	239/358	8.4E-11		T-SL-T	21/41	5.6E-05	
E-MN	238/357	1.0E-10		KV-A-E	20/38	5.8E-05	
E-DL-K	240/363	1.6E-10		LG-I-S	21/42	8.1E-05	
A-K-V-K	217/320	2.3E-10		R-S-G-P	21/42	8.1E-05	
T-E-E-T	235/356	2.8E-10		G-LS-K	19/36	8.7E-05	
GD-M	235/357	3.5E-10		Q-K-P-L	22/46	1.1E-04	
G-K-K-G	229/346	4.1E-10		L-Y-L-E	23/50	1.3E-04	
G-G-T-T	236/360	4.3E-10		N-L-R-S	22/47	1.5E-04	
D-E-V-K	216/321	4.4E-10		Q-S-S-K	21/44	1.6E-04	

[a]*expression differences of SLiMs were assessed through Pearson's Chi-squared tests on different classes;* [b]*expression of amino acids in the sequence segments containing a target SLiM; for VIP-enriched SLiMs, the positive and negative segment sets were extracted from VIPs and non-VIPs, respectively; for ‘backward’-enriched SLiMs, the positive and negative segment sets were extracted from ‘backward’ and ‘forward’ VIPs, respectively.* Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; non-VIP, HIV-1 non-interacting human protein; SLiM, protein short linear motif.

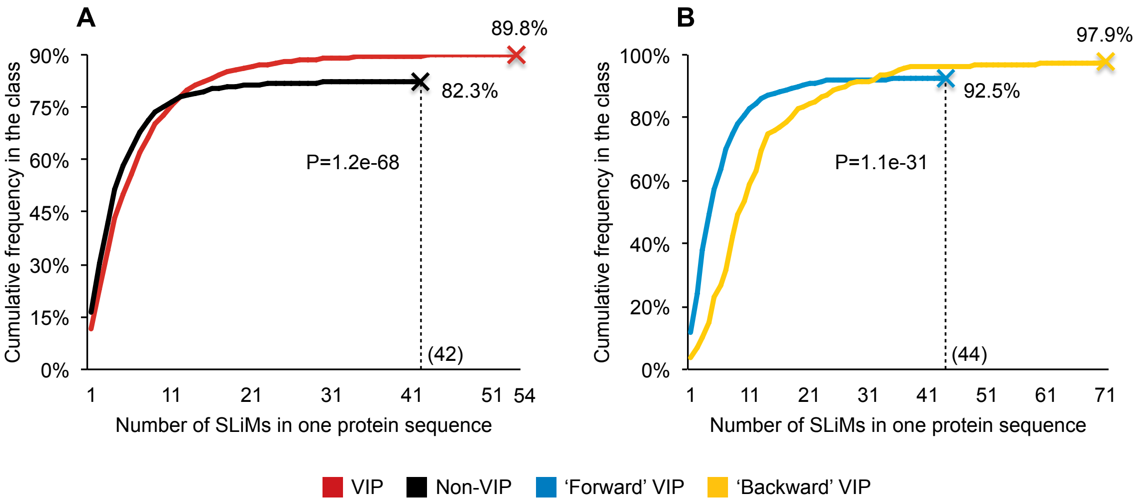


Fig 3.12 Expression of (A) 85 VIP-enriched SLiMs in human proteins and (B) 121 'backward' VIP-enriched SLiMs in 'forward' and 'backward' VIPs. Difference between these expressions are evaluated with Mann–Whitney U tests. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, HIV-1 non-interacting human proteins; SLiMs, protein short linear motifs.

Intrinsically disordered regions have a broad occurrence in proteins, allowing the same polypeptide to be involved in different PPIs [293]. Characterised by their biased amino acid composition and low sequence complexity, intrinsically disordered proteins lack the ability to fold spontaneously into stable secondary and well-packed tertiary structures. However, they still play an important role in many biological activities. Based on the result given by the IUPred [267], 92% of VIPs contained at least one disorder region while 89% of non-VIPs were disordered (Pearson's Chi-squared test: P=5.3E-5). Distributions of disorder regions were not distinguishable when comparing VIPs with non-VIPs (**Fig 3.13**) but were slightly different in VIPs with distinct directionality (**Fig 3.14**). We found that 'backward' VIPs were less likely to form disorder regions close to the beginning or end of their sequences. Disorder regions were less frequent in the middle of 'bidirectional' VIP sequences and showed great depletion at the end of the VIP sequence (**Fig 3.14A**). The representation of some amino acids, e.g., serine (S), threonine (T) (**Fig 3.14B**), E, and K (**Fig 3.14D**), were biased by the directionality of HIV-1-host molecular interactions. We assumed the results of Espritz [266] might be more useful since they could link the information of VIP-enriched SLiMs and disorder expression in the 'backward' VIPs. Amino acids K and E are important to the pattern of VIP-enriched SLiMs (**S3.3 Data**), and they had a higher chance to be found in disordered regions in the sequence of 'backward' VIPs.

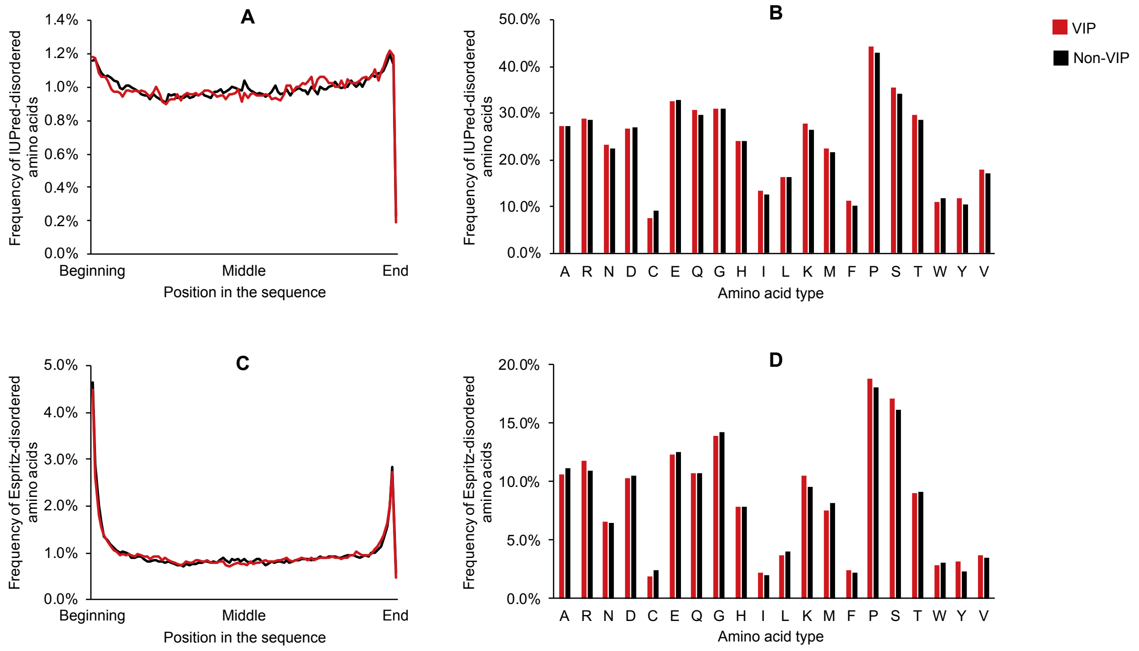


Fig 3.13 Disorder status in different human proteins. (A) and (B) showed the results calculated by IUPred [267] while (C) and (D) showed the results calculated by the Espritz [266]. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, HIV-1 non-interacting human proteins.

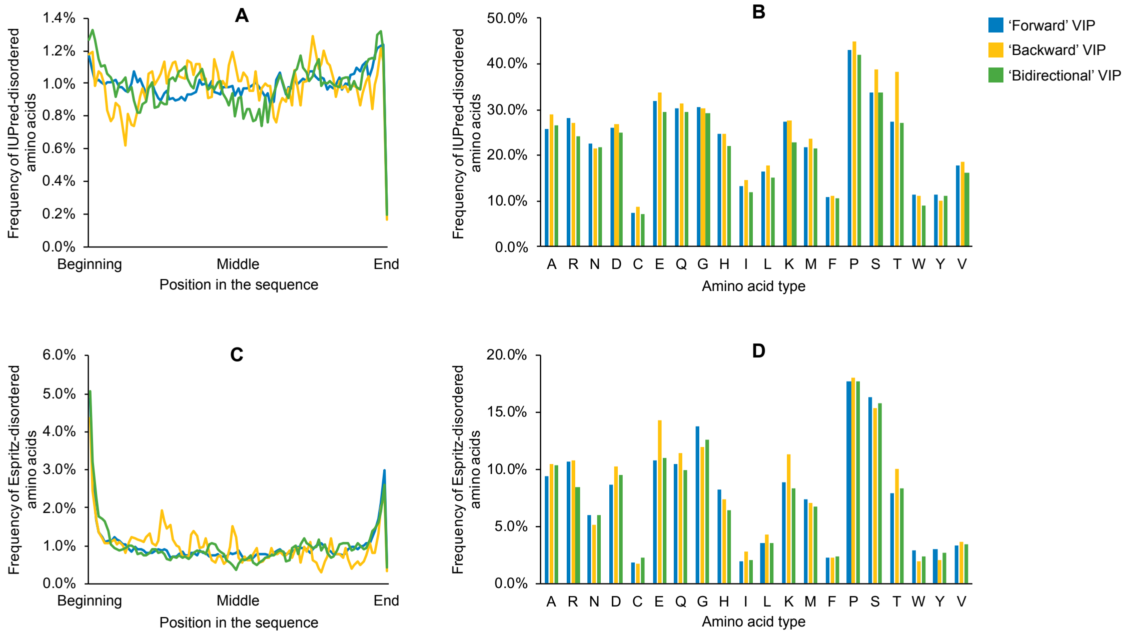


Fig 3.14 Disorder status in different VIPs. (A) and (B) showed the results calculated by IUPred [267] while (C) and (D) showed the results calculated by Espritz [266]. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins.

Briefly, VIPs and non-VIPs showed some significant differences in their sequence patterns from the nucleobase composition to SLiMs. It gave an acceptable answer to explain the reason why some human host proteins could interact with multiple HIV-1 proteins. Meanwhile, pro-viral and pro-host signs of VIPs were also reflected by special sequence patterns and intrinsic disorder status in the protein sequence.

3.4.2.3 Characterisation of features in annotation and network profiles

The Gene Ontology is a resource of knowledge unifying the representation of gene and gene product attributes. It is reported to be one of the strongest indicators of interacting proteins [294]. Compared with the annotation profile of non-VIPs, we found some GO terms, e.g., cytokine-mediated signalling pathway (GO:0019221), protein binding (GO:0005515), and nucleoplasm (GO:0005654) were highly enriched in the GO profile of VIPs. In order to reduce the dependency of our knowledge models on the annotated GO profile, we mapped the collected GO terms through the derivation tree and catalogued them into 66 domains representing child terms of biological process, molecular function and cellular component (**S3.3 Data**). Based on these 'new' GO profiles, we found an estimated 90% of VIPs were involved in the cellular process while the ratio reduced to two-thirds in non-VIPs (Pearson's Chi-squared test: P=1.9E-123) (**Fig 3.15**). The difference of binding activities was also observed between VIPs and non-VIPs (P=1.9E-84) (**Fig 3.16**). This is not surprising since the majority of VIPs were placed in key positions with a high degree or betweenness centrality within the human interactome (**S3.3 Data**). Additionally, we also found some clues in the catalogued GO profiles, which might help for the classification of VIPs with different directionality. For instance, approximately 77% of 'bidirectional' VIPs could raise a response to stimulus but the percentage for 'forward' VIPs, 'backward' VIPs, and non-VIPs only reached 47%, 29% and 20%, respectively (**Fig 3.15**, **S3.3 Data**). 'Backward' or 'bidirectional' VIPs were more likely to be found in organelles as opposed to 'forward' VIPs and non-VIPs (P=7.3E-5, 1.6E-7, 2.0E-70, respectively) (**Fig 3.17**, **S3.3 Data**).

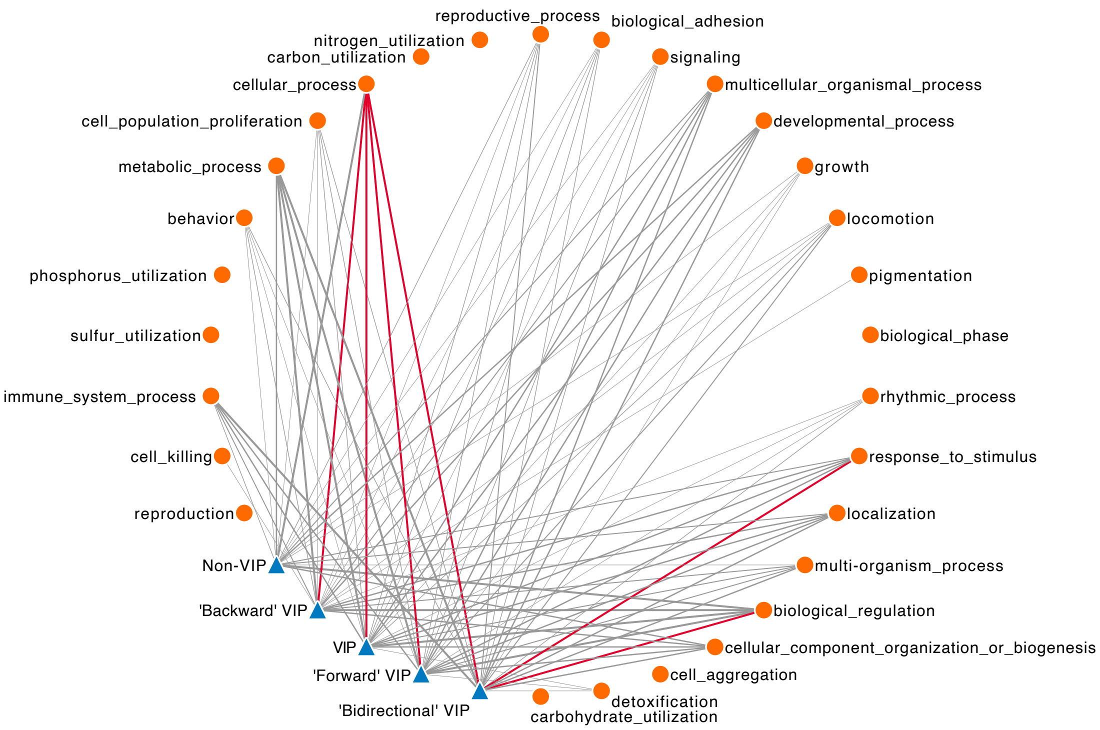


Fig 3.15 GO annotation status of different VIPs under the catalogue of biological process. Bolder lines indicate stronger relationship between human proteins and GO child terms. Red lines indicate a GO child term is related to more than 75% of human proteins. Detailed data about the GO involvement in different human proteins were provided in **S3.3 Data**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; GO, gene ontology.

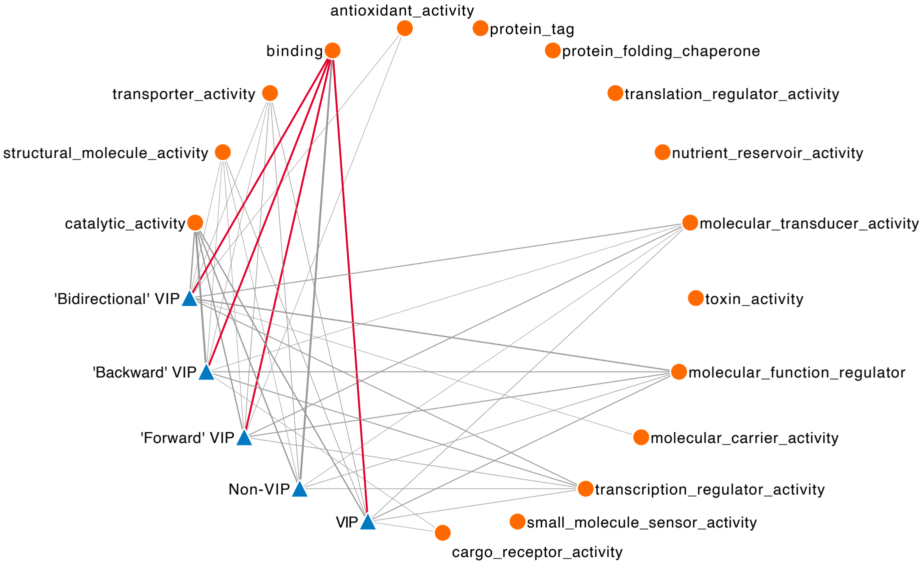

antioxidant_activity
protein_tag
binding
protein_folding_chaperone
transporter_activity
translation_regulator_activity
structural_molecule_activity
nutrient_reservoir_activity
catalytic_activity
molecular_transducer_activity
'Bidirectional' VIP
toxin_activity
'Backward' VIP
molecular_function_regulator
'Forward' VIP
molecular_carrier_activity
Non-VIP
transcription_regulator_activity
VIP
small_molecule_sensor_activity
cargo_receptor_activity

Fig 3.16 GO annotation status of different VIPs under the catalogue of molecular function. Bolder lines indicate stronger relationship between human proteins and GO child terms. Red lines indicate a GO child term is related to more than 75% of human proteins.

Detailed data about the GO involvement in different human proteins were provided in **S3.3 Data**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; GO, gene ontology.

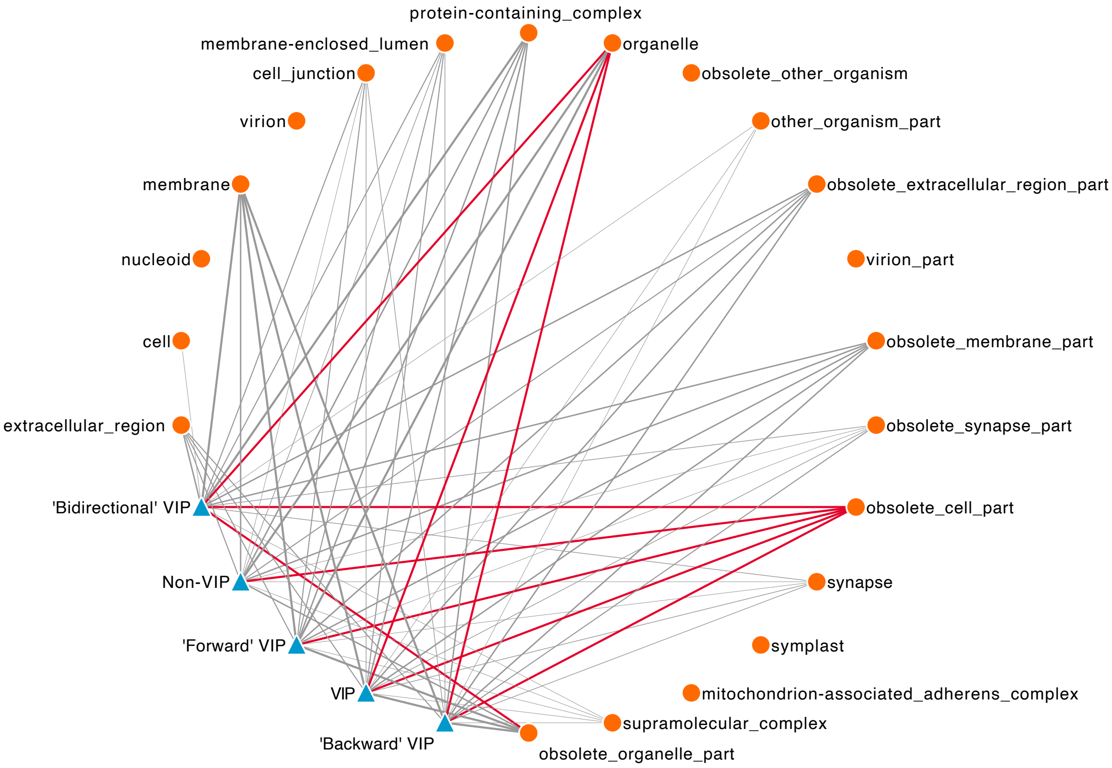


Fig 3.17 GO annotation status of different VIPs under the catalogue of cellular component. Bolder lines indicate stronger relationship between human proteins and GO child terms. Red lines indicate a GO child term is related to more than 80% of human proteins. Detailed data about the GO involvement in different human proteins were provided in **S3.3 Data**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; GO, gene ontology.

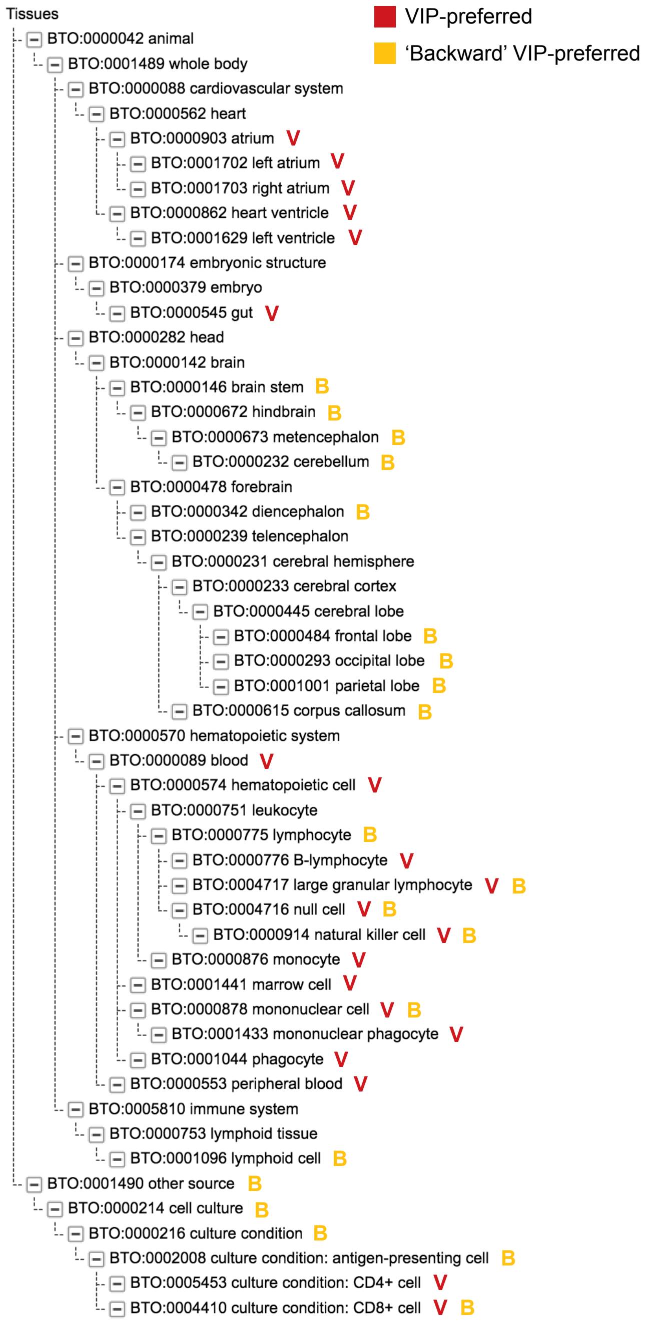

🟥 VIP-preferred
🟨 'Backward' VIP-preferred
Tissues
BTO:0000042 animal
BTO:0001489 whole body
BTO:0000088 cardiovascular system
BTO:0000562 heart
BTO:0000903 atrium V
BTO:0001702 left atrium V
BTO:0001703 right atrium V
BTO:0000862 heart ventricle V
BTO:0001629 left ventricle V
BTO:0000174 embryonic structure
BTO:0000379 embryo
BTO:0000545 gut V
BTO:0000282 head
BTO:0000142 brain
BTO:0000146 brain stem B
BTO:0000672 hindbrain B
BTO:0000673 metencephalon B
BTO:0000232 cerebellum B
BTO:0000478 forebrain
BTO:0000342 diencephalon B
BTO:0000239 telencephalon
BTO:0000231 cerebral hemisphere
BTO:0000233 cerebral cortex
BTO:0000445 cerebral lobe
BTO:0000484 frontal lobe B
BTO:0000293 occipital lobe B
BTO:0001001 parietal lobe B
BTO:0000615 corpus callosum B
BTO:0000570 hematopoietic system
BTO:0000089 blood V
BTO:0000574 hematopoietic cell V
BTO:0000751 leukocyte
BTO:0000775 lymphocyte B
BTO:0000776 B-lymphocyte V
BTO:0004717 large granular lymphocyte V B
BTO:0004716 null cell V B
BTO:0000914 natural killer cell V B
BTO:0000876 monocyte V
BTO:0001441 marrow cell V
BTO:0000878 mononuclear cell V B
BTO:0001433 mononuclear phagocyte V
BTO:0001044 phagocyte V
BTO:0000553 peripheral blood V
BTO:0005810 immune system
BTO:0000753 lymphoid tissue
BTO:0001096 lymphoid cell B
BTO:0001490 other source B
BTO:0000214 cell culture B
BTO:0000216 culture condition B
BTO:0002008 culture condition: antigen-presenting cell B
BTO:0005453 culture condition: CD4+ cell V
BTO:0004410 culture condition: CD8+ cell V B

Fig 3.18 Top 20 tissues preferred by VIPs and 'backward' VIPs. Detailed data about the top 20 tissue tropisms were provided in **S3.3 Data**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein.

From the perspective of tissue tropisms, VIPs preferred heart- or hematopoietic system-related tissue. The most significant difference between VIP and non-VIP was found in their association with monocyte (nongranular phagocytic leukocyte) (**Fig 3.18**, **S3.3 Data**) (P=1.7E-137) [295]. The same differences were found in phagocyte that engulfs foreign material [296] and immature blood cells that develop in the bone marrow [297]. In the tissue group of antigen-presenting cells, CD4+ and CD8+ cells were favoured by VIPs, which showed a strong relationship between a virus invading and the immune responses (P=5.1E-131 and 5.8E-114, respectively) [298,299]. Compared with ‘forward’ VIPs, ‘backward’ VIPs were less involved in the hematopoietic system, but they were more expressed in brain-related tissues, such as the brain stem and cerebral lobe (**S3.3 Data**). Cells originating from stem cells and differentiating in lymphoid tissues were favoured by ‘backward’ VIPs. The relationship between ‘backward’ VIPs and CD8+ cells were even more obvious, showing a clear relationship linking the virus invading and the host antiviral immune responses [300].

In summary, annotation profiles represented by the GO terms and tissue tropisms differentiated the biological environment between VIPs and non-VIPs. Some preferences of VIPs such as more involvement in cellular binding activities were also reflected in the human interactome.

3.4.3 Performance of different feature sets in the training stage

3.4.3.1 Models for predicting VIPs

In this study, we encoded 537 features for the prediction of VIPs. According to the data source from which they were extracted, we divided these features into four categories: genome-based features, proteome-based features, annotation-based features and interaction profile-based features. We first tested the performance of features in different categories on the balanced training datasets and found that annotation-based features performed the best, achieving the highest AUC value at 0.8090 on dataset S1’ (**Table 3.4**). On the same dataset, the combination of interaction profile-based features produced some good predictions even if only 10 features were included. However, the performance of proteome-based features was poor on dataset S1’. By combining all encoded 537 features, we found the classifier could further produce a better performance (AUC=0.8324) than using features by categories (AUC=0.7118, 0.6641, 0.8090, 0.7487, respectively) (**Table 3.4**). We compared SVM with another three machine learning algorithms including KNN, DT and RF [170]. We used the square root of the size of the training samples as the k-value for the KNN algorithm [301]. We found this algorithm was biased to the positive class and did not achieve a better prediction performance than the SVM. The DT algorithm was designed with a feature selection scheme, which helped it better split the dataset for lower system entropy [196]. It used 278 out of 537 features and made the worst performance among the machine learning algorithms compared. We initialised the RF algorithm with 50 trees and repeated the modelling process ten times to balance its random processes on bootstrapping the dataset and selection of features [302]. The prediction performance of the RF algorithm on S1' was promising but did not surpass that of the SVM. These results suggested that the majority of features encoded for the prediction of VIPs were contributing to the signal, but the complete feature set was not optimal for reliable prediction since it included some poorly performing features.

Table 3.4 The performance of different feature sets on the training datasets over five-cross validations.

Dataset[a]	Algorithm	Features	Features number	Threshold[e]	Sensitivity	Specificity	Accuracy	MCC	AUC
S1'	SVM	Genetic sequences	107	0.51	0.613	0.700	0.656	0.314	0.7118
	SVM	Proteomic sequences	128	0.51	0.595	0.649	0.622	0.244	0.6641
	SVM	Annotations	292	0.57	0.663	0.806	0.735	0.475	0.8090
	SVM	Interaction profiles	10	0.52	0.611	0.777	0.694	0.394	0.7487
	SVM	Combination	537	0.56	0.690	0.817	0.754	0.512	0.8324
	KNN[b]	Combination	537	0.35~0.39	0.766	0.633	0.699	0.402	0.7772
	DT[c]	Partial	278	N/A	0.633	0.642	0.637	0.275	N/A
	RF[d]	Random	Random	0.44~0.52	0.733±0.035	0.752±0.030	0.742±0.004	0.486±0.009	0.8157±0.0031
	SVM	Top-ranked 33	33	0.54	0.645	0.718	0.681	0.363	0.7468
	SVM	Top-ranked 193	193	0.48	0.748	0.751	0.750	0.499	0.8261
	KNN[b]	Optimum	441	0.43~0.48	0.689	0.720	0.705	0.410	0.7734
	SVM	Optimum	441	0.52	0.727	0.787	0.757	0.514	0.8344
S2'	SVM	Genetic sequences	107	N/A[f]	N/A	N/A	N/A	N/A	N/A
	SVM	Proteomic sequences	164	0.40	0.860	0.633	0.747	0.507	0.8023
	SVM	Annotations	292	0.46	0.767	0.520	0.643	0.296	0.6786
	SVM	Interaction profiles	21	0.51	0.740	0.633	0.687	0.375	0.7108
	SVM	Combination	584	0.46	0.807	0.553	0.680	0.372	0.7383
	KNN[b]	Combination	584	0.50~0.54	0.400	0.833	0.617	0.259	0.6501
	DT[c]	Partial	66	N/A	0.673	0.660	0.667	0.333	N/A
	RF[d]	Random	Random	0.38~0.58	0.706±0.134	0.710±0.167	0.708±0.030	0.432±0.045	0.7609±0.0270
	KNN[b]	Optimum	129	0.27~0.36	0.487	0.873	0.680	0.390	0.7509
	SVM	Optimum	129	0.44	0.853	0.680	0.767	0.542	0.8260

[a] *Dataset S1' and S2' were balanced training datasets constructed via undersampling strategy [265] from dataset S1 and S2, respectively (**Table 3.1**). Compositions of these two datasets are provided in **S3.2 Data**.* [b] *k-value here was determined as the square root of the size of the training samples in the five-fold cross validation;* [c] *the DT algorithm selected 278 and 66 features from the original feature sets for the two modelling tasks;* [d] *the RF algorithm used 50 random grown trees and the modelling and validation procedures were repeated for 10 times;* [e] *threshold was set by maximizing the value of MCC;* [f] *multiple N/A were given here as the prediction quality of the generated classifier was worse than a random guess.* Abbreviations: SVM, support vector machine; KNN, k-nearest neighbors; DT, decision tree; RF, random forest; MCC, Matthews Correlation Coefficient; AUC, area under the receiver operating characteristic curve.

In order to find a better feature subset for the prediction of VIPs, we first used the Fisher-Markov Selector [276] to calculate the importance of individual features (**Fig 3.19A**). The results demonstrated the importance of gene ontology (e.g., involvement in metabolic process, ranked the 24th) and tissue tropism features for prediction (e.g., expression in monocyte, ranked the 1st), even if they were used individually. According to the importance score of individual features, we then used our first feature selection strategy (**Fig 3.3**) and five-cross validations to optimise the prediction model. **Fig 3.19B** shows that the classifier obtained the last obvious improvement in the prediction when the top 33 features are used. With a subset of 193 features, the improvement became limited. The optimum feature subset (n = 441) was composed of 105 genome-based features, 84 proteome-based features, 243 annotation-based features and nine interaction profile-based features. In other words, after removing 96 badly performing features, the classifier became stronger and the risk of facing missing data or errors was also reduced simultaneously [303]. The distribution of prediction scores for VIPs and non-VIPs was negatively and positively skewed, with most values clustered around the right and left tails of the distribution, respectively (**Fig 3.19C**). Additionally, the classifier that used the top 193 features was also recommended for further reducing such risks and producing similar performance as the one using 441 features (**Fig 3.19C**, **Fig 3.20B**). By contrast, the model using the top 33 features was not the most promising unless only gene ontology and tissue tropism data were available for making the prediction (**Fig 3.20A**).

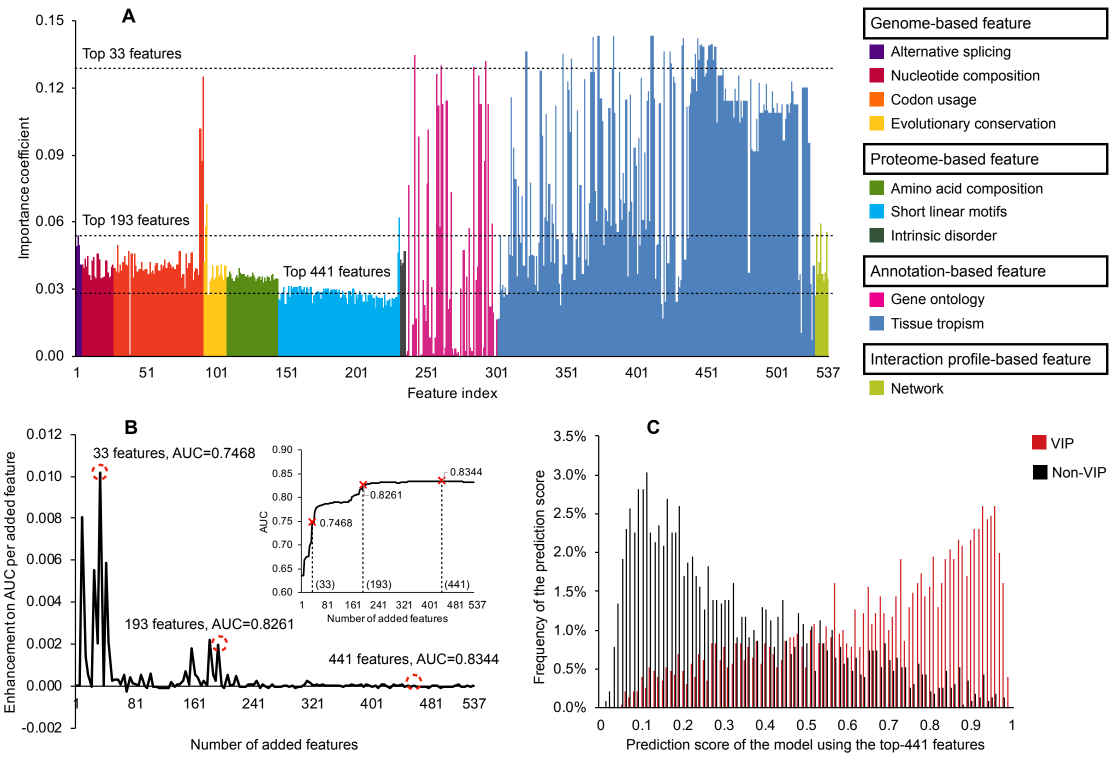


Fig 3.19 The performance of different features for the prediction of VIPs. (A) importance of different features. (B) change on AUC as the increase of ordered features. (C) the distribution of prediction scores (for VIPs and non-VIPs) generated by model using the

top 441 features. In (A) the importance of individual feature was recorded by averaging the results on the balanced training datasets generated by 10-round undersampling procedures [265] on dataset S1. The ranked list of the encoded 537 features is provided in **S3.4 Data**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, HIV-1 non-interacting human proteins; AUC, area under the receiver operating characteristic curve.

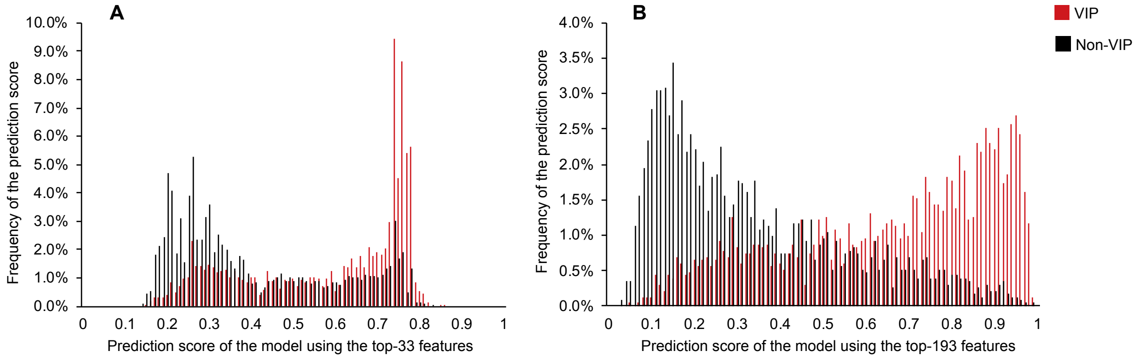


Fig 3.20 Prediction score generated by models using (A) top-33 and (B) top-193 features on dataset S1' over five-fold cross-validation. Proteins obtained higher prediction scores are more likely to be VIPs. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, HIV-1 non-interacting human proteins.

3.4.3.2 Models for predicting the directionality in HIV-1-host PPIs

We encoded 584 features for VIPs to predict their directionality in HIV-1-host molecular interactions. Predictions on the training dataset S2' were different from those on dataset S1'. The performance of genome-based features was even worse than a random prediction (**Table 3.4**). The combination of proteomic features produced a highly predictive model. The performance of annotation features was not as good as anticipated, and the combination of all features made the classifier even worse than only using proteomic features. This indicated a big difference between two prediction tasks highlighted in this study. After checking the importance of features with the Fisher-Markov Selector [276], we found the difference between the generated importance score was not obvious (**Fig 3.21, S3.4 Data**), which meant the contribution of individual features to the prediction model had not changed appreciably. The comparison results among different machine learning algorithms proved that the SVM classifier also worked on dataset S2' (**Table 3.4**). These discoveries suggested that the complementarity of these 584 features was not as good as those used for predicting VIPs. There might be a large number of noisy features involved in the complete set. Considering this, our first feature selection strategy might not work on dataset S2' (**Fig 3.22A**).

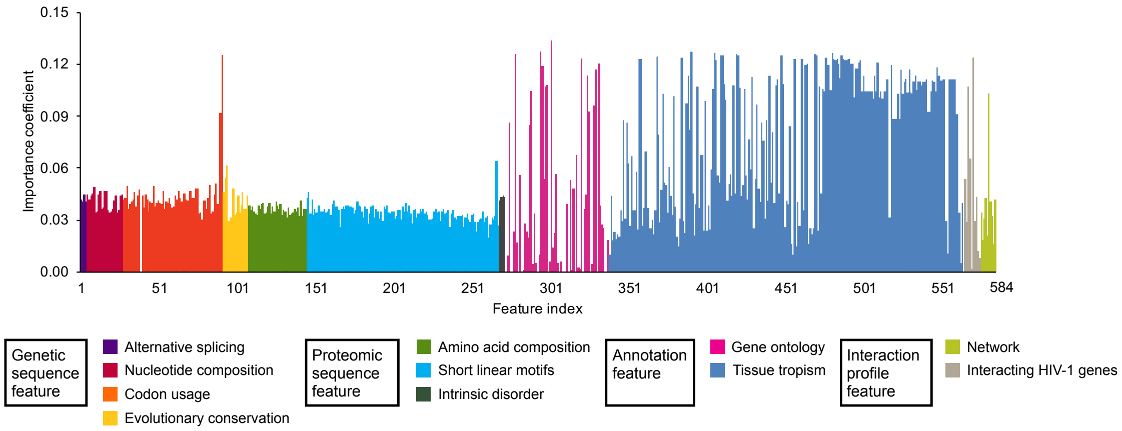


Fig 3.21 Importance of individual features for predicting 'backward' and 'forward' VIPs. The importance score of individual features is recorded by averaging the results on the balanced training datasets generated by ten-round undersampling procedures [265] on dataset S2. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins.

Instead, we applied our second feature selection strategy to optimise the prediction model (**Fig 3.4**). We assumed that the proteomic features might be an ideal set with good complementarity for the initialization as they were better-performing than features in the other catalogues (**Table 3.4**). The performance of the classifier was enhanced, however, it started to decrease after adding four non-proteomic sequence features. Next, we removed badly-performing features in the proteomic sequence feature set. We identified an optimum model generated by 129 features (**Fig 3.22B**). In that feature subset, 36 amino acid composition features, 85 SLiM features, four intrinsic disorder features, two gene ontology feature, and two tissue tropism features were included (**S3.4 Data**). Compared with the model generated from the complete feature set, the model using 129 optimum features enhanced the performance by more than 10% from the perspective of the AUC (**Table 3.4**). Likewise, on the training dataset S2', SVM was still superior to KNN, DT and RF algorithms. Additionally, the model generated from all proteomic features was also recommended as it only required the information from the protein sequence to make good predictions (**Fig 3.23**).

Interestingly, testing on the reference dataset S3 suggested that 'bidirectional' VIPs were closer to 'forward' VIPs than to 'backward' VIPs (**Fig 22D**). 'Forward' VIPs might be 'responding' to HIV-1 infection and target HIV-1, making them 'bidirectional' [304]. 'Backward' VIPs were less likely to be targeted by HIV-1 so their chance of becoming 'bidirectional' VIPs was relatively low. The recommended direction based on the prediction score generated by the model using 129 optimum features is provided in **Table 3.5**. We could confidently label 60% of the generated VIPs based on the prediction scores as ‘backward’, ‘forward’, or ‘bidirectional’. For prediction scores located in specific ranges, our confidence on the direction of HIV-1-host molecular interactions could reach as high as 89%.

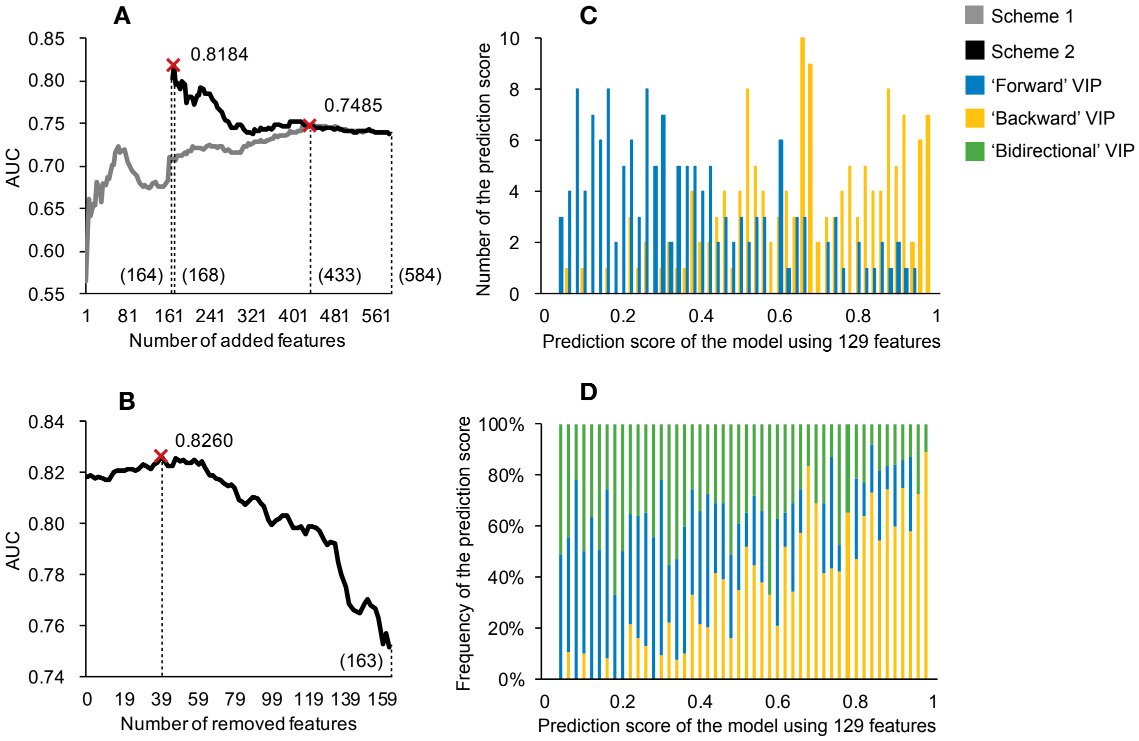


Fig 3.22 The performance of different features for predicting ‘backward’ and ‘forward’ VIPs. (A) changes on AUC as the increase of ordered features. (B) changes on AUC as the decrease of ordered proteome-based features. (C) the count of prediction scores (for pro-viral ‘forward’ VIPs and pro-host ’backward’ VIPs) generated by model using 129 optimum features. (D) the percentage of ‘forward’, ‘backward’ and ‘bidirectional’ VIPs within different regions of prediction scores (scale = 0.02) on the combined dataset of S2’ and S3. The recommended directionality of HIV-1-host molecular interactions are listed in **Table 3.5**. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIP, HIV-1 interacting human protein; AUC, area under the receiver operating characteristic curve.

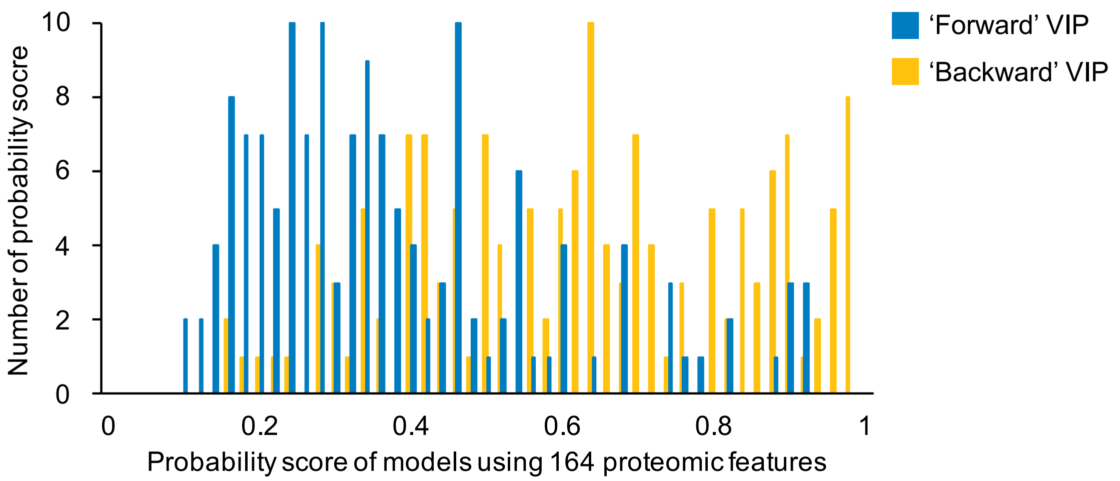


Fig 3.23 Prediction score generated by models using proteomic features on dataset S2’ over five-cross validation. Proteins obtained higher prediction scores are more likely to be ‘backward’ VIPs. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins.

Table 3.5 The recommended directionality of the prediction score for models using 129 optimum features[f].

Score range	Direction distribution ‘backward’/’forward’/ ’bidirectional’	Recommendation_1[b,c]	Confidence_1[a,d]	Recommendation_2	Confidence_2[e]
[0.00-0.02)	0.00%/0.00%/0.00%	'Forward'	N/A	'Bidirectional'	High
[0.02-0.04)	0.00%/0.00%/0.00%	'Forward'	N/A	'Bidirectional'	High
[0.04-0.06)	**0.00%/2.00%/2.09%**	**'Bidirectional'**	**51%**	**'Forward'**	**100.00%**
[0.06-0.08)	0.67%/2.67%/2.69%	'Bidirectional'	45%	'Forward'	88.93%
[0.08-0.10)	**0.00%/5.33%/1.49%**	**'Forward'**	**78%**	**'Bidirectional'**	**100.00%**
[0.10-0.12)	0.67%/2.67%/3.28%	'Bidirectional'	50%	'Forward'	89.92%
[0.12-0.14)	**0.00%/4.67%/2.69%**	**'Forward'**	**63%**	**'Bidirectional'**	**100.00%**
[0.14-0.16)	**0.00%/4.00%/3.88%**	**'Forward'**	**51%**	**'Bidirectional'**	**100.00%**
[0.16-0.18)	0.67%/5.33%/2.09%	'Forward'	66%	'Bidirectional'	91.76%
[0.18-0.20)	**0.00%/1.33%/2.69%**	**'Bidirectional'**	**67%**	**'Forward'**	**100.00%**
[0.20-0.22)	0.00%/3.33%/3.28%	'Forward'	50%	'Bidirectional'	100.00%
[0.22-0.24)	2.00%/4.00%/3.28%	'Forward'	43%	'Bidirectional'	78.46%
[0.24-0.26)	0.67%/2.00%/1.49%	'Forward'	48%	'Bidirectional'	83.97%
[0.26-0.28)	1.33%/5.33%/3.58%	'Forward'	52%	'Bidirectional'	86.99%
[0.28-0.30)	**0.00%/3.33%/2.69%**	**'Forward'**	**55%**	**'Bidirectional'**	**100.00%**
[0.30-0.32)	0.67%/4.67%/1.49%	'Forward'	68%	'Bidirectional'	90.23%
[0.32-0.34)	1.33%/1.33%/3.28%	'Bidirectional'	55%	'Forward'	77.59%
[0.34-0.36)	0.67%/3.33%/4.48%	'Bidirectional'	53%	'Forward'	92.14%
[0.36-0.38)	0.67%/3.33%/2.69%	'Forward'	50%	'Bidirectional'	90.03%
[0.38-0.40)	2.67%/3.33%/2.09%	'Forward'	41%	'Backward'	74.17%
[0.40-0.42)	1.33%/2.67%/2.09%	'Forward'	44%	'Bidirectional'	78.10%
[0.42-0.44)	1.33%/3.33%/1.79%	'Forward'	52%	'Bidirectional'	79.35%
[0.44-0.46)	2.00%/1.33%/1.49%	'Backward'	41%	'Bidirectional'	72.37%
[0.46-0.48)	2.67%/2.00%/2.09%	'Backward'	39%	'Bidirectional'	70.40%
[0.48-0.50)	0.67%/1.33%/2.09%	'Bidirectional'	51%	'Forward'	83.70%
[0.50-0.52)	2.67%/2.00%/2.99%	'Bidirectional'	39%	'Backward'	73.86%
[0.52-0.54)	5.33%/1.33%/3.58%	'Backward'	52%	'Bidirectional'	86.99%
[0.54-0.56)	3.33%/2.00%/2.09%	'Backward'	45%	'Bidirectional'	73.06%
[0.56-0.58)	2.67%/2.00%/2.39%	'Backward'	38%	'Bidirectional'	71.65%
[0.58-0.60)	**1.33%/0.00%/2.69%**	**'Bidirectional'**	**67%**	**'Backward'**	**100.00%**
[0.60-0.62)	2.00%/4.00%/3.58%	'Forward'	42%	'Bidirectional'	79.13%
[0.62-0.64)	2.67%/0.67%/1.79%	'Backward'	52%	'Bidirectional'	86.99%
[0.64-0.66)	2.00%/2.00%/1.79%	'Backward'	35%	'Forward'	69.07%

[0.66-0.68)	6.67%/2.00%/2.99%	'Backward'	57%	'Bidirectional'	82.84%
[0.68-0.70)	**6.00%/0.00%/1.19%**	**'Backward'**	**83%**	**'Bidirectional'**	**100.00%**
[0.70-0.72)	**1.33%/0.00%/0.60%**	**'Backward'**	**69%**	**'Bidirectional'**	**100.00%**
[0.72-0.74)	2.00%/1.33%/1.49%	'Backward'	41%	'Bidirectional'	72.37%
[0.74-0.76)	2.00%/2.00%/0.60%	'Backward'	44%	'Forward'	87.01%
[0.76-0.78)	2.67%/0.67%/2.99%	'Bidirectional'	47%	'Backward'	89.45%
[0.78-0.80)	**3.33%/0.00%/1.79%**	**'Backward'**	**65%**	**'Bidirectional'**	**100.00%**
[0.80-0.82)	2.00%/1.33%/0.90%	'Backward'	47%	'Forward'	78.82%
[0.82-0.84)	3.33%/0.67%/1.19%	'Backward'	64%	'Bidirectional'	87.16%
[0.84-0.86)	2.67%/0.67%/0.30%	'Backward'	73%	'Forward'	91.78%
[0.86-0.88)	2.67%/1.33%/0.90%	'Backward'	54%	'Forward'	81.71%
[0.88-0.90)	5.33%/0.67%/1.19%	'Backward'	74%	'Bidirectional'	90.73%
[0.90-0.92)	3.33%/1.33%/0.90%	'Backward'	60%	'Forward'	83.90%
[0.92-0.94)	4.67%/0.67%/0.90%	'Backward'	75%	'Bidirectional'	89.30%
[0.94-0.96)	1.33%/0.67%/0.30%	'Backward'	58%	'Forward'	87.01%
[0.96-0.98)	**4.00%/0.00%/1.49%**	**'Backward'**	**73%**	**'Bidirectional'**	**100.00%**
[0.98-1.00]	**4.67%/0.00%/0.60%**	**'Backward'**	**89%**	**'Bidirectional'**	**100.00%**

a: N/A is placed for the lack of reference proteins with predicted probability scores located within corresponding ranges; b: recommendation_1 is determined according to the dominant class scored within the corresponding range; c: in case the dominant class is not obvious, we recommend VIPs with prediction score no less than 0.5 to be 'backward' and those with prediction score less than 0.5 to be 'forward'; d: confidence_1 is calculated as the occurrence frequency of dominant class divided by that of all classes; e: confidence_2 is calculated as the occurrence frequency of the major two classes divided by that of all classes; we can't provide the exact confidence we have when getting a prediction score lower than 0.04, but the tested HIV-1 interacting human protein are very likely to have a 'forward' or 'bidirectional' tag; f: rows are coloured as red when we have confidence_1 scored higher than 50%; texts in these rows are further bold when confidence_2 reaches 100%. Abbreviations: HIV-1, human immunodeficiency virus type 1.

3.4.4 Performance on the testing datasets

In this study, we produced three models with the top-33, top-193 and top-441 features on the whole training dataset S1' for the prediction of VIPs, namely PreVIP-33, PreVIP-193 and PreVIP-441, respectively. Independent testing datasets prepared to assess the generalization capability of these three models was derived from our main dataset S1 through an undersampling strategy [265]. They consist of a random set of 577 VIPs and 4957 non-VIPs. The imbalance ratio of positives (VIPs) to negatives (non-VIPs) in this testing dataset was close to 1:8. PreVIP-33 could successfully predict 40.9% of VIPs and 88.1% of non-VIPs under a threshold of 0.73. The corresponding AUC value of PreVIP-33 was 0.7323. Under the same threshold, the sensitivity and specificity of PreVIP-193 increased to 45.6% and 91.2%, respectively. The optimum threshold for PreVIP-193 was 0.82, under which 34.7% of VIPs and more than 95% of non-VIPs were correctly predicted. Among the generated three models, PreVIP-441 achieved the best performance with an AUC valued 0.8079. It was close to the figure for PreVIP-193 and 10% higher than the figure for PreVIP-193 (**Table 3.6**). When attempting to successfully predict more than half of the VIPs, the ratios of false positives produced by PreVIP-441, PreVIP-193 and PreVIP-33 were 10%, 11% and 19%, respectively.

As for predicting the direction of HIV-1-host molecular interactions, we generated two models with the optimal 129 features and the overall 164 proteomic sequence features on the whole training dataset S2’, namely PreDIR-129 and PreDIR-164. Independent testing dataset prepared to assess the generalization capability of these two models was derived from our main dataset S2 through the undersampling strategy [265]. They consist of a random 38 VIPs and 857 non-VIPs. The imbalance ratio of positives (‘backward’ VIPs) to negatives (‘forward’ VIPs) in this testing dataset is close to 1:22. Compared with PreDIR-129, PreDIR-164 was generally a bit more powerful for achieving higher AUC, at 0.7110 (**Table 3.6**). The optimum threshold for PreDIR-129 was 0.70, under which 47.4% of ‘backward’ VIPs and 87.3% of ‘forward’ VIPs were successfully predicted. By contrast, to successfully filter the same number of negatives, PreDIR-164 produced three more false negatives, which showed its drawback in recognising ‘forward’ VIPs.

To further assess the quality of our prediction models, i.e., PreVIP-193 and PreVIP-441, we introduced two testing datasets: S5, the HIV-1 infection pathway from Reactome [150] and S6, a viral host-dependency epistasis map linked to HIV-1 infection [254]. **Fig 3.24** illustrates that our prediction models performed well in recognising VIPs with experimental evidence. On testing dataset S5, PreVIP-441 could recognise 70.1% of VIPs under a threshold of 0.73. It was about 40% stronger than its expected performance (**Table 3.6**). It was also capable of successfully predicting more than 90% of VIPs when using a threshold of 0.5. Under the same threshold, PreVIP-193 achieved a similar performance as PreVIP-441. The improvement under the threshold of 0.82 reached as high as 66% when compared with its expected sensitivity (34.7%). Thus, these results showed good generalization capabilities of our models on predicting VIPs involved in the host pathway hijacked during HIV-1 infections. Their performance on the testing dataset S6 was also promising with an estimated 20% improvement. The prediction results of PreVIP-193 and PreVIP-441 on testing datasets S5 and S6 are shown in **S3.5 Data**.

Table 3.6 The performance of features with different catalogues on the testing datasets.

Dataset	Model	Feature source	Threshold[a]	Sensitivity	Specificity	Accuracy	Precision	MCC	AUC
S1”	PreVIP-33	Annotation	0.73	0.409	0.881	0.832	0.285	0.248	0.7323
	PreVIP-193	Multiple	0.82	0.347	0.959	0.895	0.495	0.359	0.8034
	PreVIP-441	Multiple	0.73	0.492	0.911	0.867	0.391	0.365	0.8079
S2”	PreDIR-164	Proteomic sequence	0.53	0.658	0.762	0.758	0.109	0.194	0.7110
	PreDIR-129	Multiple	0.70	0.474	0.873	0.856	0.142	0.200	0.7057
S5[c]	PreVIP-193	Multiple	0.82	Sensitivity = 0.577					
	PreVIP-193	Multiple	0.50	Sensitivity = 0.906					
	PreVIP-441	Multiple	0.73	Sensitivity = 0.701					

	PreVIP-441	Multiple	0.50	Sensitivity = 0.910
S6[c]	PreVIP-193	Multiple	0.82	Sensitivity = 0.416
	PreVIP-193	Multiple	0.50	Sensitivity = 0.806
	PreVIP-441	Multiple	0.73	Sensitivity = 0.596
	PreVIP-441	Multiple	0.50	Sensitivity = 0.817

[a] *thresholds on S1'' and S2'' were set by maximizing the value of MCC. On testing dataset S5 and S6, two thresholds, i.e., 0.82 and 0.73 were set according to the best performance of PreVIP-193 and PreVIP-441 on testing dataset S1''. In addition, a neutral threshold (0.5) was added for crude assessments.* [c] *prediction results on testing dataset S5 and S6 are provided in* ***S3.5 Data****.* Abbreviations: MCC, Matthews Correlation Coefficient; AUC, area under the receiver operating characteristic curve; HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; PreVIP-33, machine learning model generated from training dataset S1' with the top 33 features for the VIP prediction task; PreVIP-193, machine learning model generated from training dataset S1' with the top 193 features for the VIP prediction task; PreVIP-441, machine learning model generated from training dataset S1' with the optimum 441 features for the VIP prediction task; PreDIR-164, machine learning model generated from training dataset S2' with 164 proteome-based features for the directionality prediction task; PreDIR-129, machine learning model generated from training dataset S2' with the optimum 129 features for the directionality prediction task.

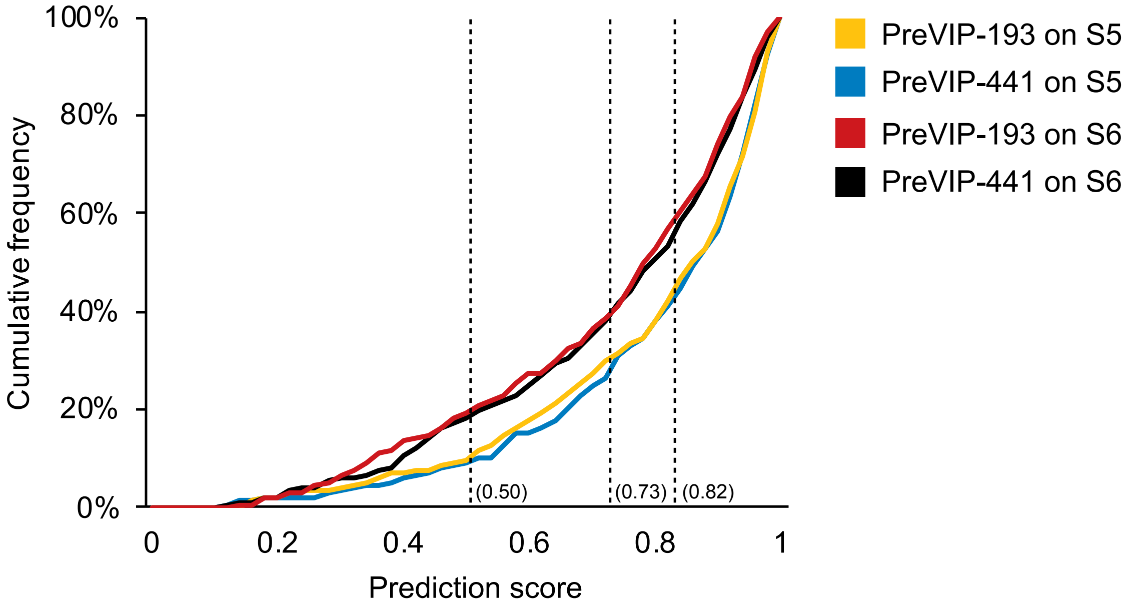


Fig 3.24 Cumulative distribution of prediction probabilities on testing datasets S5 and S6. Dataset S5 and S6 were retrieved from Reactome [150] and Gordon *et al.*'s study [254] for predicting VIPs. Composition of dataset S5 and S6 is provided in **S3.2 Data**. Prediction results on testing dataset S5 and S6 are provided in **S3.5 Data**. Abbreviations: VIPs, HIV-1 interacting human proteins; PreVIP-193, machine learning model generated from training dataset S1' with the top 193 features for the VIP prediction task; PreVIP-441, machine

learning model generated from training dataset S1' with the optimum 441 features for the VIP prediction task.

On the blind testing dataset S4, we used PreDIR-129 to predict the direction tag for 1351 'undefined' VIPs (**Fig 3.2**). According to known information about potential direction (**S3.1 Data**) and the recommendation stated in **Table 3.5**, 511, 540 and 300 'undefined' VIPs were predicted as 'backward', 'forward' and 'bidirectional', respectively (**S3.6 Data**). 66 of the 'undefined' VIPs had high probability to be 'backward' and the literature shows connections to brain-related diseases such as Autosomal recessive mental retardation [305] and Huntington disease [306] (**S3.7 Data**). 63 of them were very likely to be 'forward' VIPs and were involved in some immune system pathways (**S3.7 Data**). 50 of them showed strong signals of being 'bidirectional'. They might also be potential target of other viruses like human papillomavirus [307] and hepatitis virus [308] (**S3.7 Data**).

3.5 Discussion

In this study, we propose an *in silico* approach to investigate the HIV-1-host molecular interactions with a focus on the directionality of the virus-host interaction. We used the detailed curation of the biological nature of HIV-1-host interaction in HHID [241] to partition interactions as those required by the virus to manipulate the host molecular subsystems versus host responses to virus infection. Using this dataset, we designed a predictive system in which human proteins could be quickly evaluated for their potential to target host (a host-dependency factor), be targeted (e.g., the antiviral response), or both (bidirectional interactions). A web server is available at http://hivpre.cvr.gla.ac.uk/. It supports six different identifiers for over 80000 human peptides and is capable to produce 1000 predictions in less than 15 seconds.

In previous studies [172-174,224-229], VIPs were usually labeled based on their interacting HIV-1 status only. According to the data we retrieved from HHID [241], 1467 out of 3854 human proteins, including some key receptors (e.g. CD4, CCR5), have interactions with protein products of different HIV-1 genes (**S3.1 Data**). Such multi-target issues can be accommodated after integrating the information of molecular interactions between human proteins and the HIV-1 group. Additionally, the published prediction-based papers [172-174,224,225,227-229] have not accounted for the direction of HIV-1-host molecular interaction. By contrast, our consideration of the interaction direction contributes to a better understanding of the HIV-1-host interactions and the discovery of potential drug targets [309].

In experiments analysing evolution-related traits from the transcriptomic and genomics data (**Section 3.4.2.1**), we found HIV-1 were more likely to interact with human proteins encoded from genes with larger numbers of ORFs, duplication rates or evolutionary conservation. This seems reasonable as the evolutionary rates for duplicates is usually negatively correlated with the number of paralogues [310]. The accumulating exons and introns driven by conserved genes gave a higher chance for duplicate genes to have more ORFs, which were favoured by HIV-1 [311]. Additionally, we also observed the signal of host antiviral activities become stronger when the host genes showed huge number of ORFs, extremely high duplication rates, or evolutionary conservation.

We discovered 85 VIP-enriched SLiMs and 121 ‘backward’ VIP-enriched SLiMs from the proteomic sequence data (**Section 3.4.2.2**). Human proteins containing sufficient SLiMs of interest usually consisted of a higher number of residues. This phenomenon was even more evident in non-VIPs. We hypothesise that there are some motifs in the sequence of VIPs, making them more likely to target or be the target of HIV-1. Human proteins with long sequences have a higher possibility to include some special sequence patterns than those with short sequences but it does not indicate a feasible way to pay more attention to macromolecule to detect potential VIPs or ‘backward’ VIPs. For example, there are over 14000 residues in the sequence of a non-VIP, namely mucin 16 (MUC16), but only 17 VIP-enriched SLiMs were observed. However, this signal needs to be treated with caution especially when large numbers of VIP-enriched and ‘backward’ VIP-enriched SLiMs are both detected in the same non-VIP sequence, e.g. midasin (MDN1) (n = 41 and 57). Such ‘non-VIPs’ may potentially be false negatives if some of its SLiM-enriched regions are spatially exposed and attractive to HIV-1 [312].

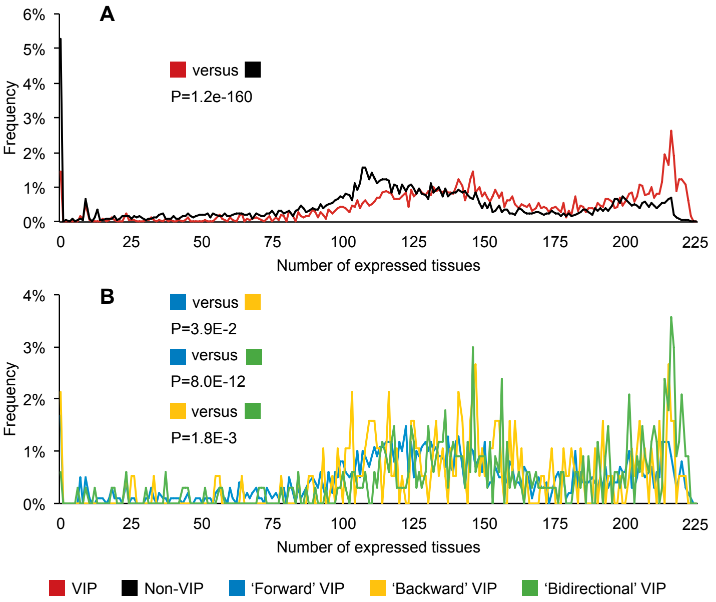


Fig 3.25 Tissue tropisms for different human proteins. Mann-Whitney U tests are applied to compare the number of expressed tissues in different classes. Abbreviations: HIV-1, human immunodeficiency virus type 1; VIPs, HIV-1 interacting human proteins; non-VIPs, HIV-1 non-interacting human proteins.

We obtained 225 experimentally verified tissue entries from the TISSUES database [245] but found some overlaps on feature values due to the hierarchy of tissues (**Section 3.4.2.3**). However, the annotation data of tissue tropisms were sufficient and such overlaps did not influence the efficacy of tissue tropisms in distinguishing VIPs from non-VIPs (**Fig 3.25**). The later machine learning experiments also proved the practical effectiveness of considering these features individually or in combination (**Fig 3.19A**, **Fig 3.21, Table 3.4**).

After finishing all prediction tasks, we assumed that some false negatives were still included in our dataset since we found some testing non-VIPs obtained very high prediction scores (**S3.8 Data**). Based on the testing result given by PreVIP-441 and PreVIP-193, 16 labelled non-VIPs might actually interact with HIV-1 proteins. We found that adapter molecule crk (CRK), TGF-beta-activated kinase 1 and MAP3K7-binding protein 1 (TAB1) and interleukin-1 receptor-associated kinase 4 (IRAK4) were involved in the pathway of HIV-1 infections in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [313] but were not included in HHID [241]. They provided further support for our machine learning approach. Some features of these human proteins also hinted at their possible roles as VIPs. For instance, alpha-synuclein (SNCA) had 16 ORFs, contained 15 VIP-enriched SLiMs within its 140-length proteomic sequence, expressed in many VIP-preferred tissues, and was highly connected with a degree of 168 in our constructed network [161]. As for the prediction of the interaction directionality, some deductions in **Table 3.5** might be ambiguous when being used individually but higher confidence could be obtained when combining the information about known directionality in **S3.1 Data**. For instance, elongin-B (ELOB) got a prediction score of 0.14 from PreDIR-129 thus was initially recommended as 'forward' VIPs according to **Table 3.5**. However, since we found 18 records on molecular interactions between ELOB and HIV-1 proteins and some outcomes of the interactions showed clear direction of 'backward', ELOB is probably 'bidirectional' rather than only 'forward' acting.

In conclusion, reliably predicting HIV-1-host molecular interactions is a difficult task and we were only superficially using the information embedded in the molecules involved. Here, we have introduced the directionality of the interaction to the task and demonstrated that there is a predictive signal embedded in the different types of molecules. We believe that better training datasets and continued development of feature representation of molecules, for example, integrating protein structure, will lead to improved predictions in the near future.

3.6 Summary

Human immunodeficiency virus type 1 (HIV-1) is the cause of acquired immunodeficiency syndrome (AIDS), a disease with no effective cure despite decades of research. A better understanding of the molecular interactions between HIV-1 and human host proteins facilitates the discovery of potential host targets which may be of great importance for the development of antiviral drugs that go beyond mere control of infection. In this study, we elucidate some host-dependency and antiviral factors that may be helpful to distinguish virus interacting human proteins (VIPs) from non-virus interacting human proteins (non-VIPs). We also consider 'directionality' in the HIV-1-host protein-protein interactions, i.e., whether the interaction is in the interest of the virus or part of the anti-viral response. We designed a machine learning framework to generate models based on the known information and used them for the classification of VIPs and non-VIPs. Our predictions have the potential to provide refined sets of human host targets aiding in the discovery of novel HIV-1 therapeutics.

3.7 Supporting information

S3.1 Data. The list of HIV-1 interacting proteins.

URL: http://hivpre.cvr.gla.ac.uk/data/VIP_3854.txt

S3.2 Data. The list of proteins sampled for training and testing.

URL: http://hivpre.cvr.gla.ac.uk/data/S2_Data.txt

S3.3 Data. Statistical data supporting discoveries in Chapter Three.

URL: http://hivpre.cvr.gla.ac.uk/data/S3_Data.xlsx

S3.4 Data. The importance and usage of features in different machine learning models.

URL: http://hivpre.cvr.gla.ac.uk/data/S4_Data.txt

S3.5 Data. Prediction results on dataset S5 and S6.

URL: http://hivpre.cvr.gla.ac.uk/data/S5_Data.txt

S3.6 Data. Prediction results for VIPs with 'undefined' directions.

URL: http://hivpre.cvr.gla.ac.uk/data/S6_Data.txt

S3.7 Data. 'Undefined' VIPs with very high probabilities indicating their directionality in HIV-1-host molecular interactions.

URL: http://hivpre.cvr.gla.ac.uk/data/S2_Table.xlsx

S3.8 Data. Non-VIPs with high prediction scores (>0.95) generated by PreVIP-441 and PreVIP-193.

URL: http://hivpre.cvr.gla.ac.uk/data/S3_Table.xlsx

CHAPTER FOUR

4. DEFINING FEATURES OF INTERFERON-ALPHA-STIMULATED HUMAN GENES

4.1 Abstract

A virus-infected cell triggers a signalling cascade resulting in the secretion of interferons (IFNs). It in turn induces the up-regulation of IFN-stimulated genes (ISGs) that play an important role in the inhibition of the viral infection. We conducted detailed analyses on 7443 features relating to evolutionary conservation, nucleotide composition, gene expression, amino acid composition, and network properties to elucidate factors associated with the stimulation of human genes in response to the typical IFN-α. We propose that ISGs are less evolutionary conserved than genes that are not significantly stimulated in IFN experiments (non-ISGs). ISGs show significant depletion of GC-content in the coding region of their canonical transcripts, which leads to differential representation in their nucleotide and amino acid compositions. ISG products tend to be implicated in key pathways of the human protein-protein interaction (PPI) network but are away from the hubs or bottlenecks. Interferon-repressed human genes (IRGs), which are down-regulated in the presence of IFNs, can have similar properties to ISGs. Meanwhile, we also propose a machine learning framework integrating the support vector machine (SVM) and feature selection algorithms. It achieves an area under the receiver operating characteristic curve (AUC) of 0.7455 for ISG prediction and demonstrates the similarity between ISGs triggered by type I and III IFNs. The model predicts several genes as potential ISGs that so far have shown no significant differential expression when stimulated with IFN in the cell/tissue types in the available databases. A webserver implementing our method is accessible at http://isgpre.cvr.gla.ac.uk/.

4.2 Introduction

Interferons (IFNs) are a family of cytokines originally defined for their capacity to interfere with viral replication. They are secreted from host cells after an infection by pathogens such as bacteria or viruses to trigger the innate immune response with the aim of inhibiting viral spread by 'warning' uninfected cells [105]. The response induced by IFNs is usually fast and feedforward, especially to synthesize new IFNs, which guarantees a full response even if the initial activation is limited [314]. In humans, several IFNs have been discovered (e.g. IFN-α/β/ε/κ/ω/γ/λ [315-320]). IFN-α, IFN-β, IFN-ε, IFN-κ, IFN-ω are grouped into type I IFNs for signalling through the common IFN-α receptor (IFNAR) complex present on target cells [315-318] (**Fig 4.1A**). IFN-α comprises 13 subtypes in humans while the remaining type I IFNs are encoded by a specific gene [321]. IFN-λ targets IFN-λ receptor 1 (IFNLR1)/interleukin-10 receptor 2 (IL-10R2) and was classified as type III IFN since its discovery in 2003 [320] (**Fig 4.1C**). Similar to type I IFNs, IFN-λ also exert antiviral properties but functions less intensely [103,322,323]. IFN-γ is classified as type II IFN and manifest its biological effects by interacting with IFN-γ receptor (IFNGR) [319] (**Fig 4.1B**). In contrast to type I and III IFNs, IFN-γ is also anti-pathogen, immunomodulatory, and proinflammatory but more focused on establishing cell immunity [103,315,319,324].

All three types of IFNs are capable of activating the Janus kinase/signal transducer and activator of transcription (JAK-STAT) pathway and inducing the transcriptional up-regulation of approximately 10% of human genes that prime cells for stronger pathogen detections and defenses [321,325,326]. Henceforth, these up-regulated human genes are referred to as IFN-stimulated genes (ISGs). They play an important role in the establishment of the cellular antiviral state, the inhibition of viral infection and the return to cellular homeostasis [315,321,325,327]. For example, the ectopic expression of heparinase (HPSE) can inhibit the attachment of multiple viruses [328,329]; interferon induced transmembrane proteins (IFITM) can impair the entry of multiple viruses and traffic viral particles to degradative lysosomes [330,331]; MX dynamin like GTPase proteins (MX) can effectively block early steps of multiple viral replication cycles [332]. Abnormality in the IFN-signalling cascade, for example, the absence of signal transducer and activator of transcription 1 (STAT1) will lead to the failure of activating ISGs, making the host cell highly susceptible to virus infections [104].

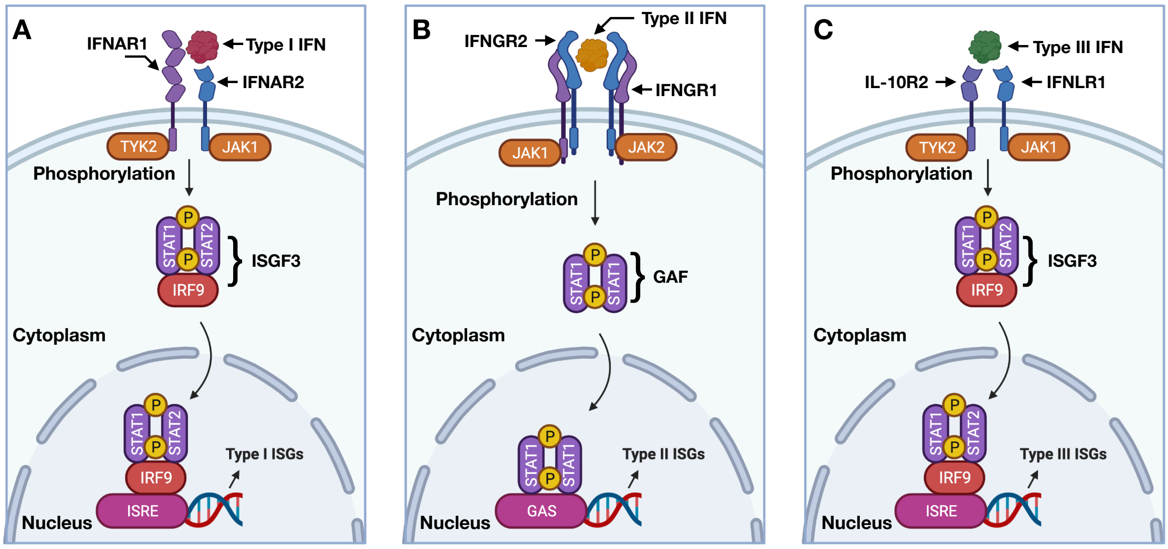


Fig 4.1 Illustration of signalling cascade triggered by different IFNs. In (A), type I IFN signals through IFNAR, JAK1, TYK2, STAT, and IRF9 to form ISGF3, and then bind to ISRE to induce the expression of type I ISGs. In (B), type II IFN signals through IFNGR, JAK1 and JAK2 to form GAF and then bind to GAS to induce the expression of type II ISGs. In (C), type III IFN signals through IFNLR1, IL-10R2, JAK1, TYK2, STAT, and IRF9 to form ISGF3, and then bind to ISRE to induce the expression of type III ISGs. Figure created using the BioRender (https://biorender.com/) under the Creative Commons licence 4.0. Abbreviations: IFNs, interferons; IFNAR, IFN-α receptor; ISGs, interferon-stimulated (up-regulated) human genes; JAK1, Janus kinase 1;; STAT, signal transducers and activators of transcription; ISGF3, IFN-stimulated gene factor 3 complex; ISRE, IFN-stimulated response elements; IFNGR, IFN-γ receptor; JAK2, Janus kinase 2; GAF, IFN-γ activation factor; GAS, gamma-activated sequence promoter elements. IFNLR1, IFN-λ receptor 1; IL-10R2, interleukin-10 receptor 2; TYK2, tyrosine kinase 2; IRF9, IFN-regulatory factor 9.

Most research on ISGs has focused on elucidating the role of ISGs in antiviral activities or discovering new ISGs within or across species [315,321,325,330,333,334]. The identification of ISGs can be achieved via various approaches. Associating gene expression with suppression of viral infection is a good strategy to identify ISGs with obvious antiviral performance, exemplified by the influenza inhibitor, MX dynamin like GTPase 1 (MX1), and the human immunodeficiency virus 1 inhibitor, MX dynamin like GTPase 2 (MX2) [332]. CRISPR screening is a loss-of-function experimental approach to identify ISGs required for IFN-mediated inhibition to viruses. It enabled the discovery of tripartite motif containing 5 (TRIM5), MX2 and bone marrow stromal cell antigen 2 (BST2) [335]. Monitoring the ectopic expression of ISGs is another instrumental way to find some ISGs (e.g., interferon stimulated exonuclease gene 20 (ISG20) and ISG15 ubiquitin like modifier

(ISG15)) that are individually sufficient for viral suppression [336]. Using RNA-sequencing [143] and fold change-based criterion to measure whether a target human gene is induced by IFN signalling now has become a well-accepted idea [334,337,338]. In most cases, a gene is defined as IFN stimulated (up-regulated) when its expression value is more than doubled with the presence of IFNs (fold change > 2) [315,334,339]. There are also many online databases to support IFN- or ISG-related research. For example, the Interferome (http://www.interferome.org) provides an excellent resource by compiling *in vivo* and *in vitro* gene expression profiles in the context of IFN stimulation [334]. The Orthologous Clusters of Interferon-stimulated Genes (OCISG, http://isg.data.cvr.ac.uk) demonstrates an evolutionary comparative approach of genes differentially expressed in type I IFN system for ten different species [315].

We notice that the same human gene may show differential response to different IFNs in different tissues or cells [334]. Despite some well-investigated ISGs, the majority of classified ISGs only have limited up-regulation following IFN stimulations [315,334]. It means that the difference between ISGs and those human genes not significantly up-regulated in the presence of IFNs (non-ISGs) may not be obvious especially when being assessed more generally. It should also be noted that, within non-ISGs, there are a group of genes down-regulated during IFN stimulations. We refer to them as interferon-repressed human genes (IRGs) and they constitute another major part of the IFN regulation system [315,340]. Collectively, the complex nature of the IFN-stimulated system results in knowledge that is far from comprehensive.

In this study, we characterise the properties of human genes stimulated by IFN-α. We propose that it is feasible to make ISG predictions on human genes with a model only compiled from the knowledge of IFN-α responses in human fibroblast cells. To achieve these ends, we first constructed a refined high-confidence dataset consisting of 620 ISGs and 874 non-ISGs by cross-checking the genes across multiple databases including the OCISG [315], Interferome [334], and Reference Sequence (RefSeq) [243]. The analyses were conducted primarily on our refined data using genome- and proteome-based features that were likely to influence the expression of human genes in the presence of IFN-α. Then based on the calculated features, we designed a machine learning framework with an optimised feature selection strategy for the prediction of putative ISGs in different IFN systems. Finally, we also developed an online web server that implemented our machine learning method at http://isgpre.cvr.gla.ac.uk/.

4.3 Methods

4.3.1 Dataset curation

In this study, we retrieved 2054 ISGs (up-regulated), 12379 non-ISGs (down-regulated or not differentially expressed), and 3944 unlabelled human genes (expression-limited genes (ELGs) with less than one count per million reads mapping across the three biological replicates [341,342]) from the OCISG (http://isg.data.cvr.ac.uk/) [315]. Gene clusters in the OCISG were built through the Ensembl Compara [343], which provided a thorough account of gene orthology based on whole genomes available in the Ensembl 103 [244]. Labels of these human genes were defined based on the fold change and a false discovery rate (FDR) following IFN-α treatments in human fibroblast cells. We searched the collected 18377 entries against the RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/) (ver. GRCh38) [243] to decipher features based on appropriate transcripts (canonical) [344] coding for the main functional isoforms of these human genes. It produced 1315, 7304, and 2217 results for ISGs, non-ISGs and ELGs, respectively. These 10836 human genes were well-annotated by multiple online databases and were used as the background dataset S1 in the analyses.

For the purpose of generating a set of human genes with high confidence of being up-regulated and non-up-regulated in response to the IFN-α, we searched the recompiled 8619 human genes (ISGs or non-ISGs) against the Interferome 2.0 (http://www.interferome.org/) [334]. We filtered out ISGs without high up-regulation (Log_2(Fold Change) > 1.0) or with obvious down-regulation (Log_2(Fold Change) < -1.0) in the presence of type I IFNs. This procedure guaranteed a refined ISGs dataset with strong levels of stimulation induced by any type I IFNs and reduced biases driven by IRGs for the analyses and predictions. We filtered out non-ISGs showing enhanced expression after type I IFN treatments (Log_2(Fold Change) > 0). The exclusion of these non-ISGs could effectively reduce the risk of involving false negatives in analyses and producing false positives in predictions. As a result, the refined dataset S2 contains 620 ISGs and 874 non-ISGs with relatively high confidence.

The training procedure in the machine learning framework was conducted on the balanced dataset S2’. It consisted of 992 randomly selected ISGs and non-ISGs from dataset S2. The remaining human genes in S2 were used for independent testing. Additionally, we also constructed another six testing datasets for the purpose of review and assessment. Dataset S3 contained 695 ISGs with low confidence compared to those ISGs in dataset S2.

Some of them could be non-ISGs or even IRGs in the type I IFN system. Dataset S4 contained 1006 IRGs from the human fibroblast cell experiments. Dataset S5, S6, and S7 were constructed based on records for experiments in type I, II, and III IFN systems from the Interferome [334]. The criterion for an ISG in the latter three datasets was a high level of up-regulation (Log$_2$(Fold Change) > 1.0) while that for non-ISGs was no up-regulation after IFN treatments (Log$_2$(Fold Change) < 0). The last testing dataset S8 was derived from our background dataset S1, containing 2217 ELGs. A breakdown of the aforementioned eight datasets is shown in **Table 4.1**. Detailed information of the human genes used in this study is provided in **S4.1 Data**.

Table 4.1 A breakdown of datasets used in this study.

Dataset	Brief description	IFN system	ISGs	Non-ISGs	ELGs
S1	Well-annotated human genes	IFN-α in fibroblast cells	1315	7304	2217
S2	Refined dataset with high confidence	IFN-α in fibroblast cells	620	874	0
S2’	Training subset of S2	IFN-α in fibroblast cells	496	496	0
S2”	Testing subset of S2	IFN-α in fibroblast cells	124	378	0
S3	ISGs with low confidence in S1	IFN-α in fibroblast cells	695	0	0
S4	IRGs divided from S1	IFN-α in fibroblast cells	0	1006	0
S5	ISGs from the Interferome [334]	Type I IFNs in all cells	1259	872	0
S6	ISGs from the Interferome [334]	Type II IFN in all cells	2229	755	0
S7	ISGs from the Interferome [334]	Type III IFN in all cells	33	1683	0
S8	ELGs divided from S1	IFN-α in fibroblast cells	0	0	2217

Abbreviations: ISGs, interferon-stimulated (up-regulated) human genes; non-ISGs, human genes not significantly up-regulated by interferons; IRGs, interferon-repressed (down-regulated) human genes, ELGs, human genes with limited expression in interferon-α experiments.

4.3.2 Generation of discrete features

We encoded 397 discrete features from aspects of evolution, nucleotide composition, transcription, amino acid composition, and network preference. From the perspective of evolution, we used the number of transcripts, open reading frames (ORFs) and protein-coding exons in the canonical transcripts to quantify the alternative splicing process. Genes with more transcripts and ORFs have higher alternative splicing diversity to produce proteins with similar or different biological functions [255,345,346]. Genes with more protein-coding exons in their canonical transcripts can produce more complex alternative splicing products [288]. Here, duplication and mutation features were measured by the number of within species paralogues and substitutions [347,348]. These data were collected from the BioMart (ver. February 2021) [244] to assess the selection on protein sequences and mutational processes affecting the human genome [259].

From the perspective of nucleotide composition, we calculated the percent of adenine, thymine, cytosine, guanine, and their four-category combinations in the coding region of the canonical transcript. The first category measured the proportion of two different nitrogenous bases out of the implied four bases, e.g., GC-content. The second category also focused on the combination of two nucleotides but added the impact of phosphodiester bonds along the 5' to 3' direction, e.g., CpG-content [257]. The third category calculated the occurrence frequency of 4-mers, e.g., 'CGCG' composition to involve some positional resolution [349]. The last category considered the co-occurrence of some short linear motifs (SLims) in the complementary DNA (SLim_DNAs). From the perspective of transcription, we calculated the usage of 61 coding codons and three stop codons in the coding region of the canonical transcripts. Codon usage biases are observed when there are multiple codons available for coding one specific amino acid. They can affect the dynamics of translation thus regulate the efficiency of translation and even the folding of the proteins [258,350].

From the perspective of amino acid composition, we calculated the percentage of 20 standard amino acids and their combinations based on their physicochemical properties [351]. Patterns in the amino acid level are considered to have a direct impact on the establishment of biological functions or to reflect the result of strong purifying selection [153]. Based on the chemical properties of the side chain, we grouped amino acids into seven classes including aliphatic, aromatic, sulfur, hydroxyl, acidic, amide, and basic amino acids. We also grouped amino acids based on geometric volume, hydropathy, charge status, and polarity, but found some overlaps among these features. For instance, amino acids with basic side chains are all positively charged. Aromatic amino acids all have large geometric volumes (volume > 180 cubic angstroms). Likewise, we also considered the co-occurrence of some SLims at the protein level. These co-occurring SLims in the protein sequence (SLim_AAs) may relate to potential mechanisms regulating the expression of ISGs [352].

When trying to measure the network preference for the gene products, we constructed a human PPI network based on 332,698 experimentally verified interactions with at least medium confidence (HIPPIE score > 0.63) from HIPPIE 2.2 [161]. Nodes and edges of this network are provided at http://isgpre.cvr.gla.ac.uk/. Eight network-based features including the average shortest path, closeness, betweenness, stress, degree, neighbourhood connectivity, clustering coefficient, and topological coefficient were calculated from this network. Isolated nodes or proteins were not included in our network and were assigned zero value for all these eight features. The shortest path measures the average length of the shortest path between a focused node and others in the network. Closeness of a node is defined as the reciprocal of the length of the average shortest path. Proteins with a low value of the shortest paths or closeness are close to the centre of the network. Betweenness reflects the degree of control that one node exerted over the interactions of other nodes in the network [353]. Stress of a node measures the number of shortest paths passing through it. Proteins with a high value of betweenness or stress are close to the bottleneck of the network. Degree of a node counts the number of edges linked to it while neighbourhood connectivity reflected the average degree of its neighbours. Proteins with high degree or neighbourhood connectivity are close to the hub of the network. They are considered to play an important role in the establishment of the stable structure of the human interactome [354]. Clustering and topological coefficient measure the possibility of a node to form clusters or topological structures with shared neighbours. The former coefficient can be used to identify the modular organisation of metabolic networks [355] while the latter one may be helpful to find out virus mimicry targets [356].

4.3.3 Generation of categorical features

In this study, categorical features were used to check the occurrence of SLims in the genome and proteome. SLim_DNAs constructed in this study contained three to five random nucleotides, producing 708,540 alternative choices. SLim_DNAs with no restrictions on their first or last position were not taken into consideration as their patterns could be expressed in a more concise way. A SLim_DNA was picked out to encode a binary feature when its occurrence level in the coding region of the canonical ISG transcripts was significantly higher than that for non-ISGs (Pearson's chi-squared test: $p < 0.05$). SLim_AAs were constructed with three to four fixed amino acids separated by putative gaps. The gap could be occupied by at most one random amino acid, producing 1,312,000 alternative choices. Likewise, binary features were prepared for SLim_AAs showing significant enrichment in ISGs products than in non-ISG products (Pearson's chi-squared test: $P < 0.05$). Since there were lots of results rejecting the null-hypothesis, we adopted the Benjamini-Hochberg correction procedure to avoid type I error [263]. Additionally, we also encoded two features to check the co-occurrence or absence of multiple SLim_DNAs and SLim_AAs. This co-occurrence status might be a better representation of functional sites composed of short stretches of adjacent nucleobases or amino acids surrounding SLim_DNAs or SLim_AAs [153].

4.3.4 Assessment of associations between feature representation and IFN-triggered stimulations

We obtained 8619 human genes with expression data from the OCISG [315]. 4111 of them were annotated with a positive $\mathrm{Log_2}$(Fold Change) ranging from 0 to 12.6, which meant they were up-regulated after IFN-α treatments in human fibroblast cells. In order to associate the feature representation with IFN-induced stimulation ($\mathrm{Log_2}$(Fold Change) > 0), we first introduced a 0.1-length sliding-window to group human genes with similar expression in the IFN experiments. For instance, the first group contains 1116 human genes with a $\mathrm{Log_2}$(Fold Change) ranging from 0 to 0.1. Then the representation values of a focused feature f are averaged within each group, denoted as $\overline{V_\iota}$. Finally, Pearson's correlation coefficient (PCC) was introduced for the measurement:

$$PCC(f) = \frac{1}{n-1}\sum\nolimits_{i=1}^{n}\left(\frac{LFC_i - M_0}{SD_0}\right) \times \left(\frac{\overline{V_\iota} - M_f}{SD_f}\right) \tag{4.1}$$

where n is the number of divided parts that equals to 126 in this study; LFC_i and $\overline{V_\iota}$ are the average value of $\mathrm{Log_2}$(Fold Change) and feature representation in the i-th group; M_0 and SD_0 are the mean and standard deviation of $\mathrm{Log_2}$(Fold Change), which is set as 6.4 and 3.7 respectively in this study; M_f and SD_f are the mean and standard deviation of 126 $\overline{V}$ that reflect the representation of the focused feature. To make fair comparisons among features with different scales, we normalised them based on the major value of their representations:

$$Norm(f) = \begin{cases} 1, f > UB(f) \\ \dfrac{f - LB(f)}{UB(f) - LB(f)}, LB(f) < f < UB(f) \\ 0, f < LB(f) \end{cases} \tag{4.2}$$

where $LB(f)$ and $UB(f)$ are the lower and upper bound representing the 5th and 95th percentile within representation values for the target feature. The representation of feature was considered to have a stronger positive/negative association with IFN-α-triggered stimulations if the PCC calculated from the normalised features was closer to 1.0/-1.0 and the p value calculated by the Student t-test was lower than 0.05.

4.3.5 Machine learning and optimisation

We designed a machine learning framework for the prediction of ISGs. Firstly, all features were encoded and normalised based on their major representations (**Equation 4.2**). Then we used an under-sampling procedure [265] to generate a balanced dataset from dataset S2 for training and modelling. Support vector machine (SVM) with radial basis function (RBF) [200] was used as the basic classifier. It maps the normalised feature space to a higher dimension to generate a space plane to better classify the majority of positive and negative samples. Since there were usually lots of noisy data distributed in the feature space, it was necessary to remove disruptive features. This effectively reduced the dimensionality of the feature space and made it easier for the SVM model to generate a more appropriate classification plane that involved fewer false positives and false negatives. Here, we propose a subtractive iteration algorithm driven by the change of area under the receiver operating characteristic curve (AUC). It is close to a backward feature elimination strategy [357] but reserves non-disruptive features in each iteration (**Fig 4.2**). In each iteration, we traversed the features and removed those that do not improve the AUC of the prediction results. Theoretically, this algorithm can greatly optimise the feature space and remove all disruptive features after multiple iterations. In the testing procedure, we encoded the optimum features for testing samples and place them in the optimised feature space. Samples with longer distance to the optimised classification plane indicated a stronger signal of being ISGs or non-ISGs. They were more likely to get higher prediction scores (close to 0 or 1) from the SVM model.

```
BEGIN
Initialisation: Balanced dataset $S_0 = \{(1, v_1^0), .. (1, v_n^0), (0, v_{n+1}^0) ... (0, v_{2n}^0)\}$, dimension of the feature vector $D_0$, machine learning algorithm $A$, number of disruptive feature $d_0 = D_0$, and iteration round $i = 0$.
  While $d_0 > 0$ ($i^{th}$ iteration):
    1) Use five-fold cross validation on dataset $S_i$, prediction $P_i = A(S_i)$;
    2) Evaluate the $P_i$ with the criterion of AUC;
    3) Remove one feature from feature vector $v^i$ and generate a temporary dataset $T_i$;
    4) Use five-fold cross validation on dataset $T_i$, prediction $P'_i = A(T_i)$;
    5) Evaluate the $P'_i$ with the criterion of AUC;
    6) Repeat 4) and 5) for the traversal of $D_i$ features;
    7) Traverse $v^i$ and remove $m$ features helpful to improve AUC of $P'_i$, $d_i = m$;
    8) Update dataset $S_{i+1} = \{(1, v_1^{i+1}), .. (1, v_n^{i+1}), (0, v_{n+1}^{i+1}) ... (0, v_{2n}^{i+1})\}$, $D_{i+1} = D_i - m$.
  End
  Output: dataset $S_{i-1}$ encoded by $D_{i-1}$ features.
END
```

Fig 4.2 The pseudo-code of the AUC-driven subtractive iteration algorithm. Abbreviations: AUC, area under the receiver operating characteristic curve.

4.3.6 Performance evaluation

In this study, the prediction results were evaluated with three threshold-dependent criteria including sensitivity (SN), specificity (SP), and Matthews correlation coefficient (MCC) [279] and two threshold-independent criteria: SN_n and AUC. SN and SP were used to assess the quality of the machine learning model in recognising ISGs and non-ISGs respectively while MCC provided a comprehensive evaluation for both positives and negatives. The number of ‘n’ in the SN_n criterion was determined based on the number of ISGs used for testing. It was used to measure the upper limit of the prediction model as well as to check the existence of important false positives close to the class of ISGs from the perspective of data expression. Finally, AUC was a widely used criterion to evaluate the prediction ability of a binary classifier system. The group of interest was almost unpredictable in a specific binary classifier system if the AUC of the classifier was close to 0.5.

4.4 Results

4.4.1 Evolutionary characteristics of ISGs

In this study, we constructed a dataset consisting of 620 ISGs and 874 non-ISGs (dataset S2) from 10836 well-annotated human genes (dataset S1). Human genes in dataset S2 were considered to have higher confidence based on their records in both the OCISG [315] and Interferome [334] databases. Human genes in both dataset S1 or S2 were evolutionarily unrelated as they were retrieved from the OCISG [315] that compiled clusters of orthologous genes based on whole-genome alignments. However, they might still have inherent characteristics that resulted in their different expressions in response to IFNs-α. Here, we explored features relating to alternative splicing [345], duplication [347] and mutation [348]. We used the number of ORFs and transcripts in a human gene to represent the diversity of alternative splicing for a human gene and use the number of protein-coding exons in the canonical transcript to reflect the complexity of the alternative splicing. By calculating the average number of ORFs with respect to different Log_2(Fold Change) levels of expression (window size = 0.1) in the presence of IFN-α, we found that human genes with higher Log_2(Fold Change) tended to have less ORFs (PCC = -0.287, **Fig 4.3A**). Although small number of ORFs seemed to be associated with obvious IFN-α up-regulation, it was not a necessary condition. Compared to the background human genes in dataset S1, we found that ISGs tended to have more ORFs, but these differences did not reach a statistical significance

(Mann-Whitney U test: $p > 0.05$). As for the latter two features related to the transcripts and protein-coding exons, similar negative relationships were observed when Log2(Fold Change) increased (**Fig 4.3B & 5.3C**). Particularly, as the lowest value of Log2(Fold Change) for human genes not differentially expressed only reached around -0.9. Points placed left to the boundary (x = -0.9) are all IRGs. They are generally placed below those non-ISGs with a Log2(Fold Change) around zero, suggesting the three features (number of ORFs, transcripts and exons) are all differentially represented in some IRGs compared to the remaining non-ISGs. This distribution also indicates that some IRGs had similar feature patterns to ISGs, especially to those highly up-regulated in the presence of IFN-α (right part of the scatter plots in **Fig 4.3**).

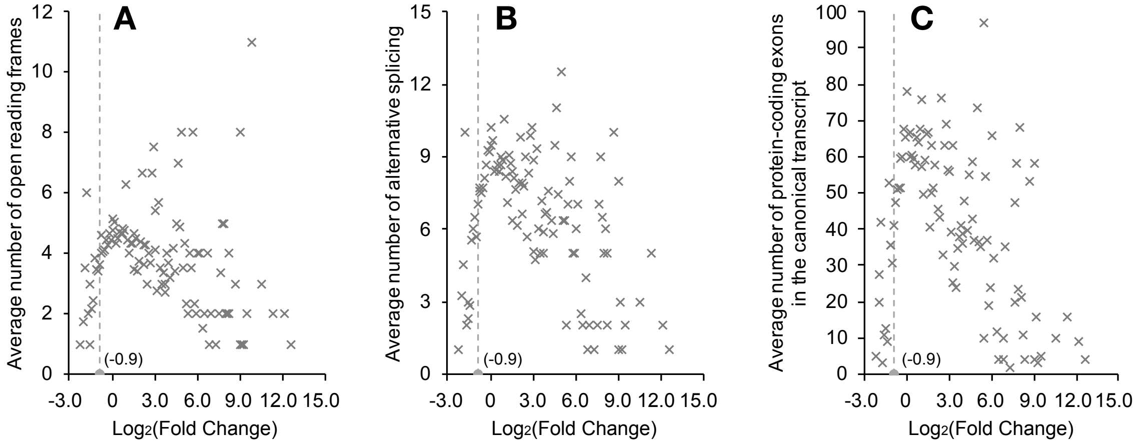


Fig 4.3 The average representation of features associated with IFN-α stimulations in human fibroblast cells. (A) The number of ORFs and (B) transcripts as measurements of the diversity of alternative splicing process. (C) The number of protein-coding exons in the canonical transcript as a measurement of the complexity of alternative splicing process. These three plots are drawn based on the expression data of 8619 human genes with valid fold change in the IFN-α experiments (**S4.1 Data**). ELGs are excluded as they had insufficient read coverage to determine a fold change in the experiments. Points in the scatter plot are located based on the average feature representation of genes with similar expression performance in experiments. Abbreviations: IFN, interferon; ORFs, open reading frames; ELGs, human genes with limited expression in IFN-α experiments.

To determine whether ISGs tend to originate from duplications, we counted the number of within human paralogs of each gene (**Fig 4.4A**). The results showed that there were around 22% of singletons in our main dataset, whilst ISGs had 15% and non-ISGs had 26%. The result of a Mann-Whitney U test [358] indicated that the number of paralogs was significantly under-represented in ISGs compared to the background human genes in dataset S1 (M_1 = 10.5, M_2 = 11.5, p = 8.8E-03). We hypothesized that such a difference was mainly caused by the imbalanced distribution of singletons in ISGs and non-ISGs. The differences became smaller when singletons are excluded from the test (M_1 = 12.4, M_2 = 14.6, $p > 0.05$). Next, we used the number of non-synonymous substitutions per non-synonymous site (dN) and synonymous substitutions per synonymous site (dS) within human paralogues as a measurement of differences in mutational signatures between different classes [359]. As shown in **Fig 4.4B**, non-synonymous substitutions are more frequently observed in ISGs than in background human genes (M_1 = 0.62, M_2 = 0.55, p = 4.0E-03). On the other hand, ISGs also have a higher frequency of synonymous substitutions than background human genes (M_1 = 37.7, M_2 = 34.6, p = 1.1E-02) (**Fig 4.4C**) but the difference is not as obvious as for non-synonymous substitutions. In **Fig 4.4D**, the distribution of dN/dS ratios within human paralogues indicates that most human genes are constrained by natural selection but ISGs, in general, tend to be less conserved (M_1 = 0.036, M_2 = 0.045, p = 8.3E-03). When eliminating the influence of duplication events, ISGs are still less conserved than non-ISGs but the difference in the dN/dS ratio is not significant (M_1 = 0.053, M_2 = 0.031, $p > 0.05$).

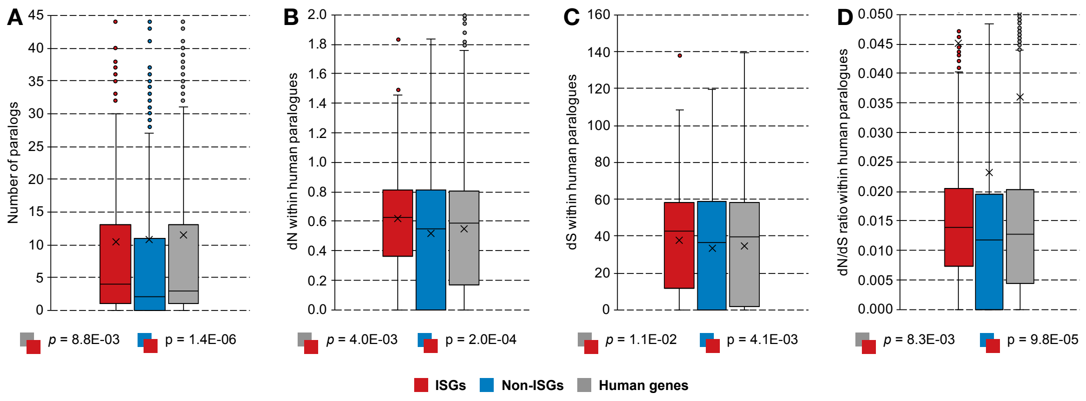


Fig 4.4 Differences in the evolutionary constraints of human genes. (A) Paralogues within *Homo sapiens*. (B) Non-synonymous substitutions within human paralogues. (C) Synonymous substitutions within human paralogues. (D) dN/dS ratios within human paralogues. Here, ISGs and non-ISGs are taken from dataset S2 while the background human genes are from dataset S1 (**Table 4.1**). Mann-Whitney U tests are applied to compare the feature distribution of different classes. Boxes in the plot represent the major distribution of values (from the first to the third quartile); outliers are added for values higher than two-fold of the third quartile; cross symbol marks the position of the average value including the outliers; upper and lower whiskers show the maximum and minimum values excluding the outliers. Abbreviations: ISGs, interferon-α-stimulated (up-regulated) genes; non-ISGs,

human genes not significantly up-regulated by interferon-α; dN, non-synonymous substitutions per non-synonymous site; dS, synonymous substitutions per synonymous site.

4.4.2 Differences in the coding region of the canonical transcripts

Compared to general profile features (e.g., number of ORFs), the sequences themselves provide more direct mapping to the protein function and structure [360]. Here, we encoded 344 discrete features and 7026 categorical features from complementary DNA (cDNA) of the canonical transcript to explore features specific to ISGs. We divided the discrete features into four categories and compared their representations among three different groups of human genes: recompiled ISGs from dataset S2, recompiled non-ISGs from dataset S2, and the background human genes from dataset S1 (**Fig 4.5**). Firstly, guanine and cytosine were both more depleted in ISGs than non-ISGs, leading to an under-representation of GC-content in ISGs (Mann-Whitney U test: M_1 = 52%, M_2 = 55%, p = 2.3E-11). This attribute was antithetical to the GC-biased gene conversion (gBGC), making ISGs less stable with weak evolutionary conservation (**Fig 4.4**) [361]. Additionally, the under-representation of GC-content also influenced the representation of other dinucleotide features. Among all dinucleotide depletions in ISGs, CpG composition was ranked the first followed by GpG and GpC composition (p = 2.9E-14, 4.9E-13 and 1.2E-10, respectively). In turn, adenine and thymine-related dinucleotide compositions, exemplified by ApT and TpA were more enriched in ISGs than non-ISGs (p = 8.0E-10 and 8.5E-10, respectively).

Next, we compared the usage of 64 different codons in the third category as their frequencies influence transcription efficiency [350]. Differences between ISGs and background human genes were observed in codons for 11 amino acids including leucine (L), isoleucine (I), valine (V), serine (S), threonine (T), alanine (A), glutamine (Q), lysine (K), glutamic acid (E), arginine (R), and glycine (G). The most significant difference was observed in the usage of codon 'AGA'. Among all arginine-targeted alternative codons, codon 'AGA' was usually favoured, and its usage reached an estimated 25% in ISGs, but reduced to 22% in the background human genes (p = 1.4E-05). It was even significantly lower in non-ISGs, at 18% (p = 1.9E-13). On the other hand, compared to background human genes, the codon 'CAG' coding for amino acid 'Q' was the most under-represented in ISGs. It was less favoured by ISGs than non-ISGs (M_1 = 72%, M_2 = 78%, p = 7.3E-13) although it dominated in coding patterns. As for the three stop codons, comparing with background human genes, the usage of the ochre stop codon ('TAA') was over-represented in ISGs (M_1 = 28%, M_2 = 33%, p = 9.7E-03). In this category of codon usage, the features with different frequencies between ISGs and background human genes became more discriminating when comparing ISGs with non-ISGs. Significant differences in codon usages between ISGs and non-ISGs were widely observed except for methionine (M) and tryptophan (W). Hence, despite the limited differences of codon usages between ISGs and the background human genes, these features were useful for discriminating ISGs from non-ISGs.

In the last category, we calculated the occurrence frequency of 256 nucleotide 4-mers to add some positional resolution for finding and comparing interesting organisational structures [349]. Among the 256 4-mers, 46 were differentially represented between ISGs and background human genes (**S4.2 Data**). Most of these 4-mers were over-represented by ISGs except two with the pattern 'TAAA' and 'CGCG'. Interestingly, the feature of 'TAAA' composition became a positive factor when comparing ISGs and non-ISGs (M_1 = 4.1%, M_2 = 3.7%, p = 4.1E-06), suggesting it might be a good feature to discern potential or incorrectly labelled ISGs. We found six nucleotide 4-mers: 'ACCC', 'AGTC', 'AGTG', 'TGCT', 'GACC', and 'GTGC' were over-represented in ISGs when compared to background human genes but they were not differentially represented when comparing ISGs with non-ISGs. These six features might be inherently biased for some reasons and were not powerful enough to distinguish ISGs from non-ISGs. In addition to the aforementioned 40 features (except 4-mer 'ACCC', 'AGTC', 'AGTG', 'TGCT', 'GACC', and 'GTGC') that were differentially represented in ISGs compared to background human genes, we found a further 39 features nucleotide 4-mers differentially represented between ISGs and non-ISGs (**S4.2 Data**).

To check the effect of these aforementioned 343 features on the level of stimulation in the IFN system (Log_2(Fold Change) > 0), we calculated the PCC for the normalised features (**Equation 4.2**) and found 106 features were positively related to the increase of fold change, and 34 features were suppressed when human gene were more up-regulated after IFN-α treatments (Student t-test: $p < 0.05$) (**S4.3 Data**). ApA composition showed the most obvious positive correlation with stimulation level (PCC = 0.464, p = 8.8E-06) while negative association between the representation of 4-mer 'CGCG' and IFN-α-induced up-regulation was the most significant (PCC = -0.593, p = 3.2E-09). Human genes with higher up-regulation in the presence of IFN-α contained more codons 'CAA' rather than 'CAG' for coding amino acid 'Q'. The depletion of GC-content, especially cytosine content, promotes the suppression of many nucleotide compositions in the cDNA, e.g. CpG composition.

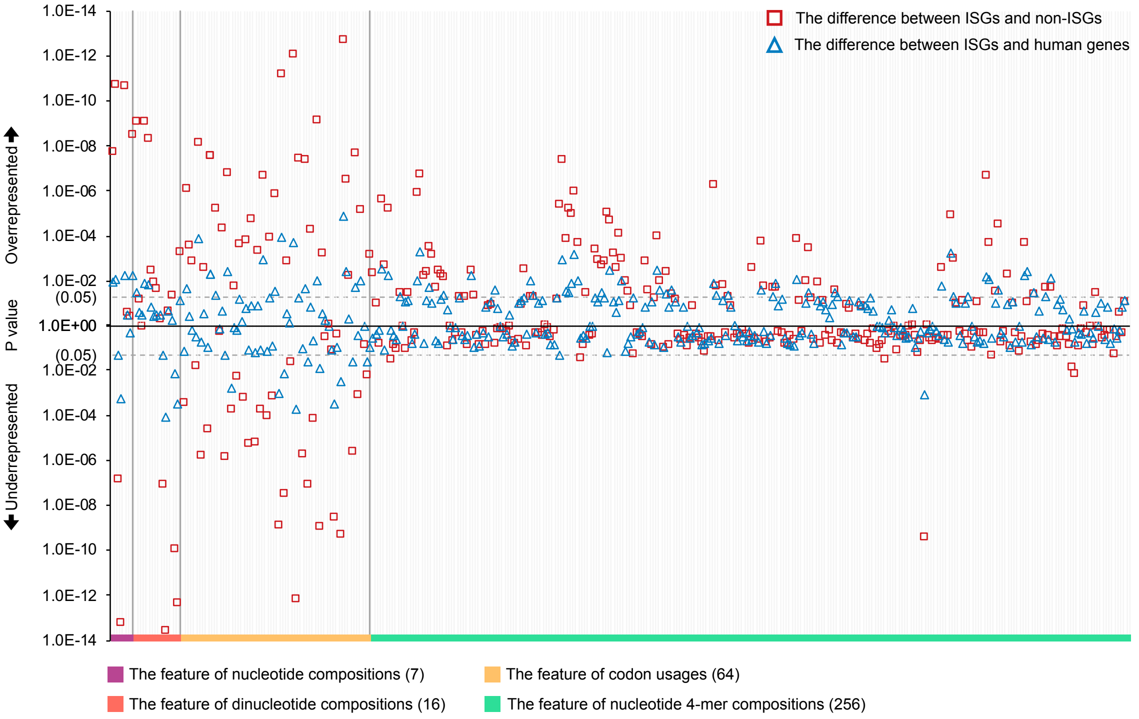


Fig 4.5 Differences in the representation of discrete features encoded from coding regions of the canonical transcript. Mann-Whitney U tests were applied for hypothesis testing and the results are provided in the **S4.2 Data**. Here, ISGs and non-ISGs are taken from dataset S2 while the background human genes are from dataset S1 (**Table 4.1**). Abbreviations: ISGs, interferon-α-stimulated (up-regulated) human genes; non-ISGs, human genes not significantly up-regulated by interferon-α.

To find conserved sequence patterns related to gene regulations [362], we checked the existence of 2940, 44100 and 661500 short linear nucleotide motifs (SLim_DNAs) consisting of three to five consecutive nucleobases in the group of ISGs and non-ISGs. By using a positive 5% difference in the occurrence frequency as cut-off threshold, we found 7884 SLim_DNAs with a maximum difference in representation around 15%. After using Pearson's chi-squared tests and Benjamini-Hochberg correction to avoid type I error in multiple hypotheses [263], 7025 SLim_DNAs remained with an adjusted p-value lower than 0.01 (**S4.4 Data**), hereon referred to as flagged SLim_DNAs. The differentially represented 7025 SLim_DNAs were ranked according to the adjusted p-value. As shown in **Fig 4.6A**, dinucleotide ‘TpA’ dominates in the top 10, top 100, top 1000, and all differentially represented SLim_DNAs even if TpA representation is suppressed in the cDNA of genes’ canonical transcripts compared to other dinucleotides. Dinucleotide ‘ApT’ and ‘ApA’ are also frequently observed in the flagged SLim_DNAs but their occurrences do not show significant difference in the top 100 SLim_DNAs (Pearson's chi-squared test: $p > 0.05$). GC-related dinucleotides, e.g., 'CpC', 'GpC' and 'GpG' are rarely observed in the flagged SLim_DNAs especially in the top 10 or top 100. In view of these, we hypothesize that the differential representation of nucleotide compositions influences and reflects on the pattern of SLim_DNAs in ISGs. By checking the co-occurrence status of the flagged SLim_DNAs, we found that these sequence patterns had a cumulative effect in distinguishing ISGs from non-ISGs especially when the number of cooccurring SLim_DNAs reached around 5320 (Pearson's chi-squared test: p = 7.9E-13, **Fig 4.6B**). There were eight (~1.3%) ISGs in the dataset S2 containing all the flagged 7025 SLim_DNAs. Their up-regulation after IFN treatment were generally low with a fold change fluctuating around 2.2. However, some of these eight genes such as desmoplakin (DSP) were clearly highly up-regulated in endothelial cells isolated from human umbilical cord veins after not only IFN-α treatments (fold change = 11.1) but also IFN-β treatments (fold change = 13.7). We also found some non-ISGs (e.g., hemicentin 1 (HMCN1)) and ELGs (e.g. tudor domain containing 6 (TDRD6)) containing the flagged SLim_DNAs, but their frequencies were lower than that in ISGs. Although there is an obvious imbalance between the number of ISGs and non-ISGs in the human genome [9-11], the curve for the background human genes in **Fig 4.6B** is still closer to that for ISGs rather than that for non-ISGs. It suggests that some genetic patterns are widely represented in the coding region of human genes, making them potentially up-regulated in the IFN-α system.

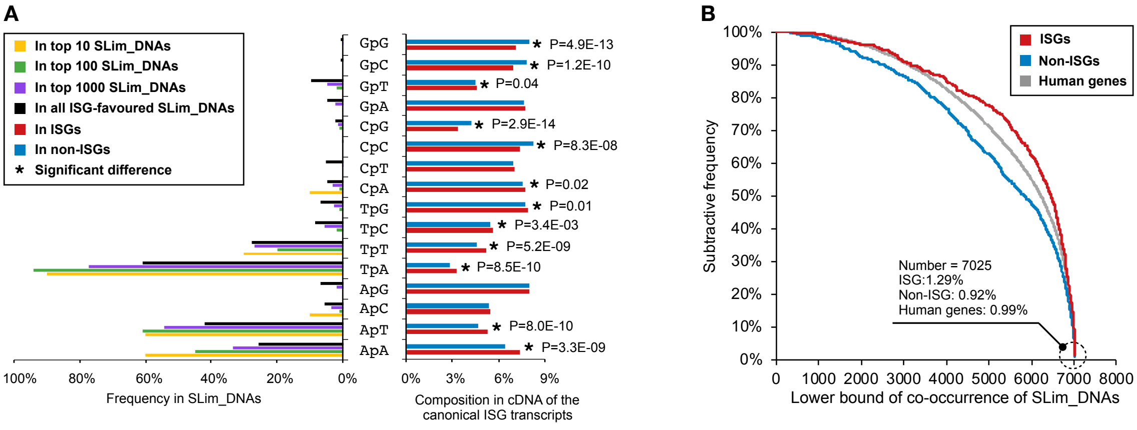


Fig 4.6 The pattern of SLim_DNAs in the coding region of the canonical transcripts. (A) Influence of dinucleotide compositions on the flagged SLim_DNAs. (B) The co-occurrence status of SLim_DNAs in different human genes. Ranks in (A) are generated based on the adjust p value given by Pearson's chi-squared tests after Benjamini-Hochberg correction procedure [264]. Detailed results of the hypothesis tests are provided in **S4.4 Data**. Here, ISGs and non-ISGs are taken from dataset S2 while the background human genes are from dataset S1 (**Table 4.1**). Abbreviations: ISGs, interferon-α-stimulated (up-regulated)

human genes; non-ISGs, human genes not significantly up-regulated by interferon-α; SLim_DNAs, short linear nucleotide motifs; cDNA, complementary DNA.

4.4.3 Differences in the protein sequence

We used the protein sequences generated by the canonical transcript to extract features at the proteomic level. In addition to the basic composition of 20 standard amino acids, we considered 17 additional features related to physicochemical (e.g., hydropathy and polarity) or geometric properties (e.g., volume) [363,364]. We found several amino acids that are either enriched or depleted in ISG products compared to background human proteins, which were produced by genes in dataset S1 (**Fig 4.7**). The differences were even more marked between protein products of ISGs and non-ISGs, highlighting some differences that were not observed when comparing ISG products to the background human proteins (e.g., isoleucine composition). The differences observed in the amino acid compositions were at least in part associated with the patterns previously observed in features encoded from genetic coding regions. For example, asparagine (N) showed significant over-representation in ISG products compared to non-ISG products or background human proteins (Mann-Whitney U test: p = 2.8E-12 and 1.2E-03, respectively). This was expected as there are only two codons, i.e., 'AAT' and 'AAC' coding for amino acid 'N', and dinucleotide 'ApA' showed a remarkable enrichment in the coding region of ISGs. A similar explanation could be given for the relationship between the deficiency of GpG content and amino acid 'G'. The translation of amino acid 'K' was also influenced by ApA composition but was not significant due to the mild representation of dinucleotide 'ApG' in the genetic coding region. Additionally, as previously mentioned, ISGs showed a significant depletion in the CpG content, and consequently, the amino acid 'A' and 'R' in ISG products were significantly under-represented. Cysteine (C) was not frequently observed in human proteins but still showed a relatively significant enrichment in ISG products (M_1 = 2.3%, M_2 = 2.5%, p = 1.8E-03).

When focusing on the composition of amino acids grouped by physicochemical or geometric properties, we found some features differentially represented between ISG products and background human proteins. The result showed that hydroxyl (amino acid 'S' and 'T'), amide (amino acid 'N' and 'Q'), or sulfur amino acids (amino acid 'C' and 'M') were more abundant in ISG products compared to the background human proteins (Mann-Whitney U test: p = 0.04, 1.0E-03 and 0.02, respectively). Small amino acids (amino acid 'N', 'C', 'T', aspartic acid (D) and proline (P), the volume ranges from 108.5 to 116.1 cubic angstroms) were more frequently observed in ISG products than in background human proteins ($M_1 = 22.1\%$, $M_2 = 21.7\%$, $p = 0.02$). These differences became more marked when comparing the representation of these features between ISG and non-ISG products. For example, features relating to chemical properties of the side chain (e.g., aliphatic), charge status and geometric volume showed differences between proteins produced by ISGs and non-ISGs. Some features such as neutral amino acids that include amino acid ‘G’, ‘P’, ‘S’, ‘T’, histidine (H) and tyrosine (Y) were not differentially represented between ISG and non-ISG products, but they indicated obvious association with the change of IFN-α-triggered stimulations (PCC = -0.556, p = 4.1E-08) (**S4.3 Data**).

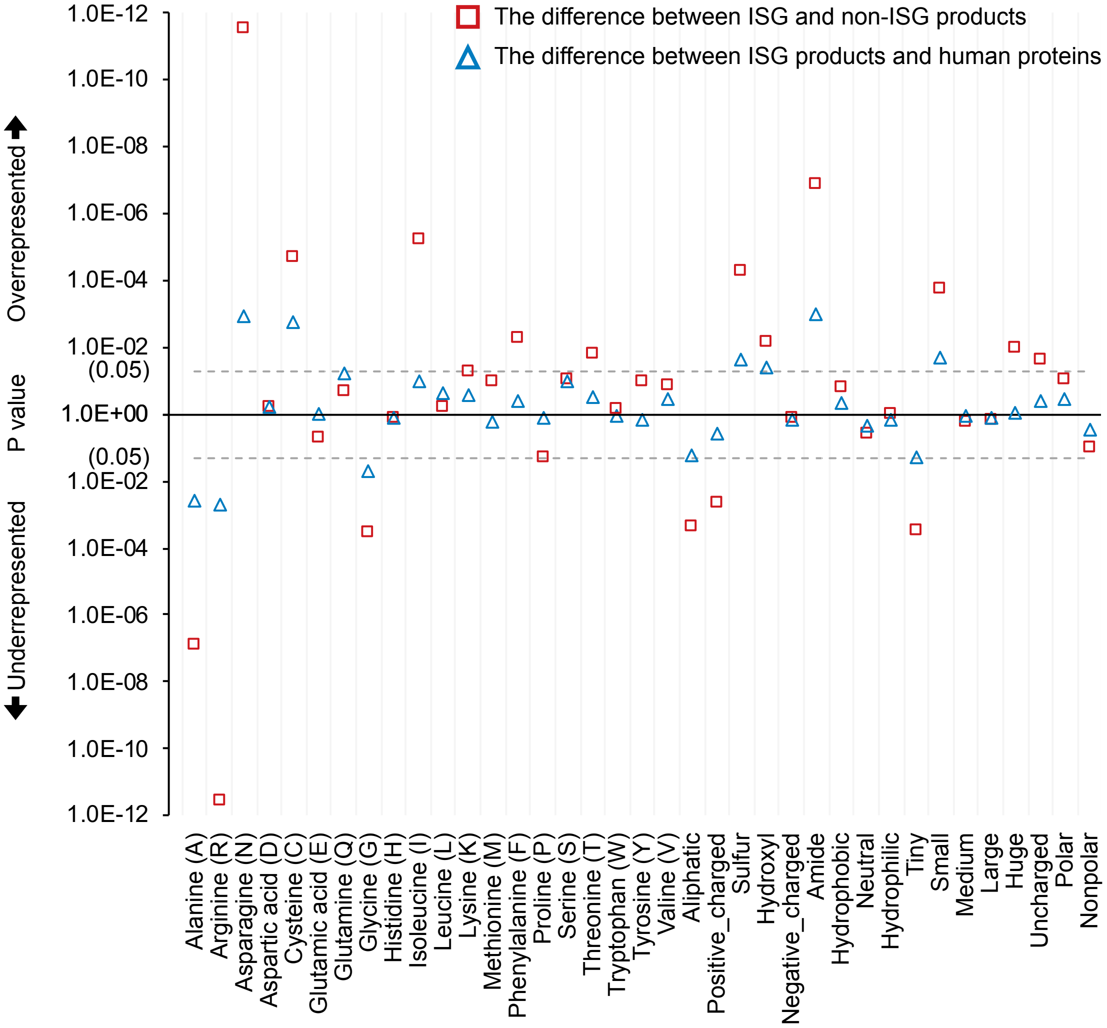


Fig 4.7 Differences in the representation of discrete features encoded from protein sequences. Mann-Whitney U tests are applied for hypothesis testing and the results were provided in the **S4.2 Data**. Here, ISGs and non-ISGs are taken from dataset S2 while the background human genes are from dataset S1 (**Table 4.1**). Aliphatic group: amino acid ‘A’, ‘G’, ‘I’, ‘L’, ‘P’ and ‘V’; aromatic/huge group: amino acid ‘F’, ‘W’ and ‘Y’ (volume > 180

cubic angstroms); sulfur group: amino acid ‘C’ and ‘M’; hydroxyl group: amino acid ‘S’ and ‘T’; acidic/negative_charged group: amino acid ‘D’ and ‘E’; amide group: amino acid ‘N’ and ‘Q’; positive_charged group: amino acid ‘R’, ‘H’ and ‘K’; hydrophobic group: amino acid ‘A’, ‘C’, ‘I’, ‘L’, ‘M’, ‘F’, ‘V’, and ‘W’ that participates to the hydrophobic core of the structural domains [351]; neutral group: amino acid ‘G’, ‘H’, ‘P’, ‘S’, ‘T’ and ‘Y’; hydrophilic group: amino acid ‘R’, ‘N’, ‘D’, ‘Q’, ‘E’ and ‘K’; Tiny group: amino acid ‘G’, ‘A’ and ‘S’ (volume < 90 cubic angstroms); small group: amino acid ‘N’, ‘D’, ‘C’, ‘P’ and ‘T’ (volume ranged from 109 to 116 cubic angstroms); medium group: amino acid ‘Q’, ‘E’, ‘H’ and ‘V’ (volume ranged within 138 to 153 cubic angstroms); large group: amino acid ‘R’, ‘I’, ‘L’, ‘K’ and ‘M’ (volume ranged within 163 to 173 cubic angstroms); uncharged group: the remaining 15 amino acids except electrically charged ones; polar group: amino acid ‘R’, ‘H’, ‘K’, ‘D’, ‘E’, ‘N’, ‘Q’, ‘S’, ‘T’ and ‘Y’; nonpolar group: the remaining 10 amino acids except polar ones. Abbreviations: ISG, interferon-α-stimulated (up-regulated) human genes; non-ISG, human genes not significantly up-regulated by interferon-α.

We then searched the sequence of ISG products against that of non-ISG products to find conserved short linear amino acid motifs (SLim_AAs), which might have resulted from strong purifying selection [153]. As opposed to the analysis on the genetic sequence, we only obtained 19 enriched sequence patterns with a Pearson's chi-squared p value ranging from 1.5E-04 to 0.02 (**Table 4.2**). These SLim_AAs were greatly influenced by four polar amino acids: ‘K’, ‘N’, ‘E’ and ‘S’, and one nonpolar amino acid: ‘L’. Some of these SLim_AAs, for example, SLim ‘NVT’ and ‘S-N-E’, were clearly over-represented in ISG products compared to background human proteins and could be used as features to differentiate ISGs from background human genes. The third column in **Table 4.2** indicates a number of patterns that are lacking in non-ISG products and hence may be the reason for the lack of up-regulation in the presence of IFN-α. Particularly, we noticed that SLim ‘KEN’ was a destruction motif that could be recognised or targeted by anaphase promoting complex (APC) for polyubiquitination and proteasome-mediated degradation [365,366]. Results shown in **Fig 4.8A** illustrate that the co-occurrence of differentially represented SLim_AAs has a cumulative effect in distinguishing ISGs from non-ISGs. This cumulative effect can be achieved with only two random SLim_AAs (Pearson's chi-squared test: $p = 4.6E\text{-}10$). The bias in the co-occurring SLim_AAs in the background human proteins towards a pattern similar to non-ISG products further proves the importance of these 19 SLim_AAs. However, their co-occurrence is not associated with the level of IFN-triggered stimulations (PCC = 0.015, $p > 0.05$) (**Fig 4.8B**).

Regions that lack stable structures under normal physiological conditions within proteins are termed intrinsically disordered regions (IDRs). They play an important role in cell signalling [367]. Compared with ordered regions, IDRs are usually more accessible and have multiple binding motifs, which can potentially bind to multiple partners [293]. According to the results calculated by the IUPred [267], we found 6721, 10510, and 119071 IDRs (IUpred score no less than 0.5) in proteins produced by ISGs, non-ISGs and background human genes respectively. We hypothesized that enriched SLims widely detected in IDRs might be important for human protein-protein interactions or potentially virus mimicry [356]. For instance, in ISG products, 29 out of 71 SLim 'S-N-T' were observed in IDRs (~40.8%), 14.9% higher than that in non-ISG products (**Table 4.2**). This difference reflected the importance of SLim 'S-N-T' for target specificity of IFN-α-induced PPIs [321] even if it was not statistically significant. By contrast, the conditional frequency of SLim 'S-N-E' discovered in IDRs of ISG and non-ISG products were almost the same, indicating that SLim 'S-N-E' might have an association with some inherent attributes of ISGs but was less likely to be involved in IFN-α-induced PPIs. SLim 'KEN' in IDRs also showed some interesting differences: in non-ISG products, 41.9% of SLim 'KEN' were observed in IDRs, 14.6% higher than that in ISG products, which provided an effective approach to distinguish ISGs from non-ISGs. When SLim 'KEN' is discovered in the ordered region of a protein sequence, statistically, the protein is more likely to be produced by an ISG, but this assumption is reversed if the SLim is located in an IDR (Pearson's chi-squared tests: $p = 0.03$). Despite the relatively low conditional frequency of SLim 'KEN' in the IDRs of ISG products, these SLim_AAs in the IDR are more likely to be functionally active than those falling within ordered globular regions [368].

Table 4.2 Representation of SLims in protein sequences and their IDRs.

SLims[a]	Frequency in ISG/non-ISG products[b]	Bias based on the frequency in human proteins	P value[c]	Conditional frequency in IDRs of ISG/non-ISG products/background human proteins[c,d]	P value[e]
S-N-E	15.2%/8.8%	+47.6%/-14.2%	1.5E-04	39.4%/40.3%/33.4%	0.90
ENE	15.0%/8.8%	+20.9%/-29.0%	2.1E-04	37.6%/42.9%/40.9%	0.49
S-N-T	11.5%/6.2%	+21.9%/-34.2%	2.9E-04	40.8%/25.9%/27.3%	0.08
SVI	15.2%/9.2%	+37.6%/-16.9%	3.6E-04	18.1%/11.3%/15.2%	0.21
L-NL	23.7%/16.4%	+13.2%/-21.9%	4.0E-04	10.2%/11.9%/9.4%	0.65
L-KL	30.8%/22.8%	+18.0%/-12.8%	4.9E-04	12.6%/10.1%/8.7%	0.43
NVT	13.7%/8.5%	+52.1%/-6.1%	1.2E-03	18.8%/21.6%/15.4%	0.66
ISS	20.5%/14.3%	+20.7%/-15.7%	1.7E-03	29.9%/25.6%/23.8%	0.44
LK-K	24.4%/17.7%	+24.5%/-9.3%	1.8E-03	14.6%/20.6%/20.0%	0.16
IK-E	14.2%/9.0%	+34.2%/-14.5%	1.8E-03	26.1%/16.5%/25.8%	0.13
EK-I	15.8%/10.4%	+31.0%/-13.7%	2.0E-03	15.3%/20.9%/16.0%	0.32

K-E-S	16.9%/11.4%	+21.9%/-17.7%	2.4E-03	36.2%/36.0%/39.2%	0.98
LNS	17.7%/12.1%	+21.2%/-17.1%	2.4E-03	20.0%/25.5%/20.5%	0.34
KEN	16.0%/10.6%	+33.5%/-11.0%	2.4E-03	27.3%/41.9%/34.8%	0.03
L-N-L	22.6%/17.5%	+14.3%/-11.4%	1.5E-02	10.7%/11.8%/9.5%	0.78
K-E-L	25.8%/20.5%	+25.7%/-0.3%	1.5E-02	18.8%/17.9%/18.7%	0.84
KLL	27.1%/21.9%	+9.9%/-11.4%	1.9E-02	11.3%/8.4%/9.9%	0.35
LKE	29.8%/24.5%	+18.2%/-3.0%	2.1E-02	19.5%/24.8%/20.1%	0.20
LK-L	33.2%/27.7%	+15.0%/-4.2%	2.1E-02	7.8%/12.4%/10.0%	0.11

*a: the dash symbol in SLims indicates one position occupied by a standard amino acid; b: here, ISGs and non-ISGs are taken from dataset S2 while the background human genes use samples in dataset S1 (**Table 4.1**); c: p values in this column use Pearson's chi-squared tests to measure the difference of SLim occurrences in ISG and non-ISG products; d: frequencies in this column are calculated based on a condition that corresponding SLims are observed in the protein sequence; e: p values in this column use Pearson's chi-squared tests to measure the difference of SLim occurrences in IDRs of ISG and non-ISG products.* Abbreviations: SLims, short linear motifs; ISGs, interferon-α-stimulated (up-regulated) human genes; non-ISGs, human genes not significantly up-regulated by interferon-α; IDRs, intrinsically disordered regions.

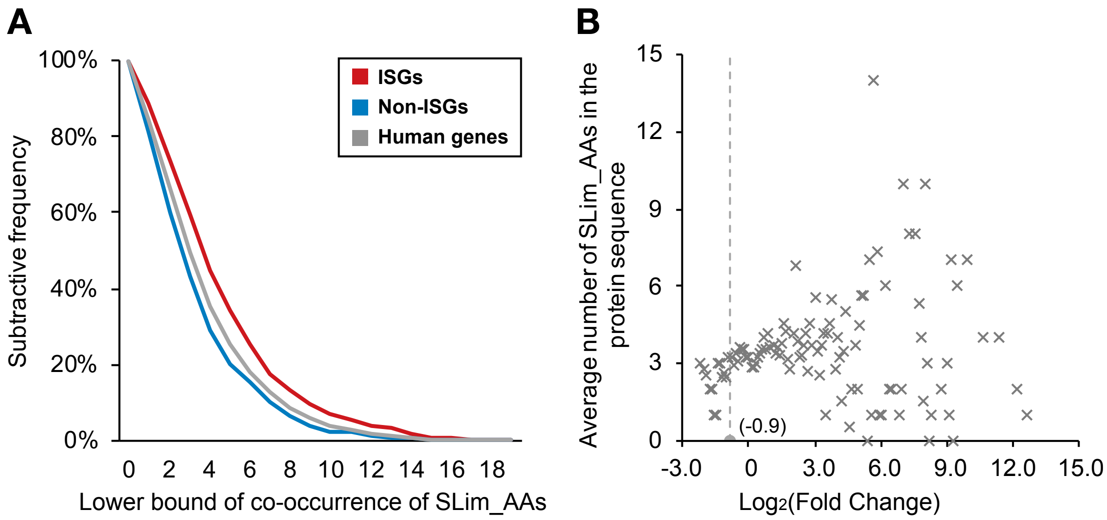


Fig 4.8 Representation of co-occurring SLim_AAs in our main dataset. (A) The co-occurrence status of SLim_AAs in different classes. (B) Relationship between co-occurrence of the marked SLim_AAs and Log$_2$(Fold Change) after IFN-α treatments. Here, ISGs and non-ISGs are taken from dataset S2 while the background human genes are from dataset S1 (**Table 4.1**). Points in (B) are located based on the average feature representation of genes with similar expression performance in IFN-α experiments. Abbreviations: IFN, interferon; ISGs, interferon-α-stimulated human genes; non-ISGs, human genes not significantly up-regulated by IFN-α; SLim_AAs, short linear amino acid motifs.

4.4.4 Differences in network profiles

We constructed a network with 332,698 experimentally verified interactions among 17603 human proteins (medium confidence, HIPPIE score > 0.63) from the HIPPIE 2.0 database [161]. 10169 out of 10836 human proteins from our background dataset S1 were included in it. Nodes and edges of this network can be downloaded from our webserver at http://isgpre.cvr.gla.ac.uk/. Based on this network, we calculated eight features including the average shortest path, closeness, betweenness, stress, degree, neighbourhood connectivity, clustering coefficient, and topological coefficient. As illustrated in **Fig 4.9**, ISG products tend to have higher values of betweenness and stress than background human proteins (Mann-Whitney U test: $p = 0.01$, and 0.03, respectively), which means they are more likely to locate at key paths connecting different nodes of the PPI network. Some ISG products with high values of betweenness and stress, e.g., tripartite motif containing 25 (TRIM25), can be considered as the shortcut or bottleneck of the network and play important roles in many PPIs including those related to the IFN-α-triggered immune activities [369,370]. However, the over-representation of betweenness does not mean ISG products are more likely to be or even be close to bottlenecks in the network compared to background human proteins. Some examples shown in **Table 4.3** indicate that ISG products are less-connected by top-ranked bottlenecks and hubs in the network than non-ISGs or background human proteins. This conclusion is not influenced by hub/bottleneck protein's performance in the IFN-α experiments. Comparing proteins produced by ISGs and non-ISGs, we found the former tends to have lower values of clustering coefficient and neighbourhood connectivity (Mann-Whitney U test: $p = 0.04$, and 7.9E-03, respectively). This discovery indicates that ISG products and the majority of their interacting proteins are less likely to be targeted by lots of proteins. It also supports the finding that ISG products are involved in many shortest paths for nodes but are away from hubs or bottlenecks in the network. To some extents, this location also increases the length of the average shortest paths through ISG products in the network.

When investigating the association between IFN-induced gene stimulation and network attributes of gene products, we only found the feature of neighbourhood connectivity was under-represented as the level of differential expression in the presence of IFN increases (PCC = -0.392, $p = 2.2E-04$). This suggests that proteins produced by genes that are highly up-regulated in response to IFN-α are further away from hubs in the PPI networks.

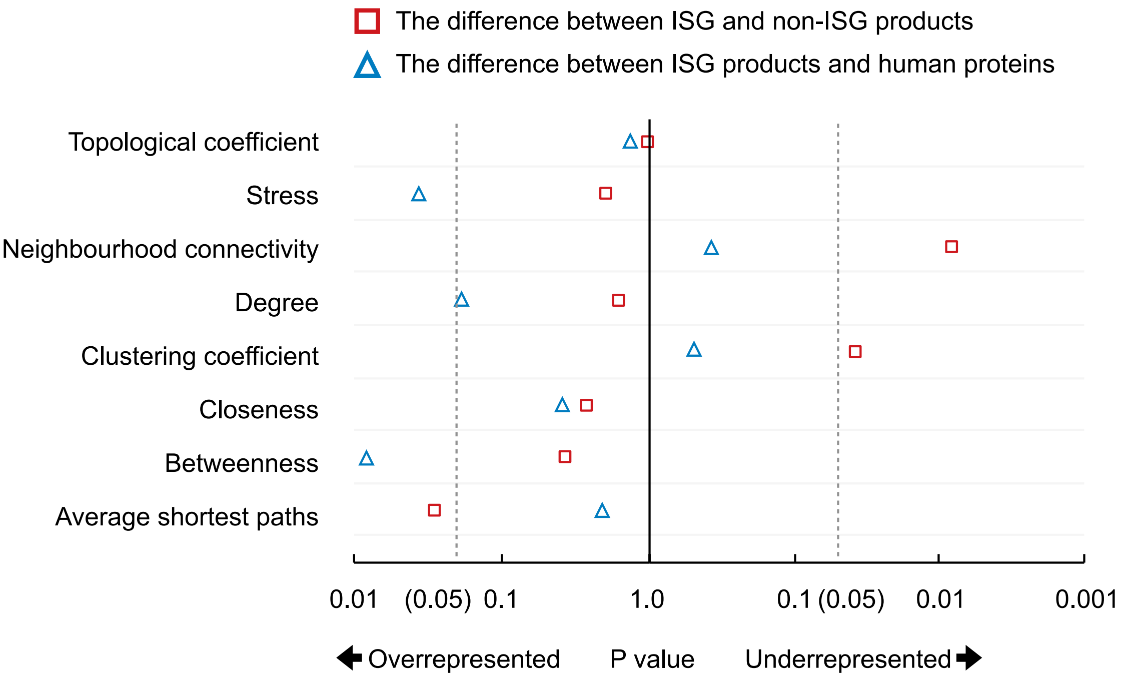


Fig 4.9 Differential network preferences of proteins coded by different human genes. Mann-Whitney U tests are applied for hypothesis testing and the results were provided in the **S4.2 Data**. Here, ISGs and non-ISGs are taken from dataset S2 while the background human genes use samples in dataset S1 (**Table 4.1**). Abbreviations: ISGs, interferon-α-stimulated (up-regulated) genes; non-ISGs, human genes not significantly up-regulated by interferon-α.

Table 4.3 Interaction profiles of human proteins connecting top hubs/bottlenecks of the HIPPIE network.

Human protein	TRIM25	ELAVL1	ESR2	NTRK1	HNRNPL
Gene class	ISG	IRG	Not included in S1[a]		
Degree (hub rank)	2295 (2nd)	1787 (4th)	2500 (1st)	1976 (3rd)	1681 (5th)
Betweenness (bottleneck rank)	0.067 (1st)	0.048 (4th)	0.051 (3rd)	0.026 (5th)	0.052 (2nd)
Difference in interacting partners (ISG products versus non-ISG)[b]	Depleted P = 0.01	P > 0.05	Depleted P = 1.1E-4	Depleted P = 5.5E-3	P > 0.05
Difference in interacting partners (ISG products versus background human proteins)[b]	P > 0.05	P > 0.05	Depleted P = 8.1E-3	Depleted P = 0.03	P > 0.05

a: ESR2 and NTRK1 are not included in dataset S1 as their expression data were not compiled in the OCISG, HNRNPL is not included in dataset S1 as its canonical isoform was uncertain when the dataset was constructed; b:differences here are measured via Pearson's chi-squared tests on human proteins interacting with the corresponding hub/bottleneck protein. Abbreviations: HIPPIE, Human Integrated Protein-Protein Interaction rEference database; TRIM25, tripartite motif containing 25; ELAVL1, embryonic lethal, abnormal vision like

RNA binding protein 1; ESR2, estrogen receptor 2; NTRK1, neurotrophic receptor tyrosine kinase 1; HNRNPL, heterogeneous nuclear ribonucleoprotein L; ISGs, interferon-α-stimulated (up-regulated) human genes; non-ISGs, human genes not significantly stimulated by interferon-α.

4.4.5 Features highly associated with the level of IFN stimulations

In this study, we encoded a total of 397 discrete and 7046 categorical features covering the aspects of evolutionary conservation, nucleotide composition, transcription, amino acid composition, and network profiles. In order to find out some key factors that may enhance or suppress the stimulation of human genes in the IFN-α system, we compared the representation of discrete features of human genes with different but positive $\mathrm{Log_2}$(Fold Change). Two features on the co-occurrence of SLims were not taken into consideration here as they were more subjective than the other discrete features and were greatly influenced by the number of focused SLims. Upon the calculation of PCC and the result of hypothesis tests, we found 168 features highly associated with the level of IFN-α-triggered stimulations (Student t-tests: $p < 0.05$) (**S4.3 Data**). Among them, 118 features showed a positive correlation (**Fig 4.10**) while the remaining 50 features showed a negative correlation (**Fig 4.11**) with the change of up-regulation in IFN-α experiments. Among these 168 features, the number of ORF, alternative splicing results, and exons in the canonical transcripts were encoded from characteristics of the gene. Average dN/dS and average dS within human paralogues were encoded based on the sequence alignment results from the Ensembl [244]. 140 and 22 features were encoded from the genetic sequence and proteomic sequence respectively. The last one, neighbourhood connectivity, was obtained from the network profile of a human interactome constructed based on experimentally verified data in the HIPPIE [161].

In the positive group, the feature of 'large' amino acid compositions that includes the composition of five amino acids with geometric volume ranged from 163 to 173 cubic angstroms was ranked the first for having the highest PCC at 0.593 (Student t-test: p = 2.8E-09). This feature was not highlighted previously as it did not have a strong signal for discriminating ISGs from non-ISGs (Mann-Whitney U test: $p > 0.05$). Similar phenomena were found on 87 features (64 positive correlations and 23 negative correlations) such as AG-content, ApG content and previously mentioned neutral amino acid composition. The strongest negative correlation between feature representation and IFN-α-triggered stimulations was found on the feature of 4-mer 'CGCG' (PCC = -0.593, p = 3.2E-09). This feature also showed a differential distribution between ISGs and non-ISGs, thus provided useful information to distinguish ISGs from non-ISGs. Similar phenomena were found on 81 features (54 positive correlations and 27 negative correlations) such as previously mentioned GC-content, CpG content and the usage of codon ‘GCG’ coding for amino acid ‘A’. Collectively, the biased effect on the basic composition of nucleotides influences the correlation between the representation of sequence-based features and IFN-α-triggered stimulations. Human genes that show over-representation in more features listed in **Fig 4.10** are expected to be more up-regulated after IFN-α treatments at least in human fibroblast cells. Meanwhile, the under-representation of features listed in **Fig 4.11** also contributes to the level of up-regulation in the IFN-α experiments.

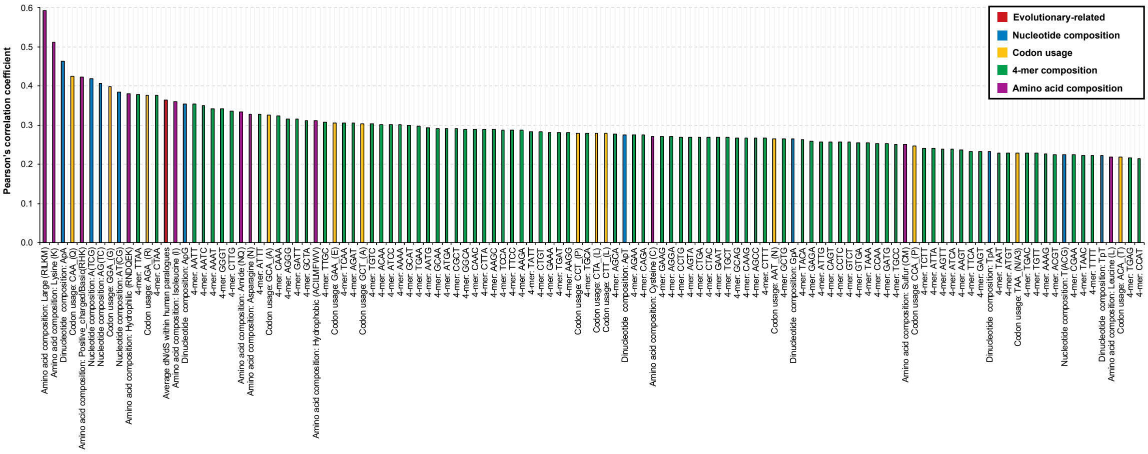


Fig 4.10 118 features positively associated with higher up-regulation after IFN-α treatments in human fibroblast cells (Student t-tests: p < 0.05). Detailed results about PCC and hypothesis tests are provided in **S4.3 Data**. Abbreviations: IFN, interferon; PCC, Pearson’s correlation coefficient; dN, non-synonymous substitutions per non-synonymous site; dS synonymous substitutions per synonymous site.

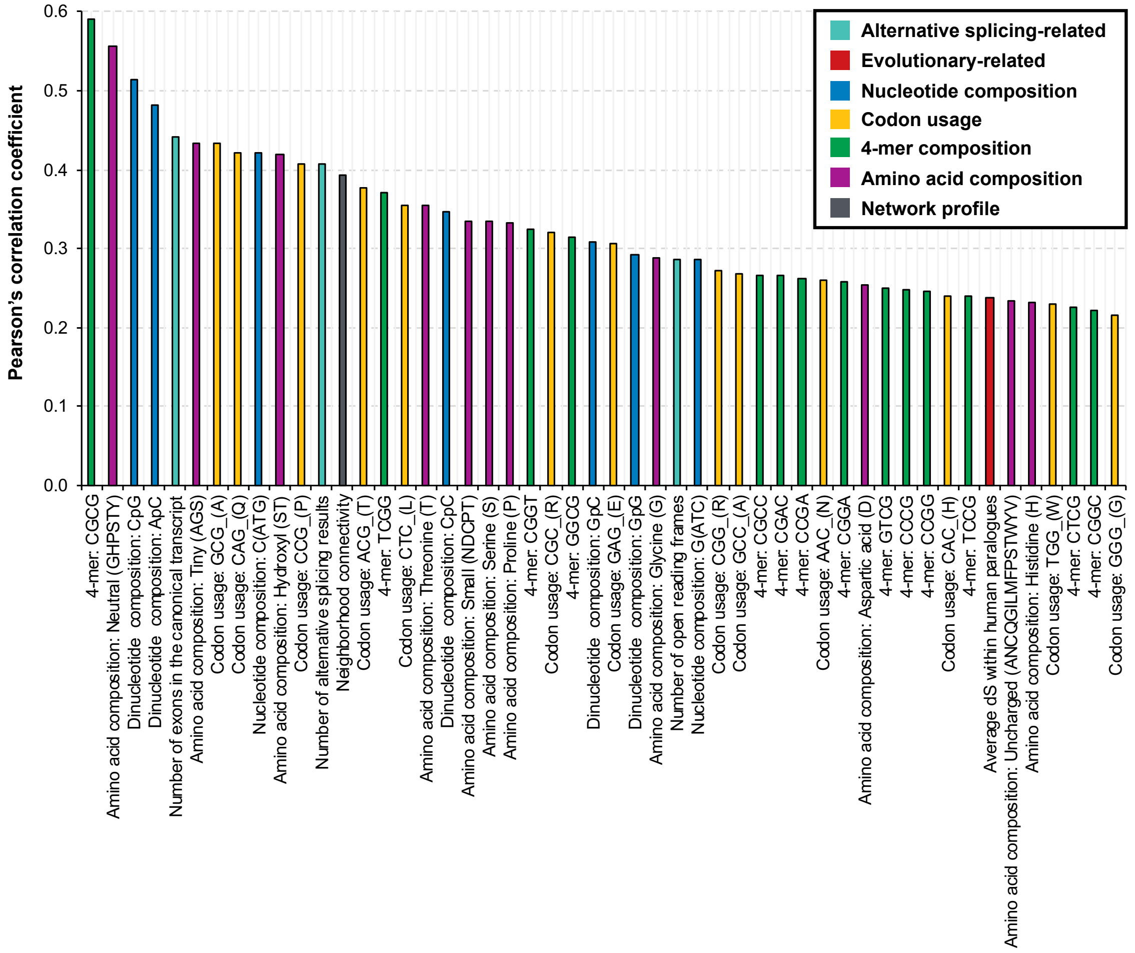


Fig 4.11 50 features negatively associated with higher up-regulation after IFN-α treatments in human fibroblast cells (Student t-tests: p < 0.05). Detailed results about PCC and hypothesis tests are provided in **S4.3 Data**. Abbreviations: IFN, interferon; PCC, Pearson's correlation coefficient; dS, synonymous substitutions per synonymous site.

4.4.6 Difference in feature representation of interferon-repressed genes and genes with low levels of expression

We grouped human genes into two classes based on their response to the IFN-α in human fibroblast cells. Genes significantly up-regulated in IFN-α experiments were included in the ISG class, while those that did not were put into the non-ISG class. However, there is also another group of human genes (IRGs) down-regulated in the presence of IFN-α. They were labelled as non-ISGs, but contain unique patterns that constitute an important aspect of the IFN response [315]. Some of these IRGs were not up-regulated in any known type I IFN systems, thus have been placed in a refined non-ISGs class for analyses and predictions. Additionally, there are a number of genes that have insufficient levels of expression in the experiments to determine a fold change. Here, we used the previously defined features to compare ISG from dataset S2 with IRGs from dataset S4 and ELGs from dataset S8 (**Table 4.1**).

As shown in **Fig 4.12,** IRGs are differentially represented to a lower extent in the majority of nucleotide 4-mer compositions than ISGs, which indicates the deficiency of some nucleotide sequence patterns in the coding region of IRGs. Note that, many nucleotide 4-mer composition features are more suppressed in ISGs than non-ISGs although the differences are small. The biased representation of these features in IRGs suggests that IRGs have characteristics similar to ISGs rather than non-ISGs. Additionally, there are a very limited number of features relating to evolutionary conservation, nucleotide compositions or codon usages showing obvious differences between ISGs and IRGs, but many of them are differentially represented when comparing ISGs with non-ISGs. Therefore, involving IRGs in the class of non-ISGs will increase the risk for machine learning models to produce more false positives. However, there are some informative features differentiating IRGs from ISGs. For example, comparing with ISGs, IRGs are more enriched in CpGs (Mann-Whitney U test: p = 5.6E-03), which is also mentioned in [371]. IRGs tend to have higher closeness centrality and neighbourhood connectivity than ISGs (Mann-Whitney U test: p = 0.04 and 6.4E-06 respectively), suggesting IRGs are closer to the centre of the human PPI network and connected to key proteins with many interaction partners. Differences in some amino acid composition features between ISGs and IRGs can also be observed in **Fig 4.12**. Therefore, good predictability is still expected when using features extracted from proteins sequences.

Collectively, **Fig 4.12** illustrates 161 features showing significant differences (Mann-Whitney U tests: p < 0.05) in the representation of ISGs and ELGs. An estimated 82% of these features were also differentially represented between ISGs and non-ISGs. 79% of these significant features displayed similar over-representation or under-representation in two comparisons, i.e., ISGs versus ELGs and ISGs versus non-ISGs. These ratios indicate that the majority of ELGs are less likely to be ISGs based on their feature profile as well as their low expression levels in cells induced with IFN-α. Network analyses showed that ELG products tended to have lower values of all calculated network features with the exception of topological coefficient than ISG products. It means that ELG products are less connected by other human proteins in the human PPI network. Particularly, their abnormal representation on the feature of average shortest paths indicating that some ELGs (e.g.

vascular cell adhesion molecule 1 (VCAM1) and ubiquitin D (UBD)) may still have high connectivity in the human PPI network.

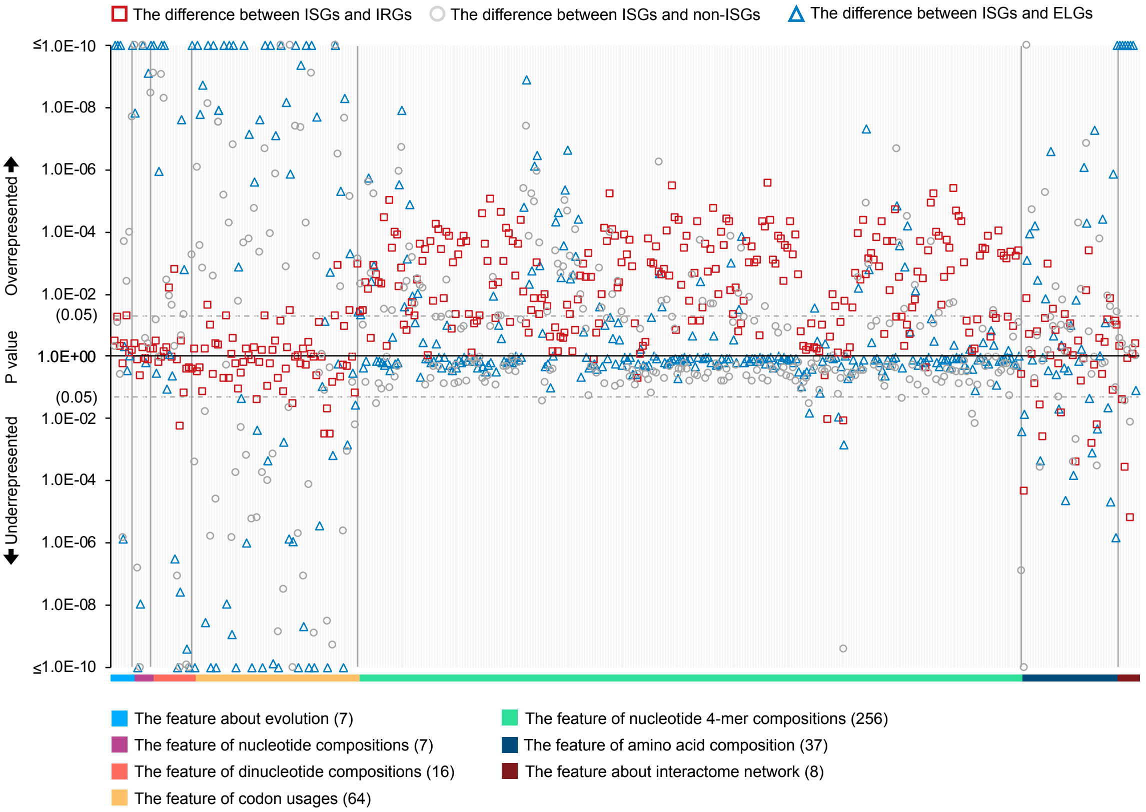


Fig 4.12 Differential expressions of discrete features between different genes and their coded proteins. Mann-Whitney U tests are applied for hypothesis testing and the results were provided in the **S2 Data**. Here, ISGs and non-ISGs are taken from dataset S2; IRGs and ELGs are taken from dataset S1 while the background human genes are from dataset S1 (**Table 4.1**). Abbreviations: ISGs, interferon-α-stimulated (up-regulated) genes; IRGs, interferon-α-repressed (down-regulated) genes; non-ISGs, human genes not significantly up-regulated by interferon-α; ELGs, expression-limited human genes in IFN-α experiments.

4.4.7 Implementation with machine learning framework

In this study, we encoded 397 discrete and 7046 categorical features for the analyses. As an excess of features will greatly increase the dimension of feature spaces and complicate the classification task for SVM [200], we limited the number of SLim_DNAs to the top 100 based on the adjusted p-value and we expected these to be sufficient to provide a picture of SLim patterns in the coding region of the canonical transcript. Accordingly, features measuring the co-occurrence status of multiple SLim_DNAs were recalculated based on the selected 100 SLim_DNAs. To reduce the impact of noisy data toward classifications, we only used the refined ISGs and non-ISGs from dataset S2 in machine learning.

Measured by SN, SP, MCC and AUC, the initial prediction results shown in **Table 4.4** indicate that proteome-based features, including those deciphered from protein sequences and the human interactome, perform much better than genome-based features presumably due to overfitting of the model [238]. Using discrete features that took the advantage of both genetic and proteomic aspects showed a good improvement in tests. The categorical features used in this study gave a binary statement for the occurrence of SLims in genetic and proteomic sequences but seemed not to perform well and disrupted the model when they were combined with discrete features. The results shown in the previous analyses also indicate that there are a considerable number of disruptive features hidden in the set (**Fig 4.5, Fig 4.7, and Fig 4.9**). The similar attributes of ISGs and IRGs (shown in **Fig 4.12**) led to lots of noisy data biasing the classifiers. This situation was not ameliorated and became more difficult when using other machine learning algorithms such as k-nearest neighbors (KNN), decision tree (DT), random forest (RF) (**Table 4.4**) [194,372]. As some genes respond to IFNs in a cell-specific manner [314], it is hard to produce predictions unless we detect key discriminating features, which are robust to the change of biological environment.

Considering these drawbacks, we designed an AUC-driven subtractive iteration algorithm (ASI) (**Fig 4.2**) to remove as many disruptive features as possible (**Fig 4.13A**). Pre-processing using the ASI algorithm showed that there were at least 28% of features disrupting the prediction model. They included 34% of features on codon usages and 50% of SLim features, thus, explaining the poor performance of the model trained with categorical features (**Table 4.4**). However, the loss of some of the individual nucleotide 4-mer feature seemed not to influence the performance of the classifier at this stage, but the similarities between IRGs and ISGs (**Fig 4.12**) particularly in these 4-mer features was a cause for concern when the model was used to predict new data especially unknown IRGs. When using the ASI algorithm, the number of disrupting features did not stabilise until the algorithm reached the 11-th iterations. The remaining 74 features constituted our optimum feature set for the prediction of ISGs (**Table 4.5**). Among them, 14 and 9 features displayed positive and negative correlations with the level of up-regulation in IFN-α experiments. During the procedure, the AUC kept increasing steadily and reached 0.7479 at the end. The MCC also showed an overall improvement although it fluctuated slightly during the last few iterations. By degressively ranking the score calculated by the prediction model, we found

68.1% of the 496 genes (equal to the number of ISGs in the training dataset) were successfully predicted as ISGs. **Fig 4.13B** illustrates the distribution of prediction scores generated by the ASI-optimised model for human genes with different expressions in IFN-α experiments. Human genes with higher up-regulation in IFN-α experiments tend to obtain higher prediction score from our optimised machine learning model (PCC = 0.243, p = 4.2E-10). However, there were also some ISGs incorrectly predicted by our model even though they were highly up-regulated, for example, basic leucine zipper ATF-like transcription factor 2 (BATF2, prediction score = 0.34). The model produced 33 ISGs with a prediction score higher than 0.8 but such figure for non-ISGs reduced to six, including one IRG (tripartite motif containing 59 (TRIM59)). The highest prediction score within non-ISGs was found on ubiquitin conjugating enzyme E2 R2 (UBE2R2, probability score = 0.88). It contains many features similar to ISGs but was not differentially expressed in the presence of IFN-α in human fibroblast cells [315]. The lowest prediction score within ISGs was found on cap methyltransferase 1 (CMTR1, probability score = 0.12) due to the weak signal from its features. For instance, CMTR1 protein does not contain any ISG-favoured SLim_AA listed in **Table 4.2**. The influence of IRGs on the prediction was reflected in the training dataset but was not significant. Compared with human genes not differentially expressed in the IFN-α experiments (non-ISGs but not IRGs), there were slightly more IRGs unsuccessfully classified when using a threshold of 0.549 (Pearson's chi-squared tests: M_1 = 27%, M_2 = 24%, $p > 0.05$).

Table 4.4 The performance of different feature combinations on the training dataset S2' via five-fold cross validation.

Method	Features	No.	Threshold-dependent					Threshold-independent	
			Score range	Threshold[a]	SN	SP	MCC	SN_496[b]	AUC
SVM	Genetic	452	0.359~0.623	0.402	0.769	0.355	0.169	0.579	0.6058
SVM	Proteomic	66	0.261~0.730	0.560	0.425	0.778	0.218	0.605	0.6360
SVM	Discrete	397	0.305~0.760	0.529	0.595	0.665	0.261	0.621	0.6573
SVM	Categorical	121	0.368~0.605	0.487	0.653	0.504	0.159	0.573	0.5736
SVM	All	518	0.328~0.743	0.542	0.567	0.681	0.250	0.615	0.6509
KNN[c]	All	518	0.100~0.900	0.500~0.550	0.593	0.621	0.214	0.607±0.014	0.6305
DT	Partial	182[d]	0 or 1	N/A	0.546	0.548	0.095	0.546	N/A
RF[e]	Random	Random	0.080~0.900	0.380~0.579	0.590±0.168	0.617±0.183	0.219±0.019	0.600±0.007	0.6413±0.0082
SVM	Optimum	74	0.098~0.918	0.549	0.623	0.750	0.376	0.681	0.7479

*a: this threshold is provided by maximum the value of MCC; b: this sensitivity is measured among tested genes with the top 496 prediction probabilities; c: k-value here is set as the square root of the size of the training samples in five-fold cross validation, i.e., k = 20 [373]; d:182 out of the 518 features (**S4.5 Data**) are used for decisions during this modelling procedure as the rest ones are not helpful to better split the dataset for lower system entropy [196]; e: this random forest algorithm uses 50 random grown trees and the modelling and* *validation procedures are repeated for 10 times.* Abbreviations: SVM, support vector machine; KNN, k-nearest neighbors; DT, decision tree; RF, random forest; SN, sensitivity; SP, specificity; MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve.

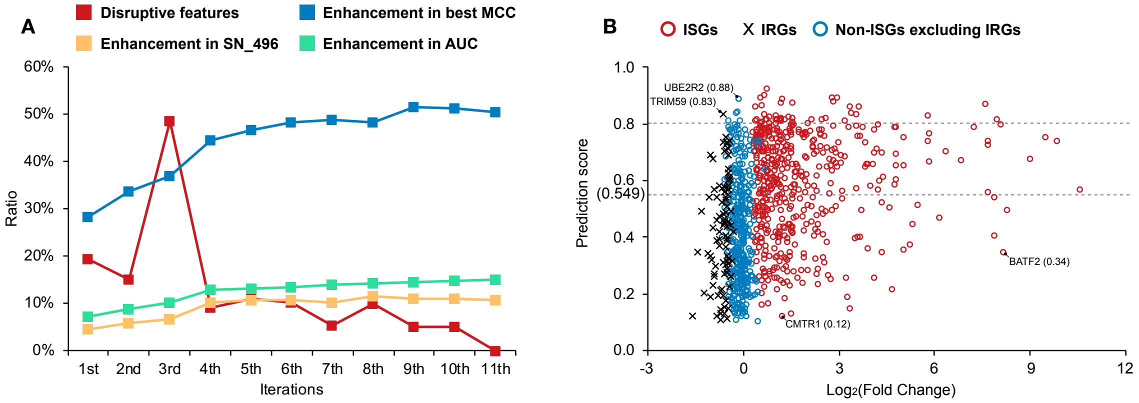


Fig 4.13 The optimisation on the machine learning model with the ASI algorithm. (A) shows the change of the prediction models based on the one generated with all 518 features (disruptive feature vector = 144, best MCC = 0.250, SN_496 = 0.615, and AUC = 0.6509). (B) shows the distribution of prediction scores generated by the ASI-optimised model for human genes with different expression levels in the IFN-α system. ISGs and non-ISGs shown in (B) are randomly selected with an undersampling strategy [265] on dataset S2. The list of gene names can be found in **S4.1 Data**. Abbreviations: SN, sensitivity; SN_496, sensitivity of predicted genes with the top 496 probability scores, MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve; ASI, AUC-driven subtractive iteration algorithm; IFN, interferon, ISGs, interferons-α-stimulated (up-regulated) human genes; IRGs, interferon-α-repressed (down-regulated) human genes; non-ISGs, human genes not significantly up-regulated by IFN-α; UBE2R2, ubiquitin conjugating enzyme E2 R2; TRIM59, tripartite motif containing 59; CMTR1, cap methyltransferase 1; BATF2, basic leucine zipper ATF-like transcription factor 2.

Table 4.5 The optimum 74 features contributing to the prediction of ISGs.

Evolutionary features (2)		
Number of human paralogues^{D}, average dS within human paralogues^{D-}.		
Codon usage features (10)		
Codon usage: CTA (L)$^{D+}$	Codon usage: ATT (I)^{D}	Codon usage: TAT (Y)^{D}
Codon usage: GCG (A)$^{D-}$	Codon usage: CAC (H)$^{D-}$	Codon usage: TGC (C)^{D}
Codon usage: CGT (R)^{D}	Codon usage: CGA (R)^{D}	Codon usage: CGG (R)$^{D-}$
Codon usage: AGA (R)$^{D+}$		
Genetic composition features (40)		
DNA AC content^{D}	Dinucleotide CpT composition^{D}	DNA 4-mer CGCG composition^{D-}
DNA 4-mer AATC composition^{D+}	DNA 4-mer TCGT composition^{D}	DNA 4-mer GATG composition^{D+}

DNA 4-mer AACA composition^{D}	DNA 4-mer TGAG composition^{D+}	DNA 4-mer GACC composition^{D}
DNA 4-mer ATAT composition^{D}	DNA 4-mer TGTA composition^{D}	DNA 4-mer GACG composition^{D}
DNA 4-mer ATGT composition^{D+}	DNA 4-mer CACG composition^{D}	DNA 4-mer GAGT composition^{D+}
DNA 4-mer ACAC composition^{D}	DNA 4-mer CTCC composition^{D}	DNA 4-mer GTAC composition^{D}
DNA 4-mer ACTA composition^{D}	DNA 4-mer CCAC composition^{D}	DNA 4-mer GTGT composition^{D}
DNA 4-mer ACTC composition^{D}	DNA 4-mer CCTA composition^{D}	DNA 4-mer GTGC composition^{D}
DNA 4-mer ACCG composition^{D}	DNA 4-mer CCTC composition^{D+}	DNA 4-mer GTGG composition^{D}
DNA 4-mer TATG composition^{D}	DNA 4-mer CCGT composition^{D}	DNA 4-mer GCAA composition^{D+}
DNA 4-mer TTCT composition^{D}	DNA 4-mer CGAG composition^{D}	DNA 4-mer GCTC composition^{D}
DNA 4-mer TTCG composition^{D}	DNA 4-mer CGTG composition^{D}	DNA 4-mer GCCT composition^{D}
DNA 4-mer TTGA composition^{D}	DNA 4-mer CGCA composition^{D}	DNA 4-mer GGGG composition^{D}
DNA 4-mer TCAT composition^{D}		
Proteomic composition features (9)		
Arginine composition^{D}, cysteine composition^{D+}, methionine composition^{D};		
Basic amino acid composition (R/H/K)$^{D+}$	Sulfur amino acid composition (C&M)$^{D+}$	
Hydroxyl amino acid composition (S&T)$^{D-}$	Small amino acid composition (N/D/C/P/T)$^{D-}$	
Large amino acid composition (R/I/L/K/M)$^{D+}$		
Uncharged amino acid composition (A/N/C/Q/G/I/L/M/F/P/S/T/W/Y/V)$^{D-}$		
Features about human interactome network (3)		
Shortest paths^{D+}, betweenness^{D}, neighborhood connectivity^{D-}.		
Motif features (8)		
SLim_DNA ATA[AG][TG]^{C}	SLim_DNA TAT[AT]T^{C}	SLim_DNA T[AT]AAA^{C}
SLim_DNA [ATG]TGTA^{C}	SLim_AA S[A-Z]N[A-Z]E^{C}	SLim_AA ENE^{C}
SLim_AA SVI^{C}	Co-occurence of SLim_AAs^{D}	

D: discrete features; C: categorical features; '+' symbol means features are positively associated with the level of up-regulation in IFN-α experiments ($p < 0.05$); '-' symbol means features are negatively associated with the level of up-regulation in IFN-α experiments ($p < 0.05$). Abbreviations: dS, synonymous substitutions per synonymous site; SLim_DNAs, short linear nucleotide motifs; SLim_AAs, short linear amino acid motifs.

4.4.8 Review of different testing datasets

In this study, we trained and optimised a SVM model from our training dataset S2', and prepared seven testing datasets (dataset S2''/S3/S4/S5/S6/S7/S8) to assess the generalisation capability of our model under different conditions (**Table 4.1**). The S2'' testing dataset was a subset of dataset S2. The prediction performance on this testing dataset was close to that in the training stage with an AUC of 0.7455 (**Fig 4.14A**). The best MCC value (0.345) was achieved when setting the judgement threshold to 0.438, which meant that the prediction model was sensitive to signals related to ISGs. In this case, it performed predictions with high sensitivity but inevitably produced many false positives, especially within IRGs.

In the S3 testing dataset, we used 695 ISGs with low confidence. The overall accuracy (equals to SN as there were no negatives) only reached 44.0% when using a judgement threshold of 0.549, about 0.18 lower than SN under the same threshold in the training dataset S2’ (**Table 4.4**). It is expected as some of their inherent attributes make them slightly up-regulated, silent or even repressed (e.g., become non-ISGs in other IFN systems) in response to some IFN-triggered signalling. On this testing dataset, our machine learning model produced 38 (5.5%) ISGs with a prediction score higher than 0.8. This number was also lower than that on the training dataset S2’. It further indicates the relatively low confidence for ISGs included in dataset S3.

The S4 testing dataset was constructed to illustrate our hypothesis that there are some patterns shared among ISGs and IRGs at least in the IFN-α system in human fibroblast cells. On this testing dataset, the prediction accuracy (equals to SP as there were no positives) was 60.2% under the judgement threshold of 0.549, about 0.15 lower than the SP under the same threshold in the training dataset S2’ (**Table 4.4**). Leucine rich repeat containing 2 (LRRC2), carbohydrate sulfotransferase 10 (CHST10) and eukaryotic translation elongation factor 1 epsilon 1 (EEF1E1) showed strong signals of being ISGs (probability score > 0.9). In total, there were 56 (5.6%) IRGs being incorrectly predicted as ISGs with probability scores higher than 0.8. This high score was found in an estimated 8.1% of ISGs but was only observed in 1.2% of human genes not differentially expressed in the IFN-α experiments (**Fig 4.13B**). These results indicate that there is a considerable number of IRGs incorrectly predicted as ISGs in the S4 testing dataset due to their close distance to the ISGs in the high-dimensional feature space. This may be the case for many other datasets including dataset S2’’, S5, S6, S7, and S8. It also supports our hypothesis about the shared patterns from the machine learning aspect and is consistent with the results shown in **Fig 4.12**.

The next three testing datasets (S5, S6, and S7) were collected from the Interferome database [334] to test the applicability of the machine learning model across different IFN types. The ISGs in these testing datasets were all highly up-regulated (Log_2(Fold Change) > 1.0) in the corresponding IFN systems while all the non-ISGs were not up-regulated after corresponding IFN treatments (Log_2(Fold Change) < 0). The results shown in **Fig 4.14** reveals that ISGs triggered by type I or III IFN signalling could still be predicted by our machine learning model, but the performance was limited to some extents (AUC = 0.6677 and 0.6754 respectively). However, it was almost impossible to make normal predictions with the current feature space for human genes up-regulated by type II IFNs (AUC = 0.5532).

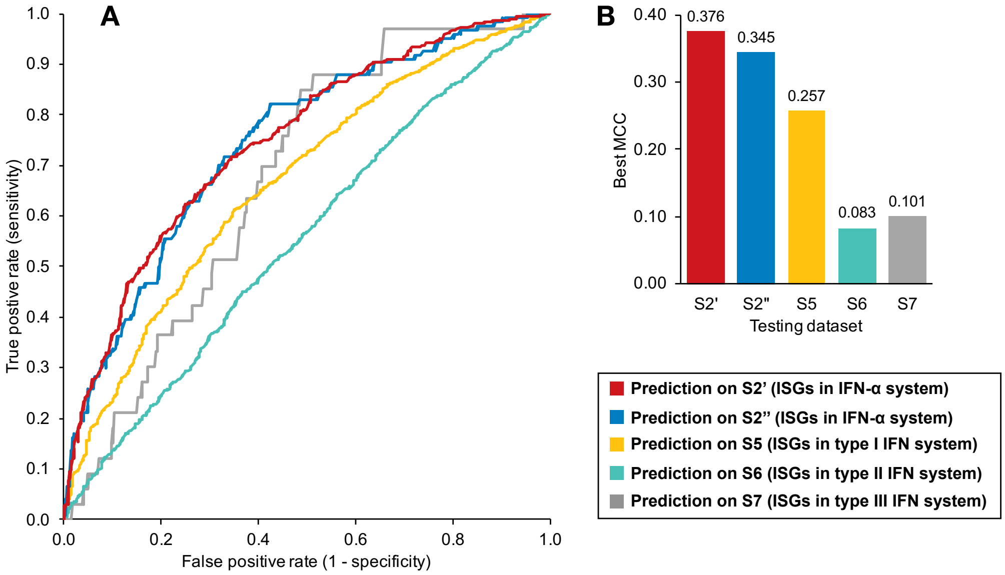


Fig 4.14 The performance of our optimised model on different datasets. (A) and (B) illustrate the AUC and best MCC. S2' is the training dataset used in this study. It randomly includes 496 ISGs and an equal number of non-ISGs from dataset S2 that contains ISGs/non-ISGs with high confidence (**Table 4.1**). Evaluation on this dataset in (A) is processed via five-fold cross validation. S2'' is the testing dataset constructed with the remaining human genes in dataset S2. S5, S6, and S7 are collected from the Interferome database [334], including human genes with different responses to the type I, II and III IFNs, respectively. The label and usage of these human genes are provided in **S4.1 Data**. Abbreviations: MCC, Matthews correlation coefficient; AUC, area under the receiver operating characteristic curve; ISGs, interferon-stimulated (up-regulated) human genes; non-ISGs, human genes not significantly up-regulated by interferons.

The S8 testing dataset consisted of 2217 human genes that were insufficiently expressed in IFN-α experiments in human fibroblast cells [315]. The results showed that there were around 41.2% ELGs being predicted as ISGs when using a judgement threshold of 0.549. This was approximately 0.21 lower than the SN under the same threshold in the training dataset S2' (**Table 4.4**). It suggests that there are more non-ISGs than ISGs in this dataset, which is consistent with the results shown in **Fig 4.12**. Particularly, we found 10 ELGs with prediction scores higher than 0.9: CD48 molecule, CD53 molecule, lipocalin 2 (LCN2), uncoupling protein 1 (UCP1), coiled-coil domain containing 68 (CCDC68), potassium calcium-activated channel subfamily M regulatory beta subunit 2 (KCNMB2), potassium voltage-gated channel interacting protein 4 (KCNIP4), zinc finger HIT-type containing 3 (ZNHIT3), serpin family B member 4 (SERPINB4), and fibrinogen silencer binding protein (FSBP). By retrieving data from the Genotype-Tissue Expression project

[374], we found that the expression of these ELGs were generally limited with the exception of CD53 and ZNHIT3 (**Fig 4.15**). The expression data of CD53 were not included in the OCISG database [315] and were also limited in the Interferome database [334]. It only showed slight up-regulation after type I IFN treatments in blood, liver, and brain but there is currently no record of its expression level in the presence of IFN-α in human fibroblast cells. ZNHIT3 is another well-expressed gene lacking information in the OCISG. In the Interferome databases [334], we found that ZNHIT3 could be up-regulated after IFN treatments in some fibroblast cells on skin. As for the remaining eight ELGs, despite their limited expression in human fibroblast cells, their features suggest that they are very likely to be IFN-α-stimulated in a currently untested cell type.

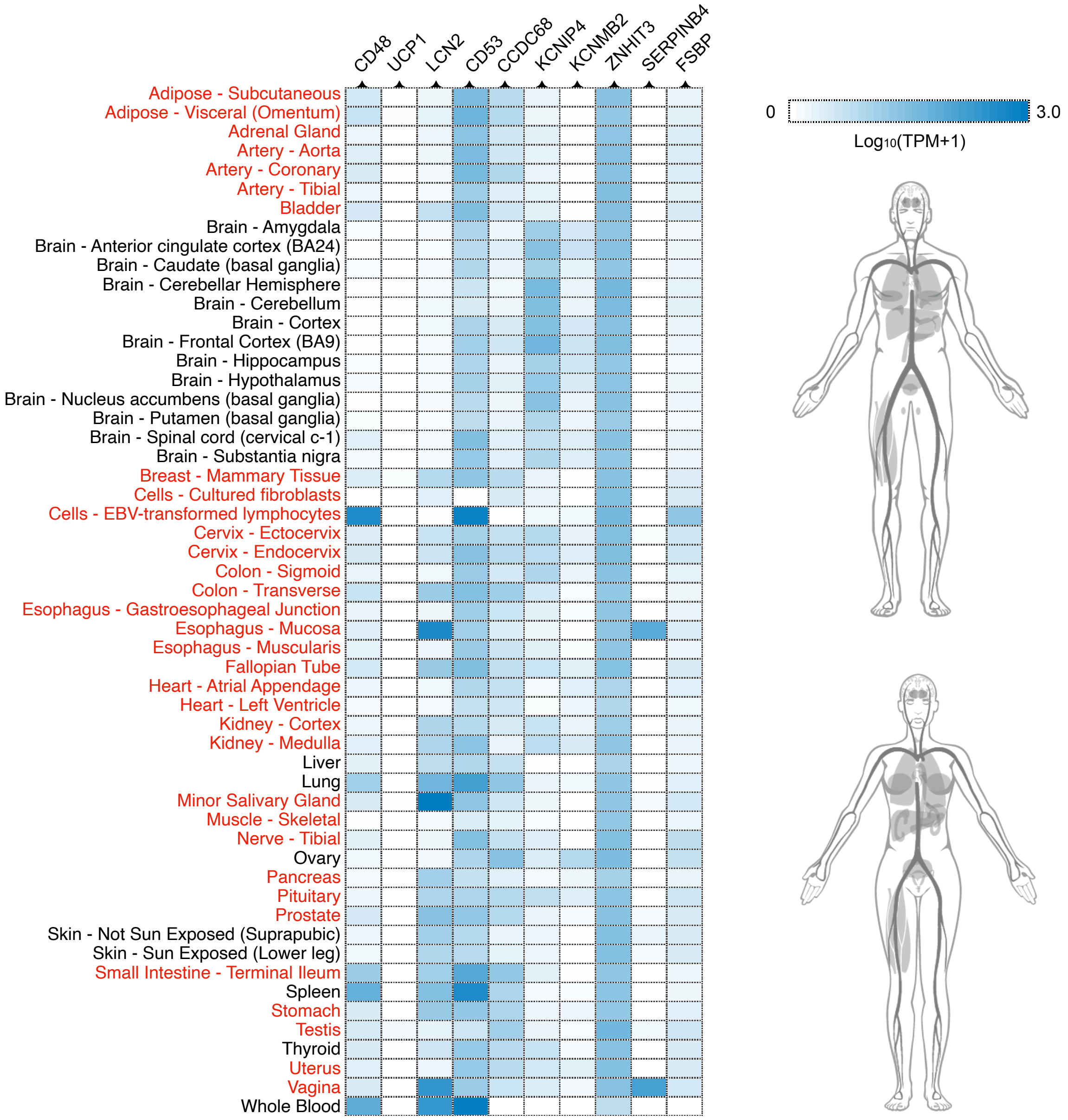


Fig 4.15 Expression of ELGs in different tissues. Expression data for ten ELGs are collected from the Genotype-Tissue Expression project (https://gtexportal.org/) [374]. The

tissues in red are not included in the Interferome database [334]. White boxes in the heatmap indicate that there is no data available for genes in the corresponding tissues. The overall expression level of these ten ELGs are reflected via human perspective photo retrieved from Expression Atlas (https://www.ebi.ac.uk/gxa) [375]. Abbreviations: ELGs, human genes with limited expression in interferon-α experiments; TPM, transcripts per million; BA, Brodmann area; EBV, Epstein-Barr virus; UCP1, uncoupling protein 1; LCN2, lipocalin 2; CCDC68, coiled-coil domain containing 68; KCNIP4, potassium voltage-gated channel interacting protein 4; KCNMB2, potassium calcium-activated channel subfamily M regulatory beta subunit 2; ZNHIT3, zinc finger HIT-type containing 3; SERPINB4, serpin family B member 4; FSBP, fibrinogen silencer binding protein.

4.5 Discussion

In this study, we investigated the characteristics that influence the expression of human genes in IFN-α experiments. We compared ISGs and non-ISGs through multiple procedures to guarantee strong signals for ISGs and to avoid cell-specific influences that resulted in the lack of ISGs expression in certain cell types [314]. Even some highly up-regulated ISGs can become down-regulated when the biological conditions change, exemplified by the performance of C-X-C motif chemokine ligand 10 (CXCL10) on liver biopsies after IFN-α treatment. This refinement is necessary as the representation of features between ISGs and the background human genes show that many non-ISGs especially IRGs have similar feature patterns to ISGs (**Fig 4.4-5.9**, **Fig 4.12**).

Generally, ISGs are less evolutionarily conserved with more human paralogues than non-ISGs. They have specific nucleotide patterns exemplified by the depletion of GC-content and have a unique codon usage preference in coding proteins. There are a number of SLim_DNAs widely observed in the cDNA of ISGs which are relatively rare in non-ISGs (**S4.4 Data**). Likewise, there are also many SLim_AAs highlighted in the sequences of ISG products that are absent or rare in non-ISGs (**Table 4.2**). In the human PPI network, ISG products tend to have higher betweenness than background human protein, indicating their more frequent interruption of the shortest path (geodesic distance) between different nodes. Abnormal expression or knockout of these proteins will increase the diameter of the network and may lead to some lethal consequences that are not tolerated in signalling pathways [376-378]. These ISG specific patterns may be the result of the evolution of the innate immune system in vertebrates and could be adaptations to the cellular environment induced by interferon following a pathogenic infection [379]. It is also possible that some of the particular SLim_DNAs and SLim_AAs may be important functionally as the cell changes from non-infected to infected. Experimental evidence will be necessary to investigate this.

Some inherent properties of ISGs facilitate or elevate their expression after IFN-α treatments but may also be used by viruses to escape from IFN-α-mediated antiviral response [104]. For instance, the representation of dN showed a more significant difference than that of dS within human paralogues. We found that higher dN/dS ratio was positively correlated with gene up-regulation following IFN-α treatments. It means the gene is less conserved with more non-synonymous or nonsense mutations, which can often be associated to inherited diseases and cancer [380]. It will also facilitate the virus to interfere with IFN-α signalling through the JAK-STAT pathway and inactivate downstream cellular factors involved in IFN-α signal transductions [104]. We found arginine was under-represented in ISG products compared to non-ISG products. As arginine is essential for the normal proliferation and maturation of human T cells [381], such depletion in ISG products may leave a risk of inhibiting T-cell function and potentially increased susceptibility to infections [382]. Furthermore, the special pattern of ISGs also promotes the representation of some features even if they are not well represented in nature, for example, the higher cysteine composition in ISGs. We hypothesize that it may be helpful to activate T-cell to regulate protein synthesis, proliferation and secretion of immunoregulatory cytokines [383,384]. There are also some features (e.g. methionine composition) not differentially represented between ISGs and non-ISGs but play important roles in IFN-α-mediated immune responses. There is evidence for the methionine content playing a role in the biosynthesis of S-Adenosylmethionine (SAM), which can improve interferon signalling in cell culture [385,386].

As previously mentioned, there were similar patterns between the feature representation of ISGs and IRGs, which led to the unclear boundary for ISGs and non-ISGs in the feature space. We found significant differences on the representation of features on evolutionary conservation (**Fig 4.4**) between ISGs and non-ISGs, but they became non-significant when comparing ISGs with IRGs. Similar phenomena were observed on many features deciphered from the canonical transcript, e.g., dinucleotide composition and codon usage features. We suggest that IRGs can be viewed as additional ISGs as they also regulate the activity of human genes in response to IFNs, only negatively. Furthermore, despite so many similarities between ISGs and IRGs, the separate classification of these genes is still possible. 4-mer compositions can be considered as the key features as most of them are differentially represented between ISGs and IRGs (**Fig 4.12**). Using proteomic features can also help to differentiate ISGs from IRGs but is not as good as using 4-mer features.

In the machine learning framework, we developed the ASI algorithm to remove disruptive features but kept features not influencing the prediction performance when being removed individually during iterations. Features might have synergistic effects thus the elimination of each feature left a different impact on the remaining ones even if these were individually useless for the improvement of the classifier. In this case, keeping as many useful features as possible seems to be a good option but will greatly increase the dimension of the feature space and increase the risk of overfitting [238]. By contrast, our ASI algorithm avoided such a risk and kept the synergistic effect of different features through iterations.

In the prediction task, we found some previously labelled non-ISGs with very high prediction scores, suggesting that they had many inherent properties enabling them to be stimulated after IFN-α treatments. Some of them, for example, ubiquitin conjugating enzyme E2 R2 (UBE2R2) has been shown to be significantly up-regulated after IFN-α treatment [387]. The non-ISG label was assigned because the relevant expression data in the presence of IFN-α were not included in the OCISG [315] and Interferome databases [334]. We also found ten ELGs with very high prediction scores (> 0.9). Literature searches on these genes indicate that they are likely to be involved in the innate immune response [388,389]. Their responses may be limited to certain tissues or cell types for which there is limited expression data in the Interferome database [334]. For example, LCN2 has been shown to mediate an innate immune response to bacterial infections by sequestering iron [388] and is induced in the central nervous system of mice infected with West Nile virus encephalitis [390]. CD48 was shown to increase in levels in the context of human IFN-α/β/γ stimulation [389]. Interestingly, CD48 is also the target of immune evasion by viruses [391] and has been captured in the genome of cytomegalovirus and undergone duplication [392]. Evidence for other ELGs is harder to assess, particularly those for which expression is absent in a range of tissues (e.g., UCP1 in **Fig 4.15**). UCP1 is a mitochondrial carrier protein expressed in brown adipose tissue (BAT) responsible for non-shivering thermogenesis [393]. It is possible that UCP1 is stimulated directly or indirectly by IFN-α in BAT, resulting in the defended elevation of body temperature in response to infection.

We developed the machine learning model based on experimental data from human fibroblast cells stimulated by IFN-α. It can be generalised to type I or III IFN systems, presumably because activations of type I and III ISGs are both controlled by ISRE [321] and aim to regulate host immune response [103,322,323]. However, our model cannot be used for predictions in the type II IFN system (AUC = 0.5532, best MCC = 0.083, **Fig 4.14**) because of the different control element and the different role in human immune activities [325].

In summary, our analyses highlight some key sequence-based features that are helpful to distinguish ISGs from non-ISGs or IRGs. Our machine learning model is able to produce a list of putative ISGs to support IFN-related research. As knowledge of ISG functions continue to be elucidated by experimentalists, the *in silico* approach applied here can in future be extended to classify the different functions of ISGs.

4.6 Summary

Interferons (IFNs) are signalling proteins secreted from host cells. IFN-triggered signalling activates the host immune system in response to intra-cellular infection. It results in the stimulation of many genes that have anti-pathogen roles in host defenses. Interferon-stimulated genes (ISGs) have unique properties that make them different from those not significantly up-regulated in response to IFNs (non-ISGs). We find the down-regulated interferon-repressed genes (IRGs) have some shared properties with ISGs. This increases the difficulty of distinguishing ISGs from non-ISGs. The use of machine learning is a sensible strategy to provide high throughput classifications of putative ISGs to support investigation with *in vivo* or *in vitro* experiments. Machine learning can also be applied to human genes for which there are insufficient expression levels before and after IFN treatments in various experiments. We expect that our study will provide new insight into better understanding the inherent characteristics of human genes that are related to response in the presence of IFN-α.

4.7 Supporting information

S4.1 Data. Basic information about human genes used in this study.

URL: http://isgpre.cvr.gla.ac.uk/data/S1.txt

S4.2 Data. The result of Mann-Whitney U tests for discrete features.

URL: http://isgpre.cvr.gla.ac.uk/data/S2_Data.txt

S4.3 Data. Association between feature representations and IFN-α stimulations.

URL: http://isgpre.cvr.gla.ac.uk/data/S3_Data.txt

S4.4 Data. The result of Pearson's chi-squared tests for sequence motifs.

URL: http://isgpre.cvr.gla.ac.uk/data/S4_Data.txt

S4.5 Data. Decision trees generated during five-cross validation on the training dataset S2'.

URL: http://isgpre.cvr.gla.ac.uk/data/S5_Data.txt

CHAPTER FIVE

5. SARS-COV-2 MIMICRY MEDIATED VIA HOST-LIKE SHORT LINEAR MOTIFS

5.1 Abstract

The ongoing coronavirus disease 2019 (COVID-19) pandemic represents a major threat to the public health. A major challenge in computational biology is to predict virus-host interactions from viral genome sequence data alone. This study focuses on the mimicry machinery that helps the virus to exploit the host system by "mimicry" of host-like short linear motifs (SLiMs). We test the hypothesis that the mimicked SLiMs are present in human proteins able to interact with the targets of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), i.e., the virus has converged on motifs that confer interaction capability with a specific host protein. Owing to functional constraints on the limits of the sequence divergence acting on these SLiMs, a convergent sequence similarity signal should be detectable in the virus genome. After screening SLiMs present in interactors of known SARS-CoV-2-host interacting proteins for disorder status, solvent exposure, subcellular localisation and lung expression, we obtain 18 high-confidence SLiMs mimicked by SARS-CoV-2. These virus SLiMs are present in envelope, nucleocapsid, open reading frame 7a (ORF7a), and non-structural proteins 8 (NSP8). They are well conserved across the diversity of SARS-CoV-2 variants including in the latest variant of concern, Omicron.

5.2 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a positive stranded RNA virus [5,394] that has triggered the most devastating pandemic in recent decades [395]. The uncontrolled spread of SARS-CoV-2 outside China has already accounted for more than 270 million infections worldwide (https://covid19.who.int/) and case numbers still show exponential growth due to the emerging SARS-CoV-2 variants of concern [396-405]. The SARS-CoV-2 genome encodes four structural proteins (spike, envelope, membrane and nucleocapsid), two polyproteins (open reading frame (ORF) 1a and 1b), and several accessory proteins (ORF3a, ORF6, ORF7a, ORF8, ORF10, etc.) [406-409]. Upon cell entry, the polyproteins are auto-proteolytically cleaved into 16 non-structural proteins (NSPs) [410].

The infection cycle of SARS-CoV-2 begins when its spike protein is primed by transmembrane serine protease 2 (TMPRSS2) [58] and ends with the release of viral progeny from the host cell [410]. During the infection, SARS-CoV-2 has to avoid the host immune defense sufficiently to ensure its replication and persistence in the host cell. For example, its membrane, ORF3a, ORF6, ORF7a, NSP1, NSP3, NSP6, NSP7, NSP12-14 antagonize host immune responses by suppressing the interferon signalling pathway [411,412]. Molecular mimicry of host-like short linear motifs (SLiMs) is a strategy adopted by the virus to interact with host molecules, necessary for efficient exploitation of the host system [356,413].

SLiMs are ubiquitous in human proteomes with more than 100,000 recorded instances [414]. They are predominantly detected in intrinsically disordered regions (IDRs) that lack stable tertiary structures [293,415,416]. They tend to have high binding affinities, be evolutionary plastic, i.e., tolerate high levels of sequence divergence while retaining function, and be ‘evolvable’ by being permissive to the addition of novel functionality by mediating novel interactions. This permits virus ‘mimicry’ of SLiMs by convergent evolution to commandeer the regulatory and signalling pathways of the host cell [4,153,417]. Additionally, their limited sequence length (usually less than ten amino acids) can lead to evolutionary degeneration of detectable sequence similarity (indicative of homology), where only physicochemical properties remain as evidence of shared function [4,417]. This simple mechanism of action is conducive not only to convergent evolution of SLiMs in human proteins without common evolutionary ancestry/homology, but also to convergent motifs in different viral species interacting with the same host proteins [418,419]. This convergent evolution can be readily achieved in viruses due to their high mutation rates [420,421].

Although mimicry motifs are expected to be prevalent in viruses, only a small fraction of host-like SLiMs have been clearly characterised in viruses. For example, SLiM ‘EHxY’ may be mimicked by virion protein 16 (VP16) of herpes simplex virus (HSV) to bind host cellular factor [422,423]. SLiM ‘TxV’ is orchestrated at the C terminus of human papillomaviruses (HPV) E6 to target PDZ-containing tumor suppressors for degradation [119]. SLiM ‘Nx[S/T]’ is utilised by envelope glycoproteins 1 (E1) of hepatitis C virus (HCV) upon partial deglycosylation with endoglycosidase H (Endo H) [424]. SLiM ‘PPxY’ are both mimicked by VP40 of Ebola virus (EBOV) and Marburg virus (MARV) to recruit the host NEDD4 E3 ubiquitin protein ligase (NEDD4) for budding [425]. SLiM 'IMxKN' may be involved in the mimicry of influenza A NS1 protein to counter the activation of protein kinase [426]. SLiM 'LYPxL' and 'LYPxxxLxxL' are thought to be mimicked by the *gag* protein of human immunodeficiency virus type 1 (HIV-1) to target programmed cell death 6 interacting protein (PDCD6IP) [427,428].

Evidence of a detectable mimicry strategy for SARS-CoV-2 is still lacking. Thus, investigating SARS-CoV-2 mimicry of host-like SLiMs in the same way as other viruses [356] use it to facilitate infection in the host [356], is of particular interest for a better understanding of virus-host interplay. Here, we propose a computational approach to systematically screen host-like SLiMs that are mimicked by SARS-CoV-2 in order to gain interaction affinity with its targets (**Fig 5.1**). While not important for their detection, we assume that the mimicked SLiMs are the result of convergent evolution but not the acquisition of host sequence [418]. Some human proteins are interactors of SARS-CoV-2-host interacting proteins; thus, they are considered potentially mimicked human proteins (PMHPs). We hypothesise that a SLiM is unlikely to arise stochastically in multiple PMHPs that share a common virus target as their interaction partners within the human protein-protein interaction (PPI) network. Its presence in multiple PMHPs would at least infer functionality and support the involvement of the SLiM in the virus mimicry especially when its enrichment in PMHPs reaches a significant level [429]. We also suggest that along with the location of the SLiM in IDRs, high solvent exposure, appropriate subcellular localisation, and expression in SARS-CoV-2-infected tissues are all important requirements to permit functionality of SLiMs and their potential for virus mimicry. By using the latter data mining strategies, we find evidence of 18 mimicking SLiMs in SARS-CoV-2 envelope, nucleocapsid, ORF7a, and NSP8.

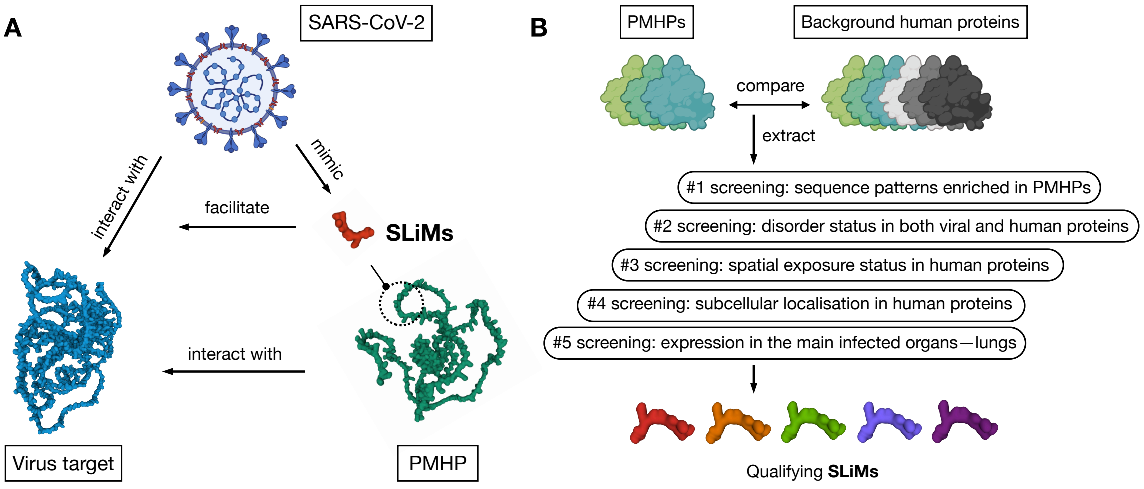

A
SARS-CoV-2
interact with
mimic
facilitate
SLiMs
interact with
Virus target
PMHP
B
PMHPs
Background human proteins
compare
extract
#1 screening: sequence patterns enriched in PMHPs
#2 screening: disorder status in both viral and human proteins
#3 screening: spatial exposure status in human proteins
#4 screening: subcellular localisation in human proteins
#5 screening: expression in the main infected organs—lungs
Qualifying SLiMs

Fig 5.1 Schematic of the principle and screening procedures used in this study. As shown in (B), to be identified as virus-mimicry, SLiMs needed to satisfy requirements about statistical enrichment, disorder, solvent exposure, subcellular localisation, and tissue expression. The example components in (A) are from BioRender (https://biorender.com/) and AlphaFold [430] predicted structures of RBMX (coloured in green), ZC3H18 (coloured in blue), and SLiM 'NxxxSRxP' (coloured in red) under the Creative Commons licence 4.0. Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; PMHP, potentially mimicked human protein; SLiM, short linear motif; RBMX, RNA binding motif protein X-linked; ZC3H18, zinc finger CCCH-type containing 18.

5.3 Methods

5.3.1 SARS-CoV-2 baseline, variants and coronaviridae members

We used SARS-CoV-2 sequence hCoV-19/USA/WA1/2020, accession MN985325 [431] (Pango lineage: A [432]), as the baseline sequence in this study. We retrieved the sequence of ten viral proteins of this isolate from NCBI Virus Variation Resource (ver. December 2021) [149] and GISAID database (ver. December 2021) [148] (**S5.1 Data**). They include four structural proteins (spike, envelope, membrane, and nucleocapsid), five accessory proteins (ORF3a, ORF6, ORF7a, ORF8, and ORF10), and one polyprotein (ORF1ab) [408] that can be auto-proteolytically processed into 16 NSPs [410]. Data on NSP11 and other ORFs were not considered in this study due to insufficient evidence supporting them and unclear annotation in the accession MN985325 [148,149,410,433]. In addition to MN985325, we also introduced six human coronaviruses (HCoVs) [434] and ten SARS-CoV-2 variants [432,435] to investigate the evolutionary conservation of potentially mimicked SLiMs (**Table 5.1**). Genome sequences of the aforementioned seven human coronaviruses and ten SARS-CoV-2 variants are provided in **S5.2 Data**.

Table 5.1 Representative of human coronaviruses and SARS-CoV-2 variants.

Coronavirus	Subfamily	Accession	Length	Ref.	Variant	Lineage[a]	Accession	Ref.
HoV-229E	Duvinacovirus	NC_002645	27317	[436]	Alpha	B.1.1.7	MZ453433	[396,397,400,437]
HCoV-NL63	Setracovirus	NC_005831	27553	[438]	Beta	B.1.351	MZ202314	[399-401,437]
HCoV-OC43	Embecovirus	NC_006213	30741	[439]	Gamma	P.1	MZ169910	[397,400,401,437]
HCoV-HKU1	Embecovirus	NC_006577	29926	[440]	Delta	B.1.617.2	OL456172	[398,401,437]
MERS-CoV	Merbecovirus	NC_019843	30119	[441]	Eta	B.1.525	OL601550	[400,401]
SARS-CoV	Sarbecovirus	KY352407	29274	[442]	Iota	B.1.526	OL615111	[400]
SARS-CoV-2	Sarbecovirus	MN985325	29882	[431]	Kappa	B.1.617.1	MZ562746	[401,443]
					Lambda	C.37	OL622097	[402,444]
					Mu	B.1.621	OK349712	[403]

Omicron	B.1.1.529	OL672836	[404]

a: lineages are defined though sequence alignment from the Pangolin project [432]. Abbreviations: SARS-CoV, severe acute respiratory syndrome coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; HCoV, human coronavirus; MERS-CoV, middle east respiratory syndrome coronavirus.

5.3.2 Interactions among SARS-CoV-2 and human proteins

We retrieved 20376 human proteins from the Swiss-Prot section of UniProtKB (ver. December 2021) [2]. We used AlphaFold to predict the tertiary structure of UniProt sequences with a length ranging from 16 to 2700 [430]. Consequently, we obtained 16,185 human proteins with a predicted tertiary structure (**S5.3 Data**). They constitute our background dataset S1. Based on Gordon *et al.*'s affinity-purification–mass spectrometry analysis [407,445,446], we recompiled high-confident physical interactions among 22 SARS-CoV-2 and 291 human proteins (**S5.4 Data**) as follows: information on NSP3 and NSP16 were not included in this study as their virus-host interactions were not available in Gordon *et al.* [407]; virus-host PPIs involving NSP11, NSP5 mutant, ORF3b, ORF7b, ORF9b, and ORF9c were removed as these viral proteins were not annotated in MN985325 [148,149,410,433]. By retrieving experimentally verified human PPIs from HIPPIE 2.2 (medium confidence, HIPPIE evidence score > 0.63) [161], we generated a network for 6727 human proteins and 291 SARS-CoV-2 targets. In this study, we considered that these 6727 human proteins are potentially mimicked by one or more SARS-CoV-2 proteins (**Table 5.2**).

Table 5.2 A breakdown of SARS-CoV-2-host interactions and PMHPs obtained from human PPIs.

Viral protein	Virus-host PPIs^a	PMHPs^b	Viral protein	Virus-host PPIs^a	PMHPs^b
Spike	2	19	NSP4	8	214
Envelope	6	379	NSP5	1	337
Membrane	30	1480	NSP6	4	63
Nucleocapsid	15	2142	NSP7	32	1012
ORF3a	8	182	NSP8	24	1446
ORF6	3	295	NSP9	16	554
ORF7a	2	93	NSP10	5	336
ORF8	47	1157	NSP12	20	607
ORF10	9	832	NSP13	40	1868
NSP1	6	190	NSP14	3	214
NSP2	7	373	NSP15	3	229

a: 291 exogenous SARS-CoV-2-host interactions are provided in ***S5.4 Data****; b: human interactors of the 291 SARS-CoV-2 targets are listed in* ***S5.5 Data****.* Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; PMHPs, potentially mimicked human proteins; ORF, open reading frame; NSP, non-structural protein.

5.3.3 Short linear motifs and their enrichment in PMHPs

SLiMs can be detected in different proteins but still show the same specificity to mediate distinct cellular processes such as targeting the protein to specific subcellular compartments and acting as binding or recognition sites in multiple signalling pathways [153,429]. Viruses may mimic host-like SLiMs to mediate the interactions with human host proteins and then radically alter the regulatory processes of a cell [4,356]. These SLiMs are usually composed of less than ten amino acid residues, of which only two or three are important for the function [4,153,356]. In this study, their lengths were restricted between two and eight. They are formulated as:

$$\begin{gathered}
[R_1][R_2] \\
[R_1][r_1][R_2] \\
[R_1][r_1][r_2][R_2] \\
\dots \\
[R_1][r_1] \dots [r_6][R_2]
\end{gathered} \tag{1}$$

where, R_n and r_n indicate a defined residue and the position occupied by a residue that is either defined or uncertain. The number of alternatives increases exponentially when including more defined residues in the SLiM pattern. For simplicity, we ignored those patterns not detected in the SARS-CoV-2 sequence (MN985325) (**S5.1 Data**). Since SLiMs are usually associated with convergent evolution [417,418], we adopted Pearson's chi-squared tests to assess the occurrence difference of SLiMs in the corresponding PMHPs (**Table 5.2**) and background dataset S1. SLiMs arose in less than five PMHPs were ignored. Benjamini-Hochberg correction procedure was applied to avoid type-I error [264]. The cut-off threshold of adjusted p-value was set as 0.01 initially and then changed to 0.05 for the final estimation. These procedures were repeated for 22 SARS-CoV-2 proteins to discover SLiMs more likely to be mimicked by each SARS-CoV-2 protein.

5.3.4 Filtering SLiMs

There are many important requirements for SLiMs to become a functional module in the protein [429]. In this study, we applied four kinds of filters on the identified SLiMs: disorder, solvent exposure, subcellular localisation, and tissue expression. Intrinsically disordered regions (IDRs) are important components of proteins. They generally lack bulky hydrophobic amino acids to form well-organized tertiary structures but can still facilitate regulation and recruit diverse binding partners [447], which may be favoured by SARS-CoV-2. In this study, IDRs of human and viral proteins were all estimated by ESpritz with short X-ray flavour [266]. It was developed based on bidirectional recursive neural networks.

Exposure and solvent accessibility measures to what extent an amino acid is accessible to the solvent surrounding the protein. Here, we used AlphaFold [430] predicted tertiary structure to calculate the accessible surface area (ASA) of amino acids with PSAIA programme [448]. Position with higher ASA values were more exposed or less buried. We hypothesise that SLiMs involving deeply buried amino acid residues (ASA = 0 square Ångstroms) are less likely to be mimicked by any SARS-CoV-2 proteins.

Subcellular localisation describes the cellular environment of proteins. This information can be used to improve reliability of virus mimicry mechanism. As SARS-CoV-2 proteins are either cytoplasmic (spike, membrane, nucleocapsid, ORF3a/6/7a/8/10, and NSP2/4/8) or both nuclear and cytoplasmic (envelope and NSP1/5/6/7/9/10/12/13/14/15) [449], SLiMs from PMHPs that are located at the same subcellular localisation are more likely to be mimicked. Here, subcellular localisation of human proteins was retrieved from UniProt [2], Reactome [150], human protein atlas (HPA) [450] and Ensembl databases [244].

Tissue expression data indicates molecular phenotypes across a wide range of tissue types [152]. Measured by the transcript per million (TPM), the expression level of a gene in the specific tissue can be catalogued into four class: high (TPM > 1000), medium (TPM between 10 to 1000), low (TPM between 0.5 to 10) and limited (TPM < 0.5) [375]. It has been suggested that SARS-CoV-2 mainly infect basal, ciliated, club, AT2, and proliferative KRT7+ epithelial cells in the lung tissue [451]. In this study, we mainly focused on SARS-CoV-2 mimicry that could take place in the lung tissue. Expression data in lung were retrieved from 578 sample tests in the Genotype-Tissue Expression (GTEx) project [152]. We hypotheses that genes with a medium TPM lower than 0.5 in lung are less likely to be involved in SARS-CoV-2 mimicry.

5.4 Results

5.4.1 PMHPs with high connectivity to the virus targets

Some human proteins are highly connected in human PPI network. For example, among 6727 PMHPs, estrogen receptor 2 (ESR2) was ranked first for interacting with the largest number of SARS-CoV-2 targets (n=100, **S5.5 Data**). Its high connectivity still remained when only focusing on human proteins targeted by SARS-CoV-2 envelope (**Fig 5.2A**), NSP8, or NSP12 (n=4, 16 and 12 respectively).

Exemplified by the PPI between SARS-CoV-2 spike protein and angiotensin converting enzyme 2 (ACE2) receptor [58], each viral protein has to target some specific human proteins to hijack and exploit the host system [407,410,452]. Mimicry mechanisms of different SARS-CoV-2 proteins need to be investigated individually due to the difference in their interacting targets (**S5.4 Data**). At this stage, we temporarily filtered out human proteins connected to less than 50% of the SARS-CoV-2 targets. With respect to human interactors shared among the targets of SARS-CoV-2 spike protein, ring finger protein 4 (RNF4), hepatitis A virus cellular receptor 2 (HAVCR2), transmembrane protein with EGF like and two follistatin like domains 1 (TMEFF1), and claudin domain containing 1 (CLDND1) were remarkable for connecting all spike-interacting proteins in the collected data (**Fig 5.2B**). Comparing with the cases for SARS-CoV-2 spike or envelope proteins, we found more human proteins that might be involved in the mimicry of SARS-CoV-2 nucleocapsid phosphoprotein (**Fig 5.2C**). However, according to the endogenous human PPIs, no human protein was highly connected ($n \geq 15$) to the targets of SARS-CoV-2 membrane protein.

As for the accessory proteins of SARS-CoV-2, RNF4 and COP9 signalosome subunit 6 (COPS6) might be potential targets of ORF3a and ORF10 mimicry, respectively. Ubiquitin specific peptidase 7 (USP7) might be mimicked by ORF6. Seven human proteins were noteworthy for ORF7a mimicry. The potential of SARS-CoV-2 mimicry was also detected for NSPs with the exception of NSP3, NSP4, NSP5, NSP9, NSP11, NSP13, and NSP16. As a result, in addition to the virus targets, another 28 human proteins were screened as ‘hub’ proteins in corresponding sub-networks of human PPIs. Neurotrophic receptor tyrosine kinase 1 (NTRK1), ESR2, egl-9 family hypoxia inducible factor 3 (EGLN3), RNF4, elongation factor Tu GTP binding domain containing 2 (EFTUD2), small ubiquitin like modifier 2 (SUMO2), and interferon gamma inducible protein 16 (IFI16) required further investigation for their frequent appearance in the results of this analysis (**Table 5.3**).

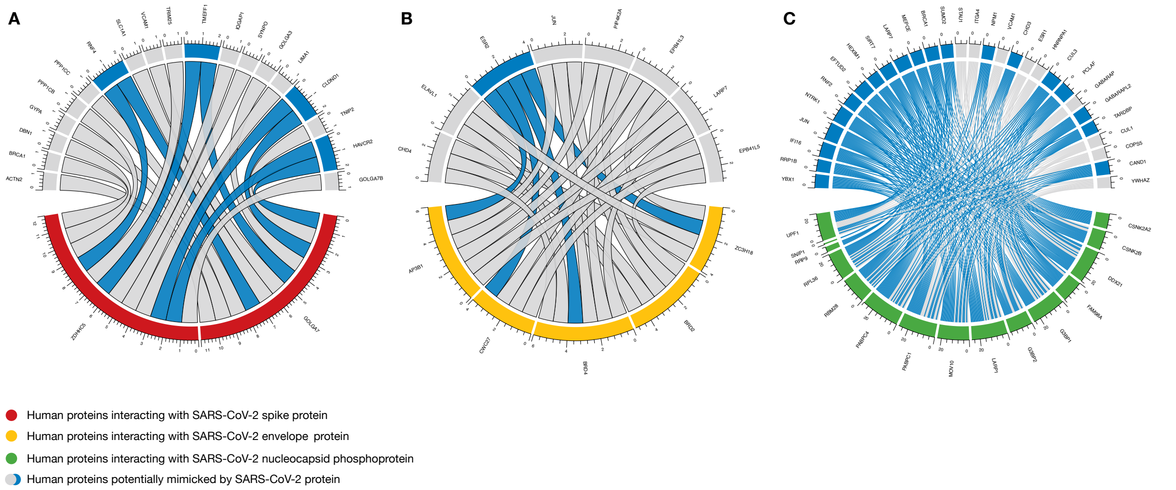


Fig 5.2 Sub-networks highlighting human proteins potentially mimicked by some structural proteins of SARS-CoV-2. Plot (A) was delineated for human some host targets of SARS-CoV-2 spike proteins while plot (B) and (C) were created for those of SARS-CoV-2 envelope protein and nucleocapsid phosphoprotein, respectively. Red, yellow and green blocks represent human proteins targeted by spike, envelope, and nucleocapsid proteins of SARS-CoV-2, respectively. Grey blocks represented human proteins interacting with 50%~60% of specific virus targets while blue blocks indicate a more connected status (>60%). More detailed endogenous interactions between PMHPs and SARS-CoV-2 targets are provided in **S5.5 Data**. Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Finally, at this stage, we detected 84 human proteins more connected to the targets of SARS-CoV-2 proteins (**Table 5.3**). They were 'hub' proteins of the corresponding human PPI sub-network and occurred in the majority of signalling pathways interfered or hijacked by the virus [429]. Mimicking patterns from these human proteins might enhance the affinity of multiple SARS-CoV-2-host PPIs and further facilitate their exploitation of the host system [356]. However, according to the current results, we still lacked convincing evidences to support the mimicry activity. It was also unclear how the putative mimicry was processed.

Table 5.3 PMHPs highly connected to SARS-CoV-2 targets.

Viral protein	PMHP (No. of interactions with virus targets/No. of virus targets)[a, c]
Spike	**RNF4**[b] (2/2), HAVCR2 (2/2), TMEFF1 (2/2), CLDND1 (2/2)
Envelope	**ESR2**[b] (4/6), **JUN**[b] (3/6), **LARP7**[b] (3/6), CHD4 (3/6), **ELAVL1**[b] (3/6), EPB41L3 (3/6), PIP4K2A (3/6), EPB41L5 (3/6)
Nucleocapsid	**JUN**[b] (13/15), **NTRK1**[b] (12/15), CUL3 (11/15), **LARP7**[b] (11/15), **EFTUD2**[b] (11/15), **IFI16**[b] (11/15), PCLAF (11/15), YBX1 (10/15), BRCA1 (10/15), RNF2 (10/15), CAND1 (10/15), SIRT7 (10/15), GABARAPL2 (10/15), **SUMO2**[b] (9/15),

	CHD3 (9/15), NPM1 (9/15), MEPCE (9/15), TARDBP (9/15), HEXIM1 (9/15), RRP1B (9/15), ESR1 (8/15), YWHAZ (8/15), HNRNPA1 (8/15), **COPS5**[b] (8/15), CUL1 (8/15), VCAM1 (8/15), ITGA4 (8/15), STAU1 (8/15), GABARAP (8/15)
ORF3a	**RNF4**[b] (5/8)
ORF6	USP7 (3/3), **ESR2**[b] (2/3), **IFI16**[b] (2/3), **NXF1**[b] (2/3), CUL7 (2/3), OBSL1 (2/3), CTNNB1 (2/3), CDC37 (2/3), FAF1 (2/3), **EGLN3**[b] (2/3), HNRNPUL1 (2/3), KPNB1 (2/3), NUP153 (2/3), NUMA1 (2/3), NUP107 (2/3), NKX2-1 (2/3), SEH1L (2/3), PTTG1 (2/3)
ORF7a	HSCB (2/2), **SUMO2**[b] (2/2), MYC (2/2), CD70 (2/2), RNF123 (2/2), SCN2B (2/2), PTP4A3 (2/2)
ORF10	COPS6 (6/9), NEDD8 (5/9), UBE2M (5/9), **COPS5**[b] (5/9), ARIH1 (5/9)
NSP1	CHM (4/6), RAE1 (4/6), CIAO1 (4/6), MMS19 (4/6), STN1 (4/6), **NTRK1**[b] (3/6), **EFTUD2**[b] (3/6), RPA2 (3/6), **SUMO2**[b] (3/6), **RPA1**[b] (3/6), RNF31 (3/6), POLA2 (3/6)
NSP2	**NTRK1**[b] (4/7), XPO1 (4/7)
NSP6	CFTR (2/4), **NXF1**[b] (2/4), **RNF4**[b] (2/4), CLN3 (2/4)
NSP7	**RNF4**[b] (19/32)
NSP8	**ESR2**[b] (16/24), **EFTUD2**[b] (12/24), **IFI16**[b] (12/24)
NSP10	**NTRK1**[b] (3/5), TRIM25 (3/5), **RPA1**[b] (3/5), RPA3 (3/5)
NSP12	**ESR2**[b] (12/20), **EGLN3**[b] (11/20)
NSP14	**EGLN3**[b] (2/3), FBXO6 (2/3), TEPSIN (2/3)
NSP15	**ELAVL1**[b] (2/3), HLA-B (2/3), CACYBP (2/3), UBE2D3 (2/3)

*a: detailed PPIs between PMHPs and virus targets are provided in **S5.5 Data**; b: bold PMHPs are highly connected to the targets of multiple SARS-CoV-2 proteins; c: number of virus targets are listed in **Table 5.2**.*
Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; PMHP, potentially mimicked human protein; ORF, open reading frame; NSP, non-structural protein; PPI, protein-protein interactions.

5.4.2 SLiMs significantly enriched in PMHPs

SLiMs can be considered as protein functional modules that play an important role in numerous cellular processes such as post-translational modification [453] and protein-protein interactions [454]. As a result of convergent evolution, they can be detected in different proteins and keep their specificity to interact with the target biomolecules [418,429].

Here, by comparing with the background dataset S1, we initially discovered 7957 distinct host-like SLiMs significantly enriched in PMHPs (adjusted P<0.01, **S5.6 Data**). Some of these SLiMs might not be relevant to SARS-CoV-2 mimicry but were less likely to be an incidental result of genetic coding. Amino acid compositions of PMHPs indicated an enrichment of acidic, negatively charged or hydrophilic residues (**Fig 5.3**). Such propensity inevitably influenced the sequence pattern of SLiMs enriched in PMHPs. For example, we found that arginine (R) and lysine (K) were over-represented in PMHPs of SARS-CoV-2 envelope. These biased amino acid compositions led to the R/K enrichment in the conserved positions of the corresponding SLiMs. Despite this, we also noticed the impact of disorder-promoting amino acid residues on the enriched SLiMs (**Fig 5.3**). We found that some amino acid residues in the non-conserved or active positions of SLiMs were not placed randomly. Exemplified by SLiM ‘ENxxxxLD’, it was detected for 14 times in PMHPs of SARS-CoV-2 ORF6 (**Table 5.4**). This motif contained four non-conserved positions. Its second and last non-conserved positions tended to be occupied by R. However, this pattern was not observed frequently enough to get a significant signal (Pearson's chi-squared test: P>0.05) warranting a more complex SLiM such as ‘ENxRxxLD’, ‘ENxxxRLD’ or ‘ENxRxRLD’.

Table 5.4 shows 22 representative SLiMs and their representations in the corresponding SARS-CoV-2 proteins. Some SLiMs (e.g., ‘CxFQ’ and ‘AFVxF’) were not located in the IDR of the viral protein. The conserved part of these SLiMs were not globally enriched in disorder-promoting residues, i.e., proline (P), glutamic acid (E), serine (S), glutamine (Q), and K [415,416]. Additionally, viral sequence regions that contained these SLiMs were more enriched with order-promoting residues including cysteine (C), tryptophan (W), isoleucine (I), tyrosine (Y), phenylalanine (F), leucine (L), histidine (H), valine (V), asparagine (N), and methionine (M) [415,416]. Hence, these SLiMs were unlikely to be evolutionarily convergent in viral proteins, which in turn might not promote the virus invasion into the network of interacting host proteins [356]. In view of this, they were less likely to be mimicked. Since similar cases were widely observed when matching our detected SLiMs to the viral sequence, further screening was required to discern SLiMs truly involved in the mimicry machinery.

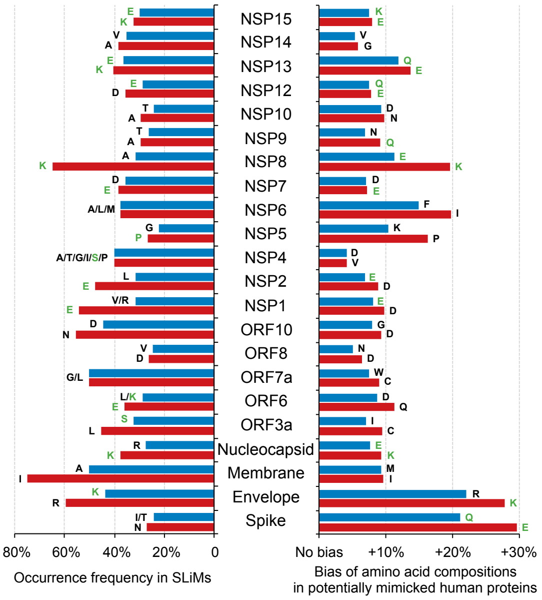


Fig 5.3 Over-representation of some amino acid residues in SLiMs enriched in PMHPs. The red and blue bar indicated the most and second most obvious bias of amino acid compositions in the SLiM set (left) when comparing with their representation in the background dataset S1 (right). Disorder-promoting residues (P, E, S, Q, K) are coloured in green. Abbreviations: SLiM, short linear motif; PMHP, potentially mimicked human protein; ORF, open reading frame; NSP, non-structural protein.

Table 5.4 Representative SLiMs and their representation in SARS-CoV-2 proteins.

Viral protein	SLiM[a,b]	In virus[b]	P-value[c]	Viral protein	SLiM[a,b]	In virus[b]	P-value[c]
Spike	CxFQ	CEFQ	3.0E-18	NSP4	RxxxxNGV	RRVVFNGV	1.4E-04
Envelope	AFVxF	AFVVF	1.0E-14	NSP5	DE	DE	6.9E-09
Membrane	VxxAxxxI	VLAAVYRI	3.0E-04	NSP6	AxxxMF	AFAMMF	2.6E-03
Nucleocapsid	DxxxK	DGKMK DPNFK DKKKK DDFSK	1.5E-24	NSP7	LxKxxxE	LAKDTTE	3.8E-04
ORF3a	CxHTxC	CWHTNC	3.4E-12	NSP8	KxxxxxxK	KKSLNVAK KQARSEDK	1.1E-20
ORF6	ENxxxxLD	ENKYSQLD	5.7E-08	NSP9	KV	KV	2.6E-07
ORF7a	LKRK	LKRK	5.9E-02	NSP10	VxxAxxYK	VDAAKAYK	1.0E-09
ORF8	IxxxxxG	IQYIDIG	4.5E-10	NSP12	DxxxxxxE	DLTKGPHE DNTSRYWE	6.0E-09
ORF10	NxIxQV	NYIAQV	6.1E-04	NSP13	KxxxE	KATEE KGTLE	3.3E-19
NSP1	ExxVxYxK	EIPVAYRK	1.2E-10	NSP14	KxxxVNL	KSAFVNL	1.3E-05
NSP2	GxxKSxL	GEQKSIL	9.6E-11	NSP15	IxPxPxV	IKPVPEV	5.9E-07

a: 'x' indicated a less conserved position occupied by a class of or random amino acid residue; b: amino acid residues here were coloured according to the propensity to promote intrinsic disorder, i.e., red for disorder-promoting residues (P, E, S, Q, K), blue for order-promoting residues (C, W, I, Y, F, L, H, V, N, M), and black for disorder-order neutral residues (alanine, glycine, aspartic acid, threonine, arginine) [415,416]; c: p-value *listed here was obtained via Pearson's chi-squared test and adjusted by Benjamini-Hochberg correction procedure to avoid type I error [264].* Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; ORF, open reading frame; NSP, non-structural protein; SLiM, short linear motif.

5.4.3 Involvement of SLiMs in intrinsically disordered regions

Intrinsically disordered regions are polypeptide segments that lack fixed or stable tertiary structures [455]. They are advantageous in binding to multiple partners [356] and can sustain their functionality in spite of harsh environmental conditions [456], thus become perfect targets for virus mimicry. As mentioned in the previous section, some SLiMs were not detected in the IDRs of the viral sequence. They were structured and might even be involved in some independent folding units (globular domains), which were usually conserved during evolution [356,457]. Recent research investigated the possibility of such structural mimicry for coronaviruses [120]. This goes against the rapid rewiring of virus interactions with the host molecules [356], thus is not the focus in this study. Here, we propose that SLiMs widely detected in the IDR of both human and viral sequence were more likely to be evolutionarily convergent [4]. Meanwhile, mimicry was more likely to have evolved in SLiMs of the virus to assist in specific virus-host molecular interactions.

To address this, in this study, we utilised ESpritz [266] to discern ordered and disordered regions within human and SARS-CoV-2 proteins. Compared with other IDR predictors [168,458,459], results from ESpritz generally involved more positions with potential to be disordered (**S5.1 Data**). We found that the majority of our detected SLiMs were located within at least one IDR of the human protein. It enhanced the confidence for some SLiMs to have independent functions and to evolve convergently [4,418]. By contrast, only 2779 of our detected SLiMs (~35%) were observed in the IDRs of corresponding viral sequences while the rest 5050 SLiMs (~65%) were not. When including SLiMs observed in both the IDRs of PMHPs and the corresponding SARS-CoV-2 proteins, only 224 SLiMs with significant enrichment status were retained (adjusted $P<0.01$, **S5.7 Data**). Based on the calculated ASA [448], these SLiMs did not involve any completely buried amino acid residues (ASA = 0 square Ångstroms).

Comparing with the background dataset S1, SLiM ‘NxxxSRxP’ was over-represented in the IDRs of human proteins potentially mimicked by SARS-CoV-2 envelope protein (Pearson's chi-squared test: $F_1=1.3\%$, $F_2=0.4\%$, $P=0.003$). Enrichment of SLiM ‘SGxxxGxS’ remained after adding disorder requirement in both PMHPs and SARS-CoV-2 ORF7a ($F_1=7.5\%$, $F_2=2.8\%$, $P=6.2E-03$). SLiM ‘ATAxExxE’ was the only one remaining after screening for SARS-CoV-2 NSP mimicry (F_1=0.6%, F_2=0.1%, adjusted P=5.4E-05), suggesting NSP8 mimicry. The observation of SLiM 'NxxxSRxP', 'SGxxxGxS' and 'ATAxExxE' in IDRs of PMHPs and corresponding viral proteins is illustrated in **Fig 5.4**. The latter two SLiMs were observed multiple times even in a same human protein. We also noticed that non-conserved positions of these SLiMs showed a bias in amino acid compositions. For instance, the first non-conserved position of SLiM 'SGxxxGxS' tended to be glycine (G). However, such propensity was not mimicked by the corresponding SARS-CoV-2 protein.

The remaining 221 SLiMs were highly enriched with conserved amino acid residue 'S' and 'R'. They were all screened from the IDRs of human proteins potentially mimicked by SARS-CoV-2 nucleocapsid phosphoprotein (**S5.7 Data**). They were located in eight regions of the viral protein, including (1) P6~P13, (2) R32~R41, (3) R92~R97, (4) G175~S197, (5) S202~R209, (6) P364~A376, (7) D399~K405, and (8) S410~S416.

Protein	Segment	Residue
Human_WDR33	GPGPN KGDSR GPPNH	950~965
Human_RBMX	SMNFN MSSSR GPLPV	135~149
Human_RBMXL1	SMNFN MSSSR GPLPV	135~149
Human_BCLAF1	RPVWN RRHSR SPRRG	94~108
Human_PRPF4B	SGKEN RSPSR RPGRS	335~349
Virus_Envelope	SRVKN LNSSR VPDLL	60~74
Human_SMARCAD1	SEYDS GSDVG SSLDE	358~372
Human_MAPT	EPPKS GDRSG YSSPG	504~518
Human_CHD3	LDYGS GEDDG KSDKR	593~607
Human_GSK3A	GGGPS GGGPG GSGRA	3~17
Human_GSK3A	GASSS GGGPG GSGGG	61~75
Human_HNRNPD	GGTAS GGTEG GSAES	49~63
Human_SOX2	PQQTS GGGGG NSTAA	14~28
Human_VCP	PSQGS GGGTG GSVYT	783~797
Virus_ORF7a	EPCSS GTYEG NSPFH	33~47
Human_KPNA1	RRNVA TAEEE TEEEV	47~61
Human_TUBB2A	QYQDA TADEQ GEFEE	424~438
Human_TUBB2B	QYQDA TADEQ GEFEE	424~438
Human_TUBB3	QYQDA TAEEE GEMYE	424~438
Human_TUBB4B	QYQDA TAEEE GEFEE	424~438
Human_NUMA1	AQQLA TAAEE REASL	582~596
Human_NLRP5	QQDSA TAAET KEQEI	176~190
Human_NLRP5	EQEGA TAAET EEQEI	195~209
Human_NLRP5	EQEGA TAAET EEQGH	214~228
Human_BRIX1	AKRHA TAEEV EEEER	30~44
Virus_NSP8	YAAFA TAQEA YEQAV	12~26

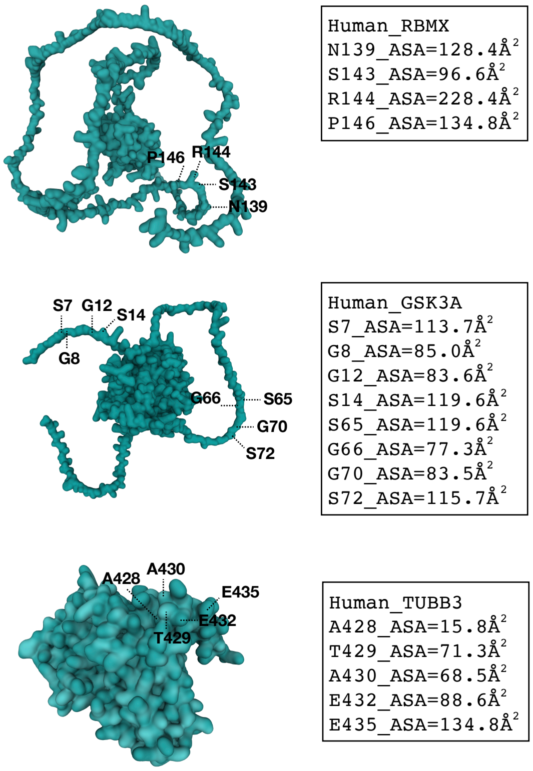


Fig 5.4 Illustration of potentially mimicked SLiMs observed in the IDR of both human and SARS-CoV-2 proteins. Human protein 'RBMX', 'GSK3A' and 'TUBB3' were chosen as representatives to show the location of corresponding SLiMs in the protein. Protein structures used here were generated by AlphaFold [430]. ASAs of the defined positions in

the exemplified SLiMs were calculated by the PSAIA programme [448] and shown in the right box. Higher ASA indicated a more exposed status. Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; IDR, intrinsically disordered region; ASA, accessible surface area; ORF, open reading frame; RBMX, RNA binding motif protein X-linked; GSK3A, glycogen synthase kinase 3 alpha; TUBB3, tubulin beta 3 class III.

5.4.4 Mimicry in SARS-CoV-2

Upon the aforementioned screening, we obtained three SLiMs for the mimicry of SARS-CoV-2 envelope, ORF7a, and NSP8, and 221 SLiMs for nucleocapsid mimicry. PMHPs with SLiM 'NxxxSRxP' arose in the nucleus region of the cell [460-462], which satisfied the condition to encounter with SARS-CoV-2 envelope proteins [449]. Most of them were involved in RNA binding, mRNA splicing, and positive regulation of transcription (**Table 5.5, S5.8 Data**). They shared SARS-CoV-2 envelope-targeting protein 'ZC3H18' as the interaction partner. These discoveries indicate that SARS-CoV-2 envelope may mimic SLiM 'NxxxSRxP' from one or some human proteins when reaching the nucleus [449] and use this host-like SLiM to interact with ZC3H18 (**Table 5.8**).

Table 5.5 Functionality of the 'qualifying' PMHPs of SARS-CoV-2 envelope protein.

PMHP	Expression[a]	Molecular function[b]
WDR33	18.05	RNA binding.
BCLAF1	43.88	DNA binding; RNA binding; transcription coregulator activity.
PRPF4B	41.52	ATP binding; protein kinase activity; protein serine kinase activity; protein threonine kinase activity; RNA binding.
RBMX	78.05	chromatin binding; identical protein binding; mRNA binding; protein domain specific binding; RNA binding; RNA polymerase II cis-regulatory region sequence-specific DNA binding.
RBMXL1	11.93	single-stranded RNA binding.

a: these data are represented by the medium TPM in lung from the GTEx project [152]; b: the function profile were retrieved from UniProt [2] and GOC [246]. Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; PMHP, potentially mimicked human protein; GTEx, Genotype-Tissue Expression; TPM, transcript per million; GOC, Gene Ontology Consortium; WDR33, WD repeat domain 33; BCLAF1, BCL2 associated transcription factor 1; PRPF4B, pre-mRNA processing factor 4B; RBMX, RNA binding motif protein X-linked; RBMXL1, RBMX like 1.

The majority of PMHPs with SLiM 'SGxxxGxS' were involved in molecule binding and the regulation of many biological processes (**Table 5.6, S5.8 Data**). As the subcellular localisation of SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1(SMARCAD1) did not match that of SARS-CoV-2

ORF7a in the infected host cell [449,463], it was less likely to be a true mimicking target. When ignoring SMARCAD1, the enrichment status of SLiM 'SGxxxGxS' in PMHPs was influenced but still at a significant level (Pearson's chi-squared test: F_1=6.5%, F_2=2.8%, P=0.03). The remaining PMHPs shared SARS-CoV-2 ORF7a-targeting protein 'midasin AAA ATPase 1 (MDN1)' as the interaction partner. These results suggest that SARS-CoV-2 ORF7a mimicry may happen in the cytoplasm. SLiM 'SGxxxGxS' may be mimicked to promote virus-host interaction with MDN1 (**Table 5.8**).

Table 5.6 Functionality of the 'qualifying' PMHPs of SARS-CoV-2 ORF7a.

PMHP	Expression[a]	Molecular function[b]
MAPT	1.27	actin binding; apolipoprotein binding; chaperone binding; DNA binding; double-stranded DNA binding; dynactin binding; enzyme binding; histone-dependent DNA binding; Hsp90 protein binding; identical protein binding; lipoprotein particle binding; microtubule binding; microtubule lateral binding; minor groove of adenine-thymine-rich DNA binding; phosphatidylinositol binding; phosphatidylinositol bisphosphate binding; protein kinase binding; protein-macromolecule adaptor activity; protein phosphatase 2A binding; RNA binding; sequence-specific DNA binding; SH3 domain binding; single-stranded DNA binding.
CHD3	68.23	ATP binding; ATP hydrolysis activity; DNA binding; DNA helicase activity; double-stranded DNA helicase activity; helicase activity; nucleosome-dependent ATPase activity; RNA binding; transcription cis-regulatory region binding; zinc ion binding.
GSK3A	43.09	ATP binding; protein kinase A catalytic subunit binding; protein serine/threonine kinase activity; protein serine kinase activity; protein threonine kinase activity; signaling receptor binding; tau protein binding; tau-protein kinase activity.
HNRNPD	135.4	chromatin binding; histone deacetylase binding; minor groove of adenine-thymine-rich DNA binding; mRNA 3'-UTR AU-rich region binding; RNA binding; telomeric DNA binding.
VCP	165.6	ADP binding; ATP binding; ATP hydrolysis activity; BAT3 complex binding; deubiquitinase activator activity; identical protein binding; K48-linked polyubiquitin modification-dependent protein binding; lipid binding; MHC class I protein binding; polyubiquitin modification-dependent protein binding; protein domain specific binding; protein phosphatase binding; RNA binding; ubiquitin-like protein ligase binding; ubiquitin protein ligase binding; ubiquitin-specific protease binding.
SOX2	0.64	DNA binding; DNA-binding transcription activator activity, RNA polymerase II-specific; DNA-binding transcription factor activity; DNA-binding transcription factor activity, RNA polymerase II-specific; miRNA binding; RNA polymerase II cis-regulatory region sequence-specific DNA binding; sequence-specific DNA binding; transcription cis-regulatory region binding.

a: these data are represented by the medium TPM in lung from the GTEx project [152]; b: the function profile were retrieved from UniProt [2] and GOC [246]. Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; PMHP, potentially mimicked human protein; ORF, open reading frame; GTEx, Genotype-Tissue Expression; TPM, transcript per million; GOC, Gene Ontology Consortium; MAPT, microtubule associated protein tau; CHD3, chromodomain helicase DNA binding protein 3; GSK3A, glycogen synthase kinase 3 alpha; HNRNPD, heterogeneous nuclear ribonucleoprotein D; VCP, valosin containing protein; SOX2, SRY-box transcription factor 2.

We found that some PMHPs containing SLiM 'ATAxExxE' were associated with GTP/tubulin binding, structural cytoskeleton, mitotic cell cycle, or neuron migration (**Table 5.7, S5.8 Data**). Biogenesis of ribosomes BRX1 (BRIX1) only occurred in nucleolus [464], from which it was less likely to be mimicked by NSP8 of SARS-CoV-2 [449]. Based on 578 sample tests from the GTEx project [152], NLR family pyrin domain containing 5 (NLRP5) did not express in lung (TPM = 0). Ignoring the occurrence of SLiM 'ATAxExxE' in BRIX1 and NLRP5 reduced the significance of SLiM enrichment in PMHPs (F_1=0.4%, F_2=0.1%, adjusted P=2.7E-03). Most of the rest PMHPs with SLiM 'ATAxExxE' shared SARS-CoV-2 NSP8-targeting protein 'La ribonucleoprotein 7 (LARP7)', and 'methylphosphate capping enzyme (MEPCE)' as their interaction partners. An exception was found on nuclear mitotic apparatus protein 1 (NUMA1). It interacted with DEAD-box helicase 10 (DDX10) and nuclear receptor binding SET domain protein 2 (NSD2), which were targeted by SARS-CoV-2 NSP8. These results suggest that SARS-CoV-2 NSP8 may mimic SLiM 'ATAxExxE' to interact with LARP7 and MEPCE (**Table 5.8**).

Table 5.7 Functionality of the 'qualifying' PMHPs of SARS-CoV-2 NSP8.

PMHP	Expression[a]	Molecular function[b]
TUBB2B	1.28	GTP binding; protein heterodimerization activity; structural constituent of cytoskeleton.
TUBB3	1.76	GTP binding; netrin receptor binding; peptide binding; structural constituent of cytoskeleton.
TUBB2A	11.13	GTP binding; structural constituent of cytoskeleton.
TUBB4B	255.4	double-stranded RNA binding; GTP binding; MHC class I protein binding; structural constituent of cytoskeleton; unfolded protein binding.
KPNA1	18.9	nuclear import signal receptor activity; nuclear localization sequence binding.
NUMA1	67.16	disordered domain specific binding; dynein complex binding; microtubule binding; microtubule minus-end binding; microtubule plus-end binding; phosphatidylinositol binding; protein-containing complex binding; protein C-terminus binding; protein domain specific binding; structural molecule activity; tubulin binding.

a: these data are represented by the medium TPM in lung from the GTEx project [152]; b: the function profile were retrieved from UniProt [2] and GOC [246]. Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; PMHP, potentially mimicked human protein; NSP, non-structural protein; GTEx, Genotype-Tissue Expression; TPM, transcript per million; GOC, Gene Ontology Consortium; TUBB2B, tubulin beta 2B class IIb; TUBB3, tubulin beta 3 class III; TUBB2A, tubulin beta 2A class IIa; TUBB4B, tubulin beta 4B class IVb; KPNA1, karyopherin subunit alpha 1; NUMA1, nuclear mitotic apparatus protein 1.

We obtained 221 SLiMs from 1121 PMHPs of SARS-CoV-2 nucleocapsid phosphoprotein. However, many of these PMHPs were not detected in the cytosol region based on the annotation data from UniProt [2], Reactome [150], HPA [450] or Ensembl databases [244]. Inconsistency of the subcellular localisation between PMHPs and SARS-CoV-2 nucleocapsid phosphoprotein reduced the confidence of virus mimicry for these SLiMs. Additionally, expression data from the GTEx [152] showed that the expression level of some PMHPs was extremely limited (medium TPM<0.5) in the lung, which also indicated that they were less likely to be involved in the mimicry of SARS-CoV-2 nucleocapsid phosphoprotein. After ignoring the occurrence of SLiMs in these ‘inappropriate’ PMHPs, 15 SLiMs remained. They had two distinct types of propensity in their amino acid compositions (SR-enriched and K-enriched) and were observed in two regions of SARS-CoV-2 nucleocapsid phosphoprotein: S184~R195 and P368~K375. The first region was included in the ‘SR/RS’ domains, which might contribute to the unpacking of coronavirus RNA and its virion assembly [465]. Here we divided these 15 SLiMs into two groups, namely ‘SR-featured’ SLiMs and ‘PK-featured’ SLiMs. 45 PMHPs were found to have ‘SR-featured’ SLiMs. Many of them were associated with DNA/RNA/ATP binding and RNA processing (**S5.8 Data**). The top-three virus targets shared among their PPI sub-networks were Smad nuclear interacting protein 1 (SNIP1), Mov10 RISC complex RNA helicase (MOV10), and G3BP stress granule assembly factor 1 (G3BP1) (**Fig 5.5A**). On the other hand, ‘PK-featured’ SLiMs arose in 25 PMHPs of SARS-CoV-2 nucleocapsid phosphoprotein. Many of them were involved in molecule binding, regulation of transcription, and innate immune response (**S5.8 Data**). Among them, MOV10, DExD-box helicase 21 (DDX21) and G3BP1 were more connected to the virus targets (**Fig 5.5B**). In view of all these, here, we propose our hypothesis of another virus mimicry: SARS-CoV-2 nucleocapsid phosphoprotein may mimic 15 ‘SR-featured’ and ‘PK-featured’ SLiMs from human proteins and orchestrate their representation in its S184~R195 (SRSSSRSRNSSR) and P368~K375 (PKKDKKKK) regions so as to enhance its interaction affinity with SNIP1, MOV10, G3BP1, DDX21, etc (**Table 5.8**).

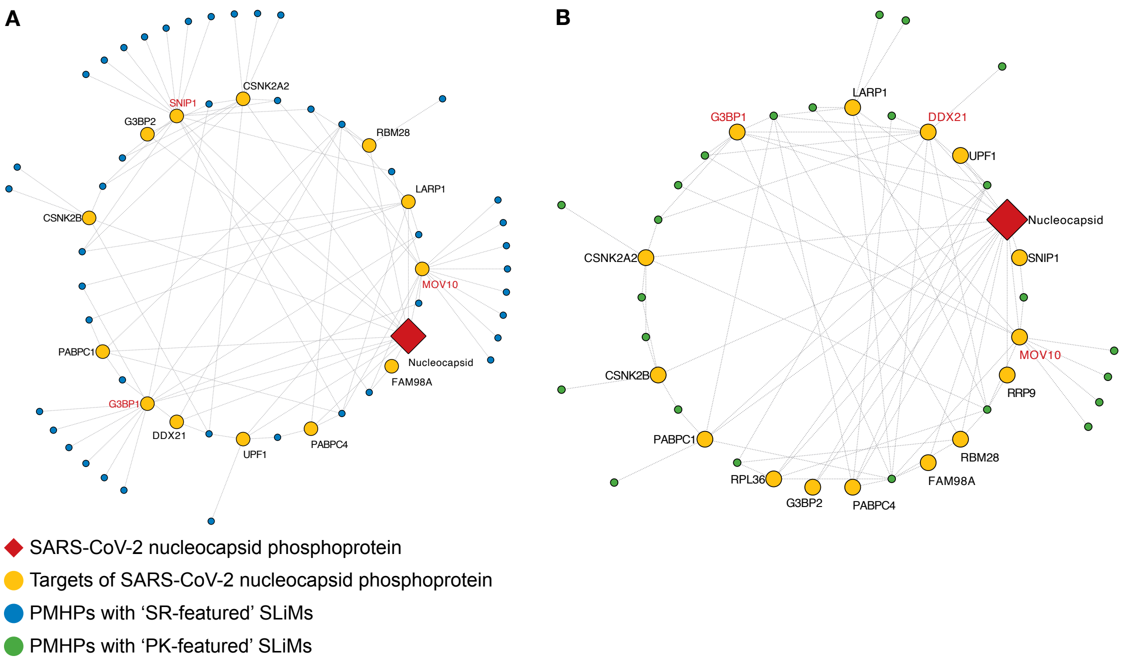


Fig 5.5 'High-confidence' interactions among SARS-CoV-2 nucleocapsid phosphoprotein, its human targets, and PMHPs with (A) 'SR-featured' and (B) 'PK-featured' SLiMs. Red label highlighted the top-three virus targets shared as the interaction partners of the corresponding PMHPs. Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; PMHP, potentially mimicked human protein; SNIP1, Smad nuclear interacting protein 1; MOV10, Mov10 RISC complex RNA helicase; G3BP1, G3BP stress granule assembly factor 1; DDX21, DExD-box helicase 21.

Table 5.8 Host-like SLiMs mimicked by SARS-CoV-2 proteins.

SLiM[a]	Detected region in viral protein	P-value[b]	Enhanced interaction affinity[c]
NxxxSRxP	Envelope: N64~P71	3.00E-03	ZC3H18
SGxxxGxS	ORF7a: S37~S44	6.20E-03	MDN1
ATAxExxE	NSP8: A16~E23	5.40E-05	LARP7, MEPCE
SRSRxxSR	Nucleocapsid: S184~R195	5.10E-03	SNIP1, MOV10, G3BP1
SRxxSRSR	Nucleocapsid: S184~R195	2.10E-02	SNIP1, MOV10, CSNK2A2
SRSRxxxR	Nucleocapsid: S184~R195	1.50E-02	SNIP1, MOV10, CSNK2A2
RSxSRSR	Nucleocapsid: S184~R195	2.10E-02	SNIP1, MOV10, CSNK2A2
RSxxRSR	Nucleocapsid: S184~R195	2.60E-02	SNIP1, MOV10, CSNK2A2
SRxRxxSR	Nucleocapsid: S184~R195	2.50E-02	SNIP1, MOV10, G3BP1
SRxxSxSR	Nucleocapsid: S184~R195	3.10E-02	SNIP1, MOV10, CSNK2A2
SxSRxxSR	Nucleocapsid: S184~R195	3.00E-02	SNIP1, MOV10, CSNK2A2
SRSxSRSR	Nucleocapsid: S184~R195	3.30E-02	SNIP1, CSNK2A2, MOV10
PKxxKKxK	Nucleocapsid: P368~K375	3.90E-02	DDX21, MOV10, G3BP1

SRSxxxSR	Nucleocapsid: S184~R195	3.80E-02	SNIP1, MOV10, G3BP1
PxKxKKxK	Nucleocapsid: P368~K375	4.40E-02	MOV10, DDX21, LARP1
SRSxSxSR	Nucleocapsid: S184~R195	4.7E-02	SNIP1, MOV10, CSNK2A2
SRSxxRSR	Nucleocapsid: S184~R195	5.0E-02	SNIP1, CSNK2A2, MOV10
RxxSRSR	Nucleocapsid: S184~R195	4.8E-02	SNIP1, MOV10, G3BP1

a: 'x' indicated a less conserved position occupied by a class of or random amino acid residue; b: p-value listed here was obtained via Pearson's chi-squared test and adjusted by Benjamini-Hochberg correction procedure to avoid type I error [264]; c: for putative SARS-CoV-2 nucleocapsid phosphoprotein mimicry, only the top three influenced virus targets were listed. Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SLiM, short linear motif; ZC3H18, zinc finger CCCH-type containing 18; MDN1, midasin AAA ATPase 1; LARP7, La ribonucleoprotein 7; MEPCE, methylphosphate capping enzyme; SNIP1, Smad nuclear interacting protein 1; MOV10, Mov10 RISC complex RNA helicase; G3BP1, G3BP stress granule assembly factor 1; CSNK2A2, casein kinase 2 alpha 2; DDX21, DExD-box helicase 21; LARP1, La ribonucleoprotein 1.

5.4.5 Association between SLiMs and typical SARS-CoV-2 mutations

SARS-CoV-2 mutations accumulate in its genome and can be spread during the pandemic [466]. A fraction of these mutations cause non-synonymous changes in the protein sequence, which may alter the infectivity, disease severity or interactions with host immunity [421,467]. After screening for intrinsic disorder, solvent exposure, subcellular localisation, and expression in the lung tissue, we obtained 18 SLiMs for the investigation of SARS-CoV-2 mimicry (**Table 5.8**). These SLiMs were exposed to the frequent mutation of SARS-CoV-2 but generally were conserved (**Fig 5.6**). SLiM 'SxSRxxSR' was even conserved across other human coronaviruses [434]. It was detected in S151~R158 (SQSRSQSR) of HoV-229E, S166~R173 (STSRQQSR) of HCoV-NL63, S195~R202 (SNSRPGSR) of HCoV-HKU1, S200~R207 (STSRTSSR) of HCoV-OC43, S177~R184 (SLSRNSSR) of MERS-CoV, and S187~R194 (SRSRGNSR) of SARS-CoV (**Fig 5.7**). 'PK-featured' SLiMs did not arise in HoV-229E, HCoV-NL63, HCoV-HKU1, and HCoV-OC43 due to the lack of proline/lysine-dominant regions. Additionally, we found that all of the nucleocapsid-mimicked SLiMs were conserved in SARS-CoV, indicating similar evolutionary constraints in Sarbecovirus subfamily members [468].

Since the screened SLiMs were only located in the envelope, nucleocapsid, ORF7a or NSP8 clades of SARS-CoV-2, mutations happened in spike, membrane, other accessory or non-structure proteins were ignored in the analyses. We noticed that envelope, ORF7a or NSP8 clades of SARS-CoV-2 were generally stable with limited non-synonymous mutations. Particularly, as beta variants of SARS-CoV-2 had 'P71L' mutation [399] in the envelope protein, SLiM ‘NxxxSRxP’ was destroyed (**Fig 5.4**). In the nucleocapsid phosphoprotein of SARS-CoV-2 variants, all of the screened 15 SLiMs were not influenced by non-synonymous substitutions.

Together, the screened 18 SLiMs are generally conserved in different SARS-CoV-2 variants. SLiM ‘SxSRxxSR’ is retained through purifying selection in human coronaviruses.

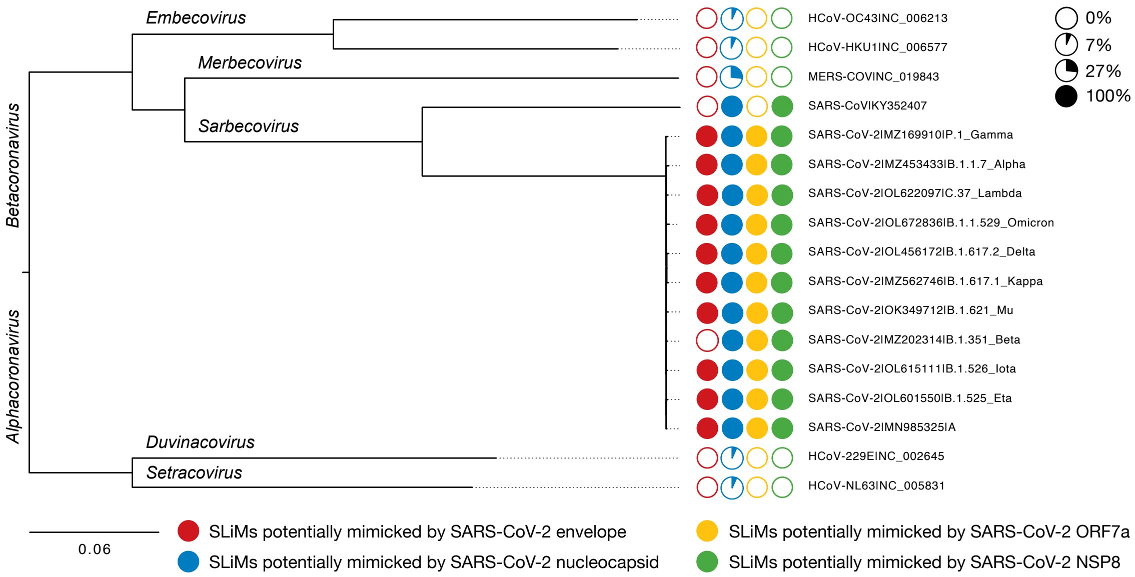

Embecovirus
Merbecovirus
Sarbecovirus
Betacoronavirus
Alphacoronavirus
Duvinacovirus
Setracovirus
HCoV-OC43|NC_006213
HCoV-HKU1|NC_006577
MERS-COV|NC_019843
SARS-CoV|KY352407
SARS-CoV-2|MZ169910|P.1_Gamma
SARS-CoV-2|MZ453433|B.1.1.7_Alpha
SARS-CoV-2|OL622097|C.37_Lambda
SARS-CoV-2|OL672836|B.1.1.529_Omicron
SARS-CoV-2|OL456172|B.1.617.2_Delta
SARS-CoV-2|MZ562746|B.1.617.1_Kappa
SARS-CoV-2|OK349712|B.1.621_Mu
SARS-CoV-2|MZ202314|B.1.351_Beta
SARS-CoV-2|OL615111|B.1.526_Iota
SARS-CoV-2|OL601550|B.1.525_Eta
SARS-CoV-2|MN985325|A
HCoV-229E|NC_002645
HCoV-NL63|NC_005831
0%
7%
27%
100%
0.06
SLiMs potentially mimicked by SARS-CoV-2 envelope
SLiMs potentially mimicked by SARS-CoV-2 nucleocapsid
SLiMs potentially mimicked by SARS-CoV-2 ORF7a
SLiMs potentially mimicked by SARS-CoV-2 NSP8

Fig 5.6 Conservation of SLiMs in the human-infecting coronaviruses and SARS-CoV-2 variants. The phylogenetic tree was generated based on the sequence alignment of representative coronavirus genomes (**Table 5.1, S5.2 Data**). Abbreviations: SARS-CoV, severe acute respiratory syndrome coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; HCoV, human coronavirus; MERS-CoV, middle east respiratory syndrome coronavirus; SLiM, short linear motif; ORF, open reading frame; NSP, non-structural protein.

```
HCoV-229E|Nucleocapsid       ---------------------MATVKWADASEPQ-----RG----------------RQG 18
HCoV-NL63|Nucleocapsid       ---------------------MASVNWADDRAA------------------------RKK 15
HCoV-HKU1|Nucleocapsid1      MSYTPGHYAGSRSSSGNRSGILKKTSWADQSERNYQTFNRGRKTQPKFTVST--QPQGNT 58
HCoV-OC43|Nucleocapsid       MSFTPGKQSSSRASSGNRSGNG-ILKWADQSDQFRNVQTRGRRAQPKQTATSQQPSGGNV 59
MERS-CoV|Nucleocapsid        -------MA--------SPAAPRAVSFADNNDI----TNTNLSR----GRGRNPKPRAAP 37
SARS-CoV-2|Nucleocapsid      MS--DNGPQ--------NQRNAPRITFGGPSDSTGSNQNGERSG----ARSKQRRPQGLP 46
SARS-CoV|Nucleocapsid        MT--DNGQQ--------GPRNAPRITFGV-SDNFDNNQNGDRTG----ARPKHRRPQGPP 45
                                                      .:.

HCoV-229E|Nucleocapsid       RIPYSLYSPLLVDS-EQPWKVIPRNLVPINKKD-KNKLIGYWNVQKR--FRTRKGKRVDL 74
HCoV-NL63|Nucleocapsid       FPPPSFYMPLLVSSDKAPYRVIPRNLVPIGKGN-KDEQIGYWNVQER--WRMRRGQRVDL 72
HCoV-HKU1|Nucleocapsid1      IPHYSWFSGITQFQKGRDFKFSDGQGVPIAFGVPPSEAKGYWYRHSRRSFKTADGQQKQL 118
HCoV-OC43|Nucleocapsid       VPYYSWFSGITQFQKGKEFEFVEGQGVPIAPGVPATEAKGYWYRHNRRSFKTADGNQRQL 119
MERS-CoV|Nucleocapsid        NNTVSWYTGLTQHGK-VPLTFPPGQGVPLNANSTPAQNAGYWRRQDRK-INTGNG-IKQL 94
SARS-CoV-2|Nucleocapsid      NNTASWFTALTQHGK-EDLKFPRGQGVPINTNSSPDDQIGYYRRATRR-IRGGDGKMKDL 104
SARS-CoV|Nucleocapsid        NNTASWFTALTQHGK-ETLTFPRGQGVPINTNSGKDDQIGYYRRASRR-VRGGDGKMKEL 103
                                 * :  :          .   : **:       .  **:    *   .   *   :*

HCoV-229E|Nucleocapsid       SPKLHFYYLGTGPHKDAKFRERVEGVVWVAVDGAKTEPT-GYGVRRKNSEPEIPH-F--N 130
HCoV-NL63|Nucleocapsid       PPKVHFYYLGTGPHKDLKFRQRSDGVVWVAKEGAKTVNT-SLGNRKRNQKPLEPK-F--S 128
HCoV-HKU1|Nucleocapsid1      LPRWYFYYLGTGPYANASYGESLEGVFWVANHQADTSTPSDVSSRDPTTQEAIPTRFPPG 178
HCoV-OC43|Nucleocapsid       LPRWYFYYLGTGPHAKDQYGTDIDGVYWVASNQADVNTPADIVDRDPSSDEAIPTRFPPG 179
MERS-CoV|Nucleocapsid        APRWYFYYTGTGPEAALPFRAVKDGIVWVHEDGATDAPS-TFGTRNPNNDSAIVTQFAPG 153
SARS-CoV-2|Nucleocapsid      SPRWYFYYLGTGPEAGLPYGANKDGIIWVATEGALNTPKDHIGTRNPANNAAIVLQLPQG 164
SARS-CoV|Nucleocapsid        SPRWYFYYLGTGPEAALPYGANKEGIVWVATEGALNTPKEHIGTRNPNNNAAIVLQLPQG 163
                              *: :*** ****     :    :*: **  . *          *    .      :  .

HCoV-229E|Nucleocapsid       QKLPNGVTVVEEPDSRA-----PSRSQSRSQSRGRGESKPQSRNPSSDRNHNSQDDIMKA 185
HCoV-NL63|Nucleocapsid       IALPPELSVVEFEDRSNNSSRASSRSSTRNNSRDSSRSTSRQQSRTRSDSNQSSSDLVAA 188
HCoV-HKU1|Nucleocapsid1      TILPQGYYVEGSGRSAS-NSR----PGSRSQSRGPNN---RSLSRSNSNFRHSD------ 224
HCoV-OC43|Nucleocapsid       TVLPQGYYIEGSGRSAP-NSR----STSRTSSRASSA---GSRSRANSGNRTPT------ 225
MERS-CoV|Nucleocapsid        TKLPKNFHIEGTGGNSQSSSRAS--SLSRNSSRSSSQ---GSRSGNSTRGTSPGPSGIGA 208
SARS-CoV-2|Nucleocapsid      TTLPKGFYAEGSRGGSQASSRSS--SRSRNSSRNSTP---GSSRGTSPARMA---GNGGD 216
SARS-CoV|Nucleocapsid        TSLPKGFYAEGSRGGSQTASRSS--SRSRGNSRNSTP---SSSRGSSPARNL---QAGGD 215
                               **                       :* .**        .

HCoV-229E|Nucleocapsid       VAAALKSLGFDKPQEKDKKSAKTGTPKPSRNQSPASSQTSAKSLARSQSSETKEQKHEMQ 245
HCoV-NL63|Nucleocapsid       VTLALKNLGFDNQSKSPSSS---GTSTPKKPNKPL-------------SQPRADKPSQLK 232
HCoV-HKU1|Nucleocapsid1      ---SIVKPDMADEIANL-VLAKLGKDS--KPQQVT----------K--QNAKEIRHKILT 266
HCoV-OC43|Nucleocapsid       ---SGVTPDMADQIASL-VLAKLGKDA-TKPQQVT----------K--HTAKEVRQKILN 268
MERS-CoV|Nucleocapsid        VGGDLLYLDLLNRLQAL-ESGKV--KQ-SQPKVIT----------K--KDAA----AAKN 248
SARS-CoV-2|Nucleocapsid      AALALLLLDRLNQLESK-MSGKG--QQ-QQGQTVT----------K--KSAA----EASK 256
SARS-CoV|Nucleocapsid        TALALLLLDRLNQLESK-VSGKT--QQ-QQPQVVT----------R--KSAS----EASK 255
                                     .  .                 : :

HCoV-229E|Nucleocapsid       KPRWKRQPNDDVTSNVTQCFGPRDLDH---NFGSAGVVANGVKAKGYPQFAELVPSTAAM 302
HCoV-NL63|Nucleocapsid       KPRWKRVPTRE--ENVIQCFGPRDFNH---NMGDSDLVQNGVDAKGFPQLAELIPNQAAL 287
HCoV-HKU1|Nucleocapsid1      KPRQKRTPNKH--CNVQQCFGKRGPSQ---NFGNAEMLKLGTNDPQFPILAELAPTPGAF 321
HCoV-OC43|Nucleocapsid       KPRQKRSPNKQ--CTVQQCFGKRGPNQ---NFGGGEMLKLGTSDPQFPILAELAPTAGAF 323
MERS-CoV|Nucleocapsid        KMRHKRTSTKS--FNMVQAFGLRGPGDLQGNFGDLQLNKLGTEDPRWPQIAELAPTASAF 306
SARS-CoV-2|Nucleocapsid      KPRQKRTATKA--YNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKHWPQIAQFAPSASAF 314
SARS-CoV|Nucleocapsid        KPRQKRTATKQ--YNVTQAFGRRGPEQTQGNFGDQELIRLGTDYKNWPQIAQFAPSASAF 313
                             * * **  .     .: *.** *.  .   *:*.  :   *..   :* :*:: *. .*:

HCoV-229E|Nucleocapsid       LFDSHIVSKESG-----------NTVVLTFTTRVTVPKDHPHLGKFLEEL----NAFTRE 347
HCoV-NL63|Nucleocapsid       FFDSEVSTDEVG-----------DNVQITYTYKMLVAKDNKNLPKFIEQI----SAFTKP 332
HCoV-HKU1|Nucleocapsid1      FFGSKLDLVKRD---SEADSPVKDVFELHYSGSIRFDSTLPGFETIMKVLEENLNAYVNS 378
HCoV-OC43|Nucleocapsid       FFGSRLELAKVQNLSGNPDEPQKDVYELRYNGAIRFDSTLSGFETIMKVLNENLNAYQQQ 383
MERS-CoV|Nucleocapsid        MGMSQFKLTHQN-----NDDHGNPVYFLRYSGAIKLDPKNPNYNKWLELLEQNIDAYKTF 361
SARS-CoV-2|Nucleocapsid      FGMSRIGMEVTP-----------SGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAYKTF 363
SARS-CoV|Nucleocapsid        FGMSRIGMEVTP-----------TGTWLTYHGAIKLDDKDPNFKDQVILLNKHIDAYKTF 362
                             :  *..                     : :   : .          :  :    .*:

HCoV-229E|Nucleocapsid       MQQHPLLNPS----ALEFNP-----------------SQTS----PATAEPVRD---EVS 379
HCoV-NL63|Nucleocapsid       SSIKEMQSQS----SHVAQN---------------TVLNAS----IPESKPLAD---DDS 366
HCoV-HKU1|Nucleocapsid1      NQN---------TDSDSLSSKPQRKRGVKQLPEQFDSLNLSA----GTQHISNDFTPEDH 425
HCoV-OC43|Nucleocapsid       DG------------MMNMSPKPQRQRGHKNGQGENDNISVAVPKSRVQQNKSRELTAEDI 431
MERS-CoV|Nucleocapsid        PKKEKKQKAPKEESTDQMSEPPKEQRVQGSITQRTRTRP------SVQPGPMIDVNTD-- 413
SARS-CoV-2|Nucleocapsid      PPTEPKKDKKKKA--DETQALPQRQK----------KQQ------TVTLLPAADLDDFSK 405
SARS-CoV|Nucleocapsid        PPTEPKKDKKKKA--DEVQPLPQRQK----------KQP------TVTLLPAAELDDFSK 404
                                               .                                  :

HCoV-229E|Nucleocapsid       IETDIIDEVN-------  389
HCoV-NL63|Nucleocapsid       AIIEIVNEVLH------  377
HCoV-HKU1|Nucleocapsid1      SLLATLDDPYVEDSVA-  441
HCoV-OC43|Nucleocapsid       SLLKKMDEPYTEDTSEI  448
MERS-CoV|Nucleocapsid        -----------------  413
SARS-CoV-2|Nucleocapsid      QLQQSMSSAD--STQA-  419
SARS-CoV|Nucleocapsid        QLQDSMNGAS--DSTQA  419
```

Fig 5.7 Alignment for nucleocapsid phosphoproteins in seven human coronaviruses. Red background highlighted regions containing conserved SLiMs within human coronaviruses. Asterisk symbol indicated positions with a single, fully conserved residue. Colon and period symbol indicated conservation between groups of strongly and weakly similar properties, respectively. Abbreviations: SARS-CoV, severe acute respiratory syndrome coronavirus; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; HCoV, human coronavirus; MERS-CoV, middle east respiratory syndrome coronavirus.

5.5 Discussion

In this study, we found 84 human proteins more connected to the targets of SARS-CoV-2 proteins (**Table 5.3**). Putative mimicry on these human proteins might promote the affinity of multiple SARS-CoV-2-host PPIs. However, according to our final results shown in **Table 5.8**, the majority of them were not involved in mimicry of SARS-CoV-2. Here, we provide a few assumptions to explain this conflict. First, we propose that the virus is trying to target highly connected 'hub' proteins rather than mimicking patterns from them. For instance, NTRK1 interacted with 91 targets of SARS-CoV-2 (HIPPIE evidence score > 0.63). It could also be implicated in interactions with the NSP3 of SARS-CoV [469], which shared 94% sequence similarity with SARS-CoV-2 NSP3 [452]. Therefore, NTRK1 was expected to be targeted or at least indirectly targeted by SARS-CoV-2 during the infection. Besides, we found that some of these 84 human proteins were not only 'hub' proteins of the corresponding sub-network but also kept their high connectivity even in the entire human PPI network. As the virus is always trying to target 'hub' proteins to affect more signalling pathways [4], they may have already been in the interacting list of SARS-CoV-2 but are well-protected by the current immune defense. This further suggests that SARS-CoV-2 mimicry cannot be easily done by ignoring the host immune system. Successful mimicry may be achieved on 'less-protected' human proteins (**Table 5.8**, **S5.7 Data**). And in turn, the virus disguises itself as a 'less-protected' human protein with the mimicked SLiMs to hijack host proteins and circumvent immune responses [4,104]. Third, innate properties especially anti-viral properties of some human proteins make them less likely to become the targets of SARS-CoV-2 for interacting or mimicry. Exemplified by tripartite motif containing 25 (TRIM25), although its mediated ubiquitination of RIG-I can be inhibited by SARS-CoV nucleocapsid phosphoprotein, no evidence has been shown to prove its 'pro-viral' role during SARS-CoV-2 infections [470].

We identified 18 SLiMs involved in the mimicry of SARS-CoV-2 envelope, nucleocapsid, ORF7a and NSP8 (**Table 5.8**). However, the number of original SLiMs for mimicry was as many as 7957. Despite the requirements on appropriate disorder status, solvent exposure, subcellular localisation, and lung expression, many SLiMs were still removed as their presences in the current dataset were not significant enough to infer functionality. However, this does not mean that they are irrelevant to the mimicry of SARS-CoV-2. In fact, we removed human proteins in the computer-annotated TrEMBL section of UniProtKB [2] and those with less than 16 or more than 2700 amino acid residues to guarantee the quality of retrieved entries. After using more data, the enrichment status of the removed SLiMs may change and even become significant and indicate another instance of convergent evolution [418]. Additionally, although mimicries of other SARS-CoV-2 proteins have not been identified in this study, they may still exist and be activated along with ongoing mutations in SARS-CoV-2 genome [421]. Benefiting from the rapid development of SARS-CoV-2 research, now it is possible to include more virus-host PPI data for a deeper investigation. For example, coronavirus disease 2019 (COVID-19) section of UniProtKB (https://covid-19.uniprot.org/) has reported more than 200 SARS-CoV-2-host PPIs, many of which have not been included in our data. The introduction of these virus-host PPIs will alter sub-networks related to different SARS-CoV-2 proteins and may produce new mimicking stories with new evidence. Additionally, variant-related mimicries may also be characterised after adding new PPI data between human proteins and SARS-CoV-2 variants.

We propose that the purpose of virus mimicry is to enhance the interaction affinity with its human host targets (**Fig 5.1**). There is another theory indicating that SARS-CoV-2 may mimic host-like SLiMs to 'set up' human proteins following its infection [471] and inadvertently instigate harmful autoimmune disorders and dysfunctions in the respiratory system [472]. Investigation of this mimicry theory requires expression data before and after the infection of SARS-CoV-2, thus it is hard to elucidate here. According to the current experiments, 'PK-featured' SLiMs were found in 25 'qualifying' PMHPs (**Fig 5.5**). Among them, seven were involved in the host anti-viral activities such as defense response to virus (GO:0051607) and innate immune response (GO:0045087) (**S5.8 Data**). As it goes against the original intention of the virus to escape from host immune defense, we suspect that this mimicry is to 'set up' human proteins for its infection in the next round. Despite 'PK-featured' SLiMs, the remaining SLiMs listed in **Table 5.8** were found in PMHPs almost irrelevant to host immune or anti-viral activities (**S5.8 Data)**, except for splicing factor proline and glutamine rich (SFPQ). Then, based on the aforementioned second mimicry theory, SARS-CoV-2 nucleocapsid phosphoprotein may mimic SLiM 'SRxRxxSR' from SFPQ to instigate autoimmune pathologies [472].

Finally, it is unclear if the mimicry strategy is also processed by other SARS-CoV-2 proteins. More research is needed to elucidate virus mimicry underlying the rapid exploitation of the host system especially in the context of ongoing pandemic or future outbreaks. It is also interesting to develop a SLiM-based backward approach to predict human proteins at risk of being targeted by SARS-CoV-2 or its variants.

5.6 Summary

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the ongoing coronavirus disease 2019 (COVID-19) pandemic. It orchestrates mimicry strategies of using host-like short linear motifs (SLiMs) to assist in SARS-CoV-2-host interactions. Investigation on this mimicry machinery is crucial to better understanding how virus exploits the host system for its replication and persistence. In this study, we discern 18 SLiMs that may be mimicked by the envelope, nucleocapsid, open reading frame 7a (ORF7a), and non-structural proteins 8 (NSP8) of SARS-CoV-2. They are considered be to the result of convergent evolution and generally conserved across SARS-CoV-2 variants. Discoveries of this study contribute to the development of sequence-based approaches for predicting virus-host interactions.

5.7 Supporting information

S5.1 Data. SARS-CoV-2 sequences and their intrinsic disorder status.

URL: https://github.com/HChai01/SARS-COV-2/blob/main/Data/MN985325.txt

S5.2 Data. Genome sequence of seven human coronaviruses and ten SARS-CoV-2 variants.

URL: https://github.com/HChai01/SARS-COV-2/blob/main/Data/Virus_Genome.txt

S5.3 Data. Sequences of human proteins in the background dataset S1.

URL: https://github.com/HChai01/SARS-COV-2/blob/main/Data/Background.txt

S5.4 Data. 291 SARS-CoV-2-host PPIs compiled in this study.

URL: https://github.com/HChai01/SARS-COV-2/blob/main/Data/Virus-host-PPIs.csv

S5.5 Data. Endogenous interaction status involving 291 SARS-CoV-2 targets.

URL: https://github.com/HChai01/SARS-COV-2/blob/main/Data/PMHPs.csv

S5.6 Data. SLiMs highly enriched in human proteins potentially mimicked by different SARS-CoV-2 proteins.

URL: https://github.com/HChai01/SARS-COV-2/blob/main/Data/Enriched_SLiMs.txt

S5.7 Data. Representation of SLiMs screened for intrinsic disorder.

URL: https://github.com/HChai01/SARS-COV-2/blob/main/Data/SLiM_224.csv

S5.8 Data. Functionality of the ‘qualifying’ PMHPs.

URL: https://github.com/HChai01/SARS-COV-2/blob/main/Data/Function_PMHPs.csv

CHAPTER VI

6. FINAL DISCUSSION

Distinct from wet-lab-based virus research, this work explores understanding of virus-host molecular interactions from *in silico* data mining and machine learning perspectives. Three Results chapters were presented: about the infection of human immunodeficiency virus type 1 (HIV-1), the immune response of interferon-α (IFN-α), and the mimicry strategy of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Collectively, they display an interactive and intimate relationship between viruses and their hosts.

As stated in **Chapter 1.1**, this work makes seven main contributions to virus research. In the HIV-1 research topic (**Chapter 3**), we propose a novel perspective of virus-host interaction prediction by relating human proteins to pro-viral or pro-host phenotypes. We found that the roles of most human host proteins in HIV-1 infections can be compiled [217]. Some of them are rapidly implicated in interaction with multiple HIV-1 proteins (e.g., *env*, *gag*, and *tat*) [241] (**Fig 3.5**). Their inactivation or dysfunction permits HIV-1 to hijack the host regulation and signaling pathway more effectively [109]. This indicates a 'pro-viral' interaction. In some cases, host proteins antagonize the infection by controlling or inhibiting the expression HIV-1 genes [473]. This indicates a 'pro-host' interaction. A 'bidirectional' role was identified when multiple types of interaction are occurring, i.e., a host molecule was involved in both pro-viral and pro-host interactions with HIV-1. Investigation on the directionality of virus-host interactions provides a novel insight for HIV-1 infection and human host immune responses. Predicting this directionality complexity of virus host molecular interactions, as done here for the first time, represents a new challenge for machine learning projects. It also has the potential to contribute to the discovery of novel anti-viral therapeutic. In the same research, we conduct systematic analyses on hosts' innate immunity properties. We propose that host innate properties can influence the interaction affinity with HIV-1. For example, we found that higher numbers of protein-coding transcripts (ORFs), duplication rates, and evolutionary conservation correlate with HIV-1-host protein-protein interactions (PPIs). Extremes of these features (e.g., huge number of ORFs and singleton cases), hint at clues for the inhibition of antiviral possibilities in the presence of HIV-1. Identification of these features in host molecules contributes to a better understanding explaining why they are implicated during the infection.

In the second research topic (**Chapter 4**), we further extend the feature scope to better characterise genes' interferon response. We propose that interferon-α-stimulated human genes (ISGs) are less evolutionary conserved than genes that are not significantly stimulated after IFN-α treatments (non-ISGs). These ISGs show significant depletion of GC-content in the coding region of their canonical transcripts, which leads to differential representation in their nucleotide and amino acid compositions. We found that ISG products are more implicated in key pathways of the human protein-protein interaction (PPI) network but tend not to be hubs or bottlenecks. These discoveries explain the differential expression of ISG in IFN-α experiments (e.g., highly-up-regulated, slightly-up-regulated, down-regulated, etc) [315]. Meanwhile, the differential features representation between ISGs and non-ISGs (**Fig 4.4-4.9**) contributes to their predictability with machine learning models. These discoveries constitute a major milestone in the development of high-throughput identification of ISGs, which in turn will provide a clear picture of host immune activities in response to viral infection. Additionally, based on the representation of multiple features (**Fig 4.12**), we found ISGs and interform-repressed genes (IRGs) are more similar to each other than ISGs to non-ISGs. It is thus worth investigating the interferon-regulated genes and their predictability in the future.

In the third research topic (**Chapter 5**), we elucidate the mimicry mechanism adopted by SARS-CoV-2. We test whether the SARS-CoV-2 envelope may mimic host-like short linear motif (SLiM) 'NxxxSRxP' from one or some human proteins when reaching the nucleus [449] and use this host-like SLiM to interact with zinc finger CCCH-type containing 18 (ZC3H18) (**Table 5.8**). SLiM 'SGxxxGxS' may be mimicked in the cytoplasm to promote virus-host interaction with midasin AAA ATPase 1 (MDN1) (**Table 5.8**). SARS-CoV-2 non-structural protein 8 (NSP8) may mimic SLiM 'ATAxExxE' to interact with La ribonucleoprotein 7 (LARP7) and methylphosphate capping enzyme (MEPCE) (**Table 5.8**). SARS-CoV-2 nucleocapsid phosphoprotein may mimic 15 'SR-featured' and 'PK-featured' SLiMs from human proteins and orchestrate their representation in its S184~R195 (SRSSSRSRNSSR) and P368~K375 (PKKDKKKK) regions so as to enhance its interaction affinity with Smad nuclear interacting protein 1 (SNIP1), Mov10 RISC complex RNA helicase (MOV10), G3BP stress granule assembly factor 1 (G3BP1), DExD-box helicase 21 (DDX21), etc (**Table 5.8**). Interesting, these mimicked SLiMs are well conserved across the diversity of SARS-CoV-2 variants, which further enhance the confidence of SARS-CoV-2 mimicry from the evolution perspective [4,418,419]. These discoveries contribute to a better understanding of molecular interactions between SARS-CoV-2 and human hosts. They also contribute to the development of a sequence-based approach for predicting virus-host interactions.

There are many restrictions that limit the scope of this work. The main restrictions are caused by insufficient data or annotations. Some important genes/proteins involved in host immune system or virus-host interactions, for example, host entry receptors (**Table 2.1**) are a key focus of biomedical research thus tend to be well-annotated. However, some genes/proteins especially those limitedly expressed in host cells are not well-investigated. This is particularly obvious in how few extensive virus-host data sets there are. The obsolescence of data is another severe risk in machine learning. For instance, we notice that many terms under the catalogue of cellular component are obsoleted by the Gene Ontology Consortium (GOC) [246], which means the performance of classifiers trained with those data will be influenced. There is also a need to integrate different types of experimental data to better capture the dynamic nature of the host system and how viruses interact and exploit the host system. For example, although we retrieved many high-confidence SARS-CoV-2-host molecular interaction data from [407], they are still an insufficient representation of infection. In fact, we found many clues about the SARS-CoV-2 mimicry (**S5.6 Data**), but only a very limited fraction of them is significant enough to infer the functionality from the statistical perspective. Although in the end we discovery some mimicry strategies used by the SARS-CoV-2 envelope, nucleocapsid, ORF7a, and NSP8, no evidence is shown to prove that mimicry is not performed by other SARS-CoV-2 proteins such as spike glycoprotein. Additionally, the variation-related mimicry machinery has not been discovered as well. Such work is not possible without the support of novel molecular interaction data between SARS-CoV-2 variants [421] and human host proteins.

The aforementioned restrictions also give some ideas to conduct new research from a different angle. For example, if we do not have sufficient SARS-CoV-2 molecular data to support a thorough investigation, it may be interesting to predict putative SARS-CoV-2-interacting proteins with the mimicked SLiMs (**Table 5.8**). In the future, we will conduct a systematic identification of virus-interacting proteins. The directionality of other virus-host molecular interactions will also be investigated.

In conclusion, this work utilises data mining and machine learning technologies to identify factors that influence the molecular interactions between viruses and their hosts. These factors can be hosts' innate properties, expression of anti-viral ISGs, achievement of virus mimicry strategies, etc. We hope the findings of this work will be used someday in the future for the development of anti-viral therapeutics.

BIBLIOGRAPHY

1. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. 2017; 45(1): 39-53. https://doi.org/10.1093/nar/gkw1002 PMID: 27899557

2. Consortium U. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021; 49(D1): D480-D489. https://doi.org/10.1093/nar/gkaa1100 PMID: 33237286

3. Zhang Z, Cai Z, Tan Z, Lu C, Jiang T, Zhang G, et al. Rapid identification of human-infecting viruses. Transbound Emerg Dis. 2019; 66(6): 2517-2522. https://doi.org/10.1111/tbed.13314 PMID: 31373773

4. Davey NE, Travé G, Gibson TJ. How viruses hijack cell regulation. Trends in biochemical sciences. 2011; 36(3): 159-169. https://doi.org/10.1016/j.tibs.2010.10.002 PMID: 21146412

5. Hulo C, De Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, et al. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res. 2011; 39(suppl_1): D576-D582. https://doi.org/10.1093/nar/gkq901 PMID: 20947564

6. Ryu W-S. Virus life cycle. Molecular Virology of Human Pathogenic Viruses. 2017: 31. https://doi.org/10.1016/B978-0-12-800838-6.00003-5

7. Rampersad S, Tennant P. Replication and expression strategies of viruses. Viruses. 2018: 55. https://doi.org/10.1016/B978-0-12-811257-1.00003-6

8. Grove J, Marsh M. The cell biology of receptor-mediated virus entry. J Cell Biol. 2011; 195(7): 1071-1082. https://doi.org/10.1083/jcb.201108131 PMID: 22123832

9. Marsh M, Helenius A. Virus entry: open sesame. Cell. 2006; 124(4): 729-740. https://doi.org/10.1016/j.cell.2006.02.007 PMID: 16497584

10. Deng X, Ueda H, Su SB, Gong W, Dunlop NM, Gao JL, et al. A synthetic peptide derived from human immunodeficiency virus type 1 gp120 downregulates the expression and function of chemokine receptors CCR5 and CXCR4 in monocytes by

activating the 7-transmembrane G-protein-coupled receptor FPRL1/LXA4R. Blood. 1999; 94(4): 1165-1173 PMID: 10438703

11. Lai WK, Sun PJ, Zhang J, Jennings A, Lalor PF, Hubscher S, et al. Expression of DC-SIGN and DC-SIGNR on human sinusoidal endothelium: a role for capturing hepatitis C virus particles. Am J Pathol. 2006; 169(1): 200-208. https://doi.org/10.2353/ajpath.2006.051191 PMID: 16816373

12. Marzi A, Gramberg T, Simmons G, Möller P, Rennekamp AJ, Krumbiegel M, et al. DC-SIGN and DC-SIGNR interact with the glycoprotein of Marburg virus and the S protein of severe acute respiratory syndrome coronavirus. J Virol. 2004; 78(21): 12090-12095. https://doi.org/10.1128/jvi.78.21.12090-12095.2004 PMID: 15479853

13. Shimojima M, Takenouchi A, Shimoda H, Kimura N, Maeda K. Distinct usage of three C-type lectins by Japanese encephalitis virus: DC-SIGN, DC-SIGNR, and LSECtin. Arch Virol. 2014; 159(8): 2023-2031. https://doi.org/10.1007/s00705-014-2042-2 PMID: 24623090

14. Davis CW, Nguyen HY, Hanna SL, Sánchez MD, Doms RW, Pierson TC. West Nile virus discriminates between DC-SIGN and DC-SIGNR for cellular attachment and infection. J Virol. 2006; 80(3): 1290-1301. https://doi.org/10.1128/jvi.80.3.1290-1301.2006 PMID: 16415006

15. Kaul M, Ma Q, Medders K, Desai M, Lipton S. HIV-1 coreceptors CCR5 and CXCR4 both mediate neuronal cell death but CCR5 paradoxically can also contribute to protection. Cell Death Differ. 2007; 14(2): 296-305. https://doi.org/10.1038/sj.cdd.4402006 PMID: 16841089

16. Short JJ, Vasu C, Holterman MJ, Curiel DT, Pereboev A. Members of adenovirus species B utilize CD80 and CD86 as cellular attachment receptors. Virus Res. 2006; 122(1-2): 144-153. https://doi.org/10.1016/j.virusres.2006.07.009 PMID: 16920215

17. Low JA, Magnuson B, Tsai B, Imperiale MJ. Identification of gangliosides GD1b and GT1b as receptors for BK virus. J Virol. 2006; 80(3): 1361-1366. https://doi.org/10.1128/JVI.80.3.1361-1366.2006 PMID: 16415013

18. Li Q, Spriggs MK, Kovats S, Turk SM, Comeau MR, Nepom B, et al. Epstein-Barr virus uses HLA class II as a cofactor for infection of B lymphocytes. J Virol. 1997;

71(6): 4657-4662. https://doi.org/10.1128/JVI.71.6.4657-4662.1997 PMID: 9151859

19. Xiao J, Palefsky JM, Herrera R, Berline J, Tugizov SM. The Epstein–Barr virus BMRF-2 protein facilitates virus attachment to oral epithelial cells. Virology. 2008; 370(2): 430-442. https://doi.org/10.1016/j.virol.2007.09.012 PMID: 17945327

20. Montgomery RI, Warner MS, Lum BJ, Spear PG. Herpes simplex virus-1 entry into cells mediated by a novel member of the TNF/NGF receptor family. Cell. 1996; 87(3): 427-436. https://doi.org/10.1016/s0092-8674(00)81363-x PMID: 8898196

21. Krummenacher C, Nicola AV, Whitbeck JC, Lou H, Hou W, Lambris JD, et al. Herpes simplex virus glycoprotein D can bind to poliovirus receptor-related protein 1 or herpesvirus entry mediator, two structurally unrelated mediators of virus entry. J Virol. 1998; 72(9): 7064-7074. https://doi.org/10.1128/JVI.72.9.7064-7074.1998 PMID: 9696799

22. Martinez WM, Spear PG. Structural features of nectin-2 (HveB) required for herpes simplex virus entry. J Virol. 2001; 75(22): 11185-11195. https://doi.org/10.1128/JVI.75.22.11185-11195.2001 PMID: 11602758

23. Wang X, Huang DY, Huong S-M, Huang E-S. Integrin αvβ3 is a coreceptor for human cytomegalovirus. Nat Med. 2005; 11(5): 515-521. https://doi.org/10.1038/nm1236 PMID: 15834425

24. Maginnis MS, Haley SA, Gee GV, Atwood WJ. Role of N-linked glycosylation of the 5-HT2A receptor in JC virus infection. J Virol. 2010; 84(19): 9677-9684. https://doi.org/10.1128/JVI.00978-10 PMID: 20660194

25. Yoon C-S, Kim K-D, Park S-N, Cheong S-W. α6 integrin is the main receptor of human papillomavirus type 16 VLP. Biochemical and biophysical research communications. 2001; 283(3): 668-673. https://doi.org/10.1006/bbrc.2001.4838 PMID: 11341777

26. Geraghty RJ, Krummenacher C, Cohen GH, Eisenberg RJ, Spear PG. Entry of alphaherpesviruses mediated by poliovirus receptor-related protein 1 and poliovirus receptor. Science. 1998; 280(5369): 1618-1620. https://doi.org/10.1126/science.280.5369.1618 PMID: 9616127

27. Li Q, Ali MA, Cohen JI. Insulin degrading enzyme is a cellular receptor mediating varicella-zoster virus infection and cell-to-cell spread. Cell. 2006; 127(2): 305-316. https://doi.org/10.1016/j.cell.2006.08.046 PMID: 17055432

28. Mori Y, Yang X, Akkapaiboon P, Okuno T, Yamanishi K. Human herpesvirus 6 variant A glycoprotein H-glycoprotein L-glycoprotein Q complex associates with human CD46. J Virol. 2003; 77(8): 4992-4999. https://doi.org/10.1128/jvi.77.8.4992-4999.2003 PMID: 12663806

29. Tang H, Serada S, Kawabata A, Ota M, Hayashi E, Naka T, et al. CD134 is a cellular receptor specific for human herpesvirus-6B entry. Proceedings of the National Academy of Sciences. 2013; 110(22): 9096-9099. https://doi.org/10.1073/pnas.1305187110 PMID: 23674671

30. Garrigues HJ, Rubinchikova YE, DiPersio CM, Rose TM. Integrin αVβ3 binds to the RGD motif of glycoprotein B of Kaposi's sarcoma-associated herpesvirus and functions as an RGD-dependent entry receptor. J Virol. 2008; 82(3): 1570-1580. https://doi.org/10.1128/JVI.01673-07 PMID: 18045938

31. Summerford C, Samulski RJ. Membrane-associated heparan sulfate proteoglycan is a receptor for adeno-associated virus type 2 virions. J Virol. 1998; 72(2): 1438-1445. https://doi.org/10.1128/JVI.72.2.1438-1445.1998 PMID: 9445046

32. Akache B, Grimm D, Pandey K, Yant SR, Xu H, Kay MA. The 37/67-kilodalton laminin receptor is a receptor for adeno-associated virus serotypes 8, 2, 3, and 9. J Virol. 2006; 80(19): 9831-9836. https://doi.org/10.1128/JVI.00878-06 PMID: 16973587

33. Weigel-Kelley KA, Yoder MC, Srivastava A. α5β1 integrin as a cellular coreceptor for human parvovirus B19: requirement of functional activation of β1 integrin for viral entry. Blood. 2003; 102(12): 3927-3933. https://doi.org/10.1182/blood-2003-05-1522 PMID: 12907437

34. Graham KL, Halasz P, Tan Y, Hewish MJ, Takada Y, Mackow ER, et al. Integrin-using rotaviruses bind α2β1 integrin α2 I domain via VP4 DGE sequence and recognize αXβ2 and αVβ3 by using VP7 during cell entry. J Virol. 2003; 77(18): 9969-9978. https://doi.org/10.1128/jvi.77.18.9969-9978.2003 PMID: 12941907

35. Barton ES, Forrest JC, Connolly JL, Chappell JD, Liu Y, Schnell FJ, et al. Junction adhesion molecule is a receptor for reovirus. Cell. 2001; 104(3): 441-451. https://doi.org/10.1016/s0092-8674(01)00231-8 PMID: 11239401

36. Maginnis MS, Forrest JC, Kopecky-Bromberg SA, Dickeson SK, Santoro SA, Zutter MM, et al. β1 integrin mediates internalization of mammalian reovirus. J Virol. 2006; 80(6): 2760-2770. https://doi.org/10.1128/JVI.80.6.2760-2770.2006 PMID: 16501085

37. Konopka-Anstadt JL, Mainou BA, Sutherland DM, Sekine Y, Strittmatter SM, Dermody TS. The Nogo receptor NgR1 mediates infection by mammalian reovirus. Cell Host Microbe. 2014; 15(6): 681-691. https://doi.org/10.1016/j.chom.2014.05.010 PMID: 24922571

38. Xiao C, Bator CM, Bowman VD, Rieder E, He Y, Hébert Bt, et al. Interaction of coxsackievirus A21 with its cellular receptor, ICAM-1. J Virol. 2001; 75(5): 2444-2451. https://doi.org/10.1128/JVI.75.5.2444-2451.2001 PMID: 11160747

39. Roivainen M, Piirainen L, Hovi T, Virtanen I, Riikonen T, Heino J, et al. Entry of coxsackievirus A9 into host cells: specific interactions with αvβ3 integrin, the vitronectin receptor. Virology. 1994; 203(2): 357-365. https://doi.org/10.1006/viro.1994.1494 PMID: 7519807

40. Williams ÇH, Kajander T, Hyypiä T, Jackson T, Sheppard D, Stanway G. Integrin αvβ6 is an RGD-dependent receptor for coxsackievirus A9. J Virol. 2004; 78(13): 6967-6973. https://doi.org/10.1128/JVI.78.13.6967-6973.2004 PMID: 15194773

41. Martino TA, Petric M, Weingartl H, Bergelson JM, Opavsky MA, Richardson CD, et al. The coxsackie-adenovirus receptor (CAR) is used by reference strains and clinical isolates representing all six serotypes of coxsackievirus group B and by swine vesicular disease virus. Virology. 2000; 271(1): 99-108. https://doi.org/10.1006/viro.2000.0324 PMID: 10814575

42. Meertens L, Carnec X, Lecoin MP, Ramdasi R, Guivel-Benhassine F, Lew E, et al. The TIM and TAM families of phosphatidylserine receptors mediate dengue virus entry. Cell Host Microbe. 2012; 12(4): 544-557. https://doi.org/10.1016/j.chom.2012.08.009 PMID: 23084921

43. Che P, Tang H, Li Q. The interaction between claudin-1 and dengue viral prM/M protein for its entry. Virology. 2013; 446(1-2): 303-313. https://doi.org/10.1016/j.virol.2013.08.009 PMID: 24074594

44. Thepparit C, Smith DR. Serotype-specific entry of dengue virus into liver cells: identification of the 37-kilodalton/67-kilodalton high-affinity laminin receptor as a dengue virus serotype 1 receptor. J Virol. 2004; 78(22): 12647-12656. https://doi.org/10.1128/JVI.78.22.12647-12656.2004 PMID: 15507651

45. Bergelson J, St John N, Kawaguchi S, Chan M, Stubdal H, Modlin J, et al. Infection by echoviruses 1 and 8 depends on the alpha 2 subunit of human VLA-2. J Virol. 1993; 67(11): 6847-6852. https://doi.org/10.1128/JVI.67.11.6847-6852.1993 PMID: 8411387

46. Bergelson JM, Chan M, Solomon KR, St John NF, Lin H, Finberg RW. Decay-accelerating factor (CD55), a glycosylphosphatidylinositol-anchored complement regulatory protein, is a receptor for several echoviruses. Proceedings of the National Academy of Sciences. 1994; 91(13): 6245-6248. https://doi.org/10.1073/pnas.91.13.6245 PMID: 7517044

47. Yang B, Chuang H, Yang KD. Sialylated glycans as receptor and inhibitor of enterovirus 71 infection to DLD-1 intestinal cells. Virol J. 2009; 6(1): 1-6. https://doi.org/10.1186/1743-422X-6-141 PMID: 19751532

48. Yamayoshi S, Yamashita Y, Li J, Hanagata N, Minowa T, Takemura T, et al. Scavenger receptor B2 is a cellular receptor for enterovirus 71. Nat Med. 2009; 15(7): 798-801. https://doi.org/10.1038/nm.1992 PMID: 19543282

49. Nishimura Y, Shimojima M, Tano Y, Miyamura T, Wakita T, Shimizu H. Human P-selectin glycoprotein ligand-1 is a functional receptor for enterovirus 71. Nat Med. 2009; 15(7): 794-797. https://doi.org/10.1038/nm.1961 PMID: 19543284

50. Feigelstock D, Thompson P, Mattoo P, Zhang Y, Kaplan GG. The human homolog of HAVcr-1 codes for a hepatitis A virus cellular receptor. J Virol. 1998; 72(8): 6621-6628. https://doi.org/10.1128/JVI.72.8.6621-6628.1998 PMID: 9658108

51. Agnello V, Ábel G, Elfahal M, Knight GB, Zhang Q-X. Hepatitis C virus and other flaviviridae viruses enter cells via low density lipoprotein receptor. Proceedings of the National Academy of Sciences. 1999; 96(22): 12766-12771. https://doi.org/10.1073/pnas.96.22.12766 PMID: 10535997

52. Evans MJ, von Hahn T, Tscherne DM, Syder AJ, Panis M, Wölk B, et al. Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry. Nature. 2007; 446(7137): 801-805. https://doi.org/10.1038/nature05654 PMID: 17325668

53. Zeisel MB, Koutsoudakis G, Schnober EK, Haberstroh A, Blum HE, Cosset FL, et al. Scavenger receptor class B type I is a key host factor for hepatitis C virus infection required for an entry step closely linked to CD81. Hepatology. 2007; 46(6): 1722-1731. https://doi.org/10.1002/hep.21994 PMID: 18000990

54. Yeager CL, Ashmun RA, Williams RK, Cardellichio CB, Shapiro LH, Look AT, et al. Human aminopeptidase N is a receptor for human coronavirus 229E. Nature. 1992; 357(6377): 420-422. https://doi.org/10.1038/357420a0 PMID: 1350662

55. Krempl C, Schultze B, Herrler G. Analysis of cellular receptors for human coronavirus OC43. Corona-and Related Viruses: Springer; 1995. p. 371-374.

56. Raj VS, Mou H, Smits SL, Dekkers DH, Müller MA, Dijkman R, et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. Nature. 2013; 495(7440): 251-254. https://doi.org/10.1038/nature12005 PMID: 23486063

57. Dimitrov DS. The secret life of ACE2 as a receptor for the SARS virus. Cell. 2003; 115(6): 652-653. https://doi.org/10.1016/s0092-8674(03)00976-0 PMID: 14675530

58. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell. 2020; 181(2): 271-280. e278. https://doi.org/10.1016/j.cell.2020.02.052 PMID: 32142651

59. Joki-Korpela P, Marjomäki V, Krogerus C, Heino J, Hyypiä T. Entry of human parechovirus 1. J Virol. 2001; 75(4): 1958-1967. https://doi.org/10.1128/jvi.75.4.1958-1967.2001 PMID: 11160695

60. Greve JM, Davis G, Meyer AM, Forte CP, Yost SC, Marlor CW, et al. The major human rhinovirus receptor is ICAM-1. Cell. 1989; 56(5): 839-847. https://doi.org/10.1016/0092-8674(89)90688-0 PMID: 2538243

61. Hofer F, Gruenberger M, Kowalski H, Machat H, Huettinger M, Kuechler E, et al. Members of the low density lipoprotein receptor family mediate cell entry of a minor-group common cold virus. Proc Natl Acad Sci U S A. 1994; 91(5): 1839-1842. https://doi.org/10.1073/pnas.91.5.1839 PMID: 8127891

62. Bochkov YA, Watters K, Ashraf S, Griggs TF, Devries MK, Jackson DJ, et al. Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. Proc Natl Acad Sci U S A. 2015; 112(17): 5485-5490. https://doi.org/10.1073/pnas.1421178112 PMID: 25848009

63. Huang P, Farkas T, Marionneau S, Zhong W, Ruvoën-Clouet N, Morrow AL, et al. Noroviruses bind to human ABO, Lewis, and secretor histo-blood group antigens: identification of 4 distinct strain-specific patterns. J Infect Dis. 2003; 188(1): 19-31. https://doi.org/10.1086/375742 PMID: 12825167

64. Mendelsohn CL, Wimmer E, Racaniello VR. Cellular receptor for poliovirus: molecular cloning, nucleotide sequence, and expression of a new member of the immunoglobulin superfamily. Cell. 1989; 56(5): 855-865. https://doi.org/10.1016/0092-8674(89)90690-9 PMID: 2538245

65. Cong H, Jiang Y, Tien P. Identification of the myelin oligodendrocyte glycoprotein as a cellular receptor for rubella virus. J Virol. 2011; 85(21): 11038-11047. https://doi.org/10.1128/jvi.05398-11 PMID: 21880773

66. Wang KS, Kuhn RJ, Strauss EG, Ou S, Strauss JH. High-affinity laminin receptor is a receptor for Sindbis virus in mammalian cells. J Virol. 1992; 66(8): 4992-5001. https://doi.org/10.1128/jvi.66.8.4992-5001.1992 PMID: 1385835

67. Ludwig GV, Kondig JP, Smith JF. A putative receptor for Venezuelan equine encephalitis virus from mosquito cells. J Virol. 1996; 70(8): 5592-5599. https://doi.org/10.1128/jvi.70.8.5592-5599.1996 PMID: 8764073

68. Schmidt K, Keller M, Bader BL, Korytář T, Finke S, Ziegler U, et al. Integrins modulate the infection efficiency of West Nile virus into cells. J Gen Virol. 2013; 94(Pt 8): 1723-1733. https://doi.org/10.1099/vir.0.052613-0 PMID: 23658209

69. Hamel R, Dejarnac O, Wichit S, Ekchariyawat P, Neyret A, Luplertlop N, et al. Biology of Zika Virus Infection in Human Skin Cells. J Virol. 2015; 89(17): 8880-8896. https://doi.org/10.1128/jvi.00354-15 PMID: 26085147

70. Shimojima M, Takada A, Ebihara H, Neumann G, Fujioka K, Irimura T, et al. Tyro3 family-mediated cell entry of Ebola and Marburg viruses. J Virol. 2006; 80(20): 10109-10116. https://doi.org/10.1128/jvi.01157-06 PMID: 17005688

71. Carette JE, Raaben M, Wong AC, Herbert AS, Obernosterer G, Mulherkar N, et al. Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. Nature. 2011; 477(7364): 340-343. https://doi.org/10.1038/nature10348 PMID: 21866103

72. Kondratowicz AS, Lennemann NJ, Sinn PL, Davey RA, Hunt CL, Moller-Tank S, et al. T-cell immunoglobulin and mucin domain 1 (TIM-1) is a receptor for Zaire Ebolavirus and Lake Victoria Marburgvirus. Proc Natl Acad Sci U S A. 2011; 108(20): 8426-8431. https://doi.org/10.1073/pnas.1019030108 PMID: 21536871

73. Radoshitzky SR, Kuhn JH, Spiropoulou CF, Albariño CG, Nguyen DP, Salazar-Bravo J, et al. Receptor determinants of zoonotic transmission of New World hemorrhagic fever arenaviruses. Proc Natl Acad Sci U S A. 2008; 105(7): 2664-2669. https://doi.org/10.1073/pnas.0709254105 PMID: 18268337

74. Gavrilovskaya IN, Shepley M, Shaw R, Ginsberg MH, Mackow ER. beta3 Integrins mediate the cellular entry of hantaviruses that cause respiratory failure. Proc Natl Acad Sci U S A. 1998; 95(12): 7074-7079. https://doi.org/10.1073/pnas.95.12.7074 PMID: 9618541

75. Bishop KA, Stantchev TS, Hickey AC, Khetawat D, Bossart KN, Krasnoperov V, et al. Identification of Hendra virus G glycoprotein residues that are critical for receptor binding. J Virol. 2007; 81(11): 5893-5901. https://doi.org/10.1128/jvi.02022-06 PMID: 17376907

76. Fujioka Y, Nishide S, Ose T, Suzuki T, Kato I, Fukuhara H, et al. A Sialylated Voltage-Dependent Ca(2+) Channel Binds Hemagglutinin and Mediates Influenza A Virus Entry into Mammalian Cells. Cell Host Microbe. 2018; 23(6): 809-818.e805. https://doi.org/10.1016/j.chom.2018.04.015 PMID: 29779930

77. Ng WC, Londrigan SL, Nasr N, Cunningham AL, Turville S, Brooks AG, et al. The C-type Lectin Langerin Functions as a Receptor for Attachment and Infectious Entry of Influenza A Virus. J Virol. 2016; 90(1): 206-221. https://doi.org/10.1128/jvi.01447-15 PMID: 26468543

78. Carroll SM, Higa HH, Paulson JC. Different cell-surface receptor determinants of antigenically similar influenza virus hemagglutinins. J Biol Chem. 1981; 256(16): 8357-8363 PMID: 6167577

79. Lugovtsev VY, Smith DF, Weir JP. Changes of the receptor-binding properties of influenza B virus B/Victoria/504/2000 during adaptation in chicken eggs. Virology.

2009; 394(2): 218-226. https://doi.org/10.1016/j.virol.2009.08.014 PMID: 19766280

80. Rogers GN, Herrler G, Paulson JC, Klenk HD. Influenza C virus uses 9-O-acetyl-N-acetylneuraminic acid as a high affinity receptor determinant for attachment to cells. J Biol Chem. 1986; 261(13): 5947-5951 PMID: 3700379

81. Fukushima K, Takahashi T, Ito S, Takaguchi M, Takano M, Kurebayashi Y, et al. Terminal sialic acid linkages determine different cell infectivities of human parainfluenza virus type 1 and type 3. Virology. 2014; 464-465: 424-431. https://doi.org/10.1016/j.virol.2014.07.033 PMID: 25146600

82. Kunz S. Receptor binding and cell entry of Old World arenaviruses reveal novel aspects of virus-host interaction. Virology. 2009; 387(2): 245-249. https://doi.org/10.1016/j.virol.2009.02.042 PMID: 19324387

83. Shimojima M, Kawaoka Y. Cell surface molecules involved in infection mediated by lymphocytic choriomeningitis virus glycoprotein. J Vet Med Sci. 2012; 74(10): 1363-1366. https://doi.org/10.1292/jvms.12-0176 PMID: 22673088

84. Goodman LB, Lyi SM, Johnson NC, Cifuente JO, Hafenstein SL, Parrish CR. Binding site on the transferrin receptor for the parvovirus capsid and effects of altered affinity on cell uptake and infection. J Virol. 2010; 84(10): 4969-4978. https://doi.org/10.1128/jvi.02623-09 PMID: 20200243

85. Tatsuo H, Ono N, Tanaka K, Yanagi Y. SLAM (CDw150) is a cellular receptor for measles virus. Nature. 2000; 406(6798): 893-897. https://doi.org/10.1038/35022579 PMID: 10972291

86. Noyce RS, Richardson CD. Nectin 4 is the epithelial cell receptor for measles virus. Trends Microbiol. 2012; 20(9): 429-439. https://doi.org/10.1016/j.tim.2012.05.006 PMID: 22721863

87. Negrete OA, Levroney EL, Aguilar HC, Bertolotti-Ciarlet A, Nazarian R, Tajyar S, et al. EphrinB2 is the entry receptor for Nipah virus, an emergent deadly paramyxovirus. Nature. 2005; 436(7049): 401-405. https://doi.org/10.1038/nature03838 PMID: 16007075

88. Negrete OA, Wolf MC, Aguilar HC, Enterlein S, Wang W, Mühlberger E, et al. Two key residues in ephrinB3 are critical for its use as an alternative receptor for Nipah

virus. PLoS Pathog. 2006; 2(2): e7. https://doi.org/10.1371/journal.ppat.0020007 PMID: 16477309

89. Guo Y, Duan M, Wang X, Gao J, Guan Z, Zhang M. Early events in rabies virus infection-Attachment, entry, and intracellular trafficking. Virus Res. 2019; 263: 217-225. https://doi.org/10.1016/j.virusres.2019.02.006 PMID: 30772332

90. Lozach PY, Kühbacher A, Meier R, Mancini R, Bitto D, Bouloy M, et al. DC-SIGN as a receptor for phleboviruses. Cell Host Microbe. 2011; 10(1): 75-88. https://doi.org/10.1016/j.chom.2011.06.007 PMID: 21767814

91. Finkelshtein D, Werman A, Novick D, Barak S, Rubinstein M. LDL receptor and its family members serve as the cellular receptors for vesicular stomatitis virus. Proc Natl Acad Sci U S A. 2013; 110(18): 7306-7311. https://doi.org/10.1073/pnas.1214441110 PMID: 23589850

92. Wilen CB, Tilton JC, Doms RW. HIV: cell binding and entry. Cold Spring Harb Perspect Med. 2012; 2(8). https://doi.org/10.1101/cshperspect.a006866 PMID: 22908191

93. Menéndez-Arias L, Sebastián-Martín A, Álvarez M. Viral reverse transcriptases. Virus Res. 2017; 234: 153-176. https://doi.org/10.1016/j.virusres.2016.12.019 PMID: 28043823

94. Trivedi V, Von Lindern J, Montes-Walters M, Rojo DR, Shell EJ, Parkin N, et al. Impact of human immunodeficiency virus type 1 reverse transcriptase inhibitor drug resistance mutation interactions on phenotypic susceptibility. AIDS Res Hum Retroviruses. 2008; 24(10): 1291-1300. https://doi.org/10.1089/aid.2007.0244 PMID: 18844463

95. Engelman AN, Singh PK. Cellular and molecular mechanisms of HIV-1 integration targeting. Cell Mol Life Sci. 2018; 75(14): 2491-2507. https://doi.org/10.1007/s00018-018-2772-5 PMID: 29417178

96. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021; 49(D1): D545-d551. https://doi.org/10.1093/nar/gkaa970 PMID: 33125081

97. Rheinemann L, Sundquist WI. Virus Budding. Encyclopedia of Virology. 2021: 519. https://doi.org/10.1016/B978-0-12-814515-9.00023-0

98. Bird SW, Kirkegaard K. Escape of non-enveloped virus from intact cells. Virology. 2015; 479-480: 444-449. https://doi.org/10.1016/j.virol.2015.03.044 PMID: 25890822

99. Votteler J, Sundquist WI. Virus budding and the ESCRT pathway. Cell Host Microbe. 2013; 14(3): 232-241. https://doi.org/10.1016/j.chom.2013.08.012 PMID: 24034610

100. Singh O, Su EC. Prediction of HIV-1 protease cleavage site using a combination of sequence, structural, and physicochemical features. BMC Bioinformatics. 2016; 17(Suppl 17): 478. https://doi.org/10.1186/s12859-016-1337-6 PMID: 28155640

101. Freed EO. HIV-1 assembly, release and maturation. Nature Reviews Microbiology. 2015; 13(8): 484-496. https://doi.org/10.1038/nrmicro3490 PMID: 26119571

102. Koyama S, Ishii KJ, Coban C, Akira S. Innate immune response to viral infection. Cytokine. 2008; 43(3): 336-341. https://doi.org/10.1016/j.cyto.2008.07.009 PMID: 18694646

103. Fensterl V, Sen GC. Interferons and viral infections. Biofactors. 2009; 35(1): 14-20. https://doi.org/10.1002/biof.6 PMID: 19319841

104. García-Sastre A. Ten strategies of interferon evasion by viruses. Cell Host Microbe. 2017; 22(2): 176-184. https://doi.org/10.1016/j.chom.2017.07.012 PMID: 28799903

105. Rönnblom L. The type I interferon system in the etiopathogenesis of autoimmune diseases. Ups J Med Sci. 2011; 116(4): 227-237. https://doi.org/10.3109/03009734.2011.624649 PMID: 22066971

106. Kino T, Slobodskaya O, Pavlakis GN, Chrousos GP. Nuclear receptor coactivator p160 proteins enhance the HIV-1 long terminal repeat promoter by bridging promoter-bound factors and the Tat-P-TEFb complex. J Biol Chem. 2002; 277(4): 2396-2405. https://doi.org/10.1074/jbc.M106312200 PMID: 11704662

107. Yang X, Gabuzda D. Regulation of human immunodeficiency virus type 1 infectivity by the ERK mitogen-activated protein kinase signaling pathway. J Virol. 1999; 73(4): 3460-3466. https://doi.org/10.1128/jvi.73.4.3460-3466.1999 PMID: 10074203

108. Muthumani K, Choo AY, Hwang DS, Premkumar A, Dayes NS, Harris C, et al. HIV-1 Nef-induced FasL induction and bystander killing requires p38 MAPK activation. Blood. 2005; 106(6): 2059-2068. https://doi.org/10.1182/blood-2005-03-0932 PMID: 15928037

109. Sanchez DJ, Miranda D, Jr., Marsden MD, Dizon TM, Bontemps JR, Davila SJ, et al. Disruption of Type I Interferon Induction by HIV Infection of T Cells. PLoS One. 2015; 10(9): e0137951. https://doi.org/10.1371/journal.pone.0137951 PMID: 26375588

110. Yim HC, Li JC, Lau JS, Lau AS. HIV-1 Tat dysregulation of lipopolysaccharide-induced cytokine responses: microbial interactions in HIV infection. AIDS. 2009; 23(12): 1473-1484. https://doi.org/10.1097/QAD.0b013e32832d7abe PMID: 19622906

111. Romero IA, Teixeira A, Strosberg AD, Cazaubon S, Couraud PO. The HIV-1 nef protein inhibits extracellular signal-regulated kinase-dependent DNA synthesis in a human astrocytic cell line. J Neurochem. 1998; 70(2): 778-785. https://doi.org/10.1046/j.1471-4159.1998.70020778.x PMID: 9453574

112. Kawasaki T, Kawai T. Toll-like receptor signaling pathways. Front Immunol. 2014; 5: 461. https://doi.org/10.3389/fimmu.2014.00461

113. Del Cornò M, Donninelli G, Varano B, Da Sacco L, Masotti A, Gessani S. HIV-1 gp120 activates the STAT3/interleukin-6 axis in primary human monocyte-derived dendritic cells. J Virol. 2014; 88(19): 11045-11055. https://doi.org/10.1128/jvi.00307-14 PMID: 25008924

114. Gangwani MR, Noel RJ, Jr., Shah A, Rivera-Amill V, Kumar A. Human immunodeficiency virus type 1 viral protein R (Vpr) induces CCL5 expression in astrocytes via PI3K and MAPK signaling pathways. J Neuroinflammation. 2013; 10: 136. https://doi.org/10.1186/1742-2094-10-136 PMID: 24225433

115. Herbein G, Varin A, Larbi A, Fortin C, Mahlknecht U, Fulop T, et al. Nef and TNFalpha are coplayers that favor HIV-1 replication in monocytic cells and primary macrophages. Curr HIV Res. 2008; 6(2): 117-129. https://doi.org/10.2174/157016208783884985 PMID: 18336259

116. Huang W, Rha GB, Chen L, Seelbach MJ, Zhang B, András IE, et al. Inhibition of telomerase activity alters tight junction protein expression and induces transendothelial migration of HIV-1-infected cells. Am J Physiol Heart Circ Physiol. 2010; 298(4): H1136-1145. https://doi.org/10.1152/ajpheart.01126.2009 PMID: 20139322

117. Miller S, Krijnse-Locker J. Modification of intracellular membrane structures for virus replication. Nat Rev Microbiol. 2008; 6(5): 363-374. https://doi.org/10.1038/nrmicro1890 PMID: 18414501

118. Wu JJ, Li W, Shao Y, Avey D, Fu B, Gillen J, et al. Inhibition of cGAS DNA Sensing by a Herpesvirus Virion Protein. Cell Host Microbe. 2015; 18(3): 333-344. https://doi.org/10.1016/j.chom.2015.07.015 PMID: 26320998

119. Nagasaka K, Kawana K, Osuga Y, Fujii T. PDZ domains and viral infection: versatile potentials of HPV-PDZ interactions in relation to malignancy. BioMed research international. 2013; 2013. https://doi.org/10.1155/2013/369712 PMID: 24093094

120. Lasso G, Honig B, Shapira SD. A Sweep of Earth's virome reveals host-Guided viral protein structural mimicry and points to Determinants of human disease. Cell Syst. 2021; 12(1): 82-91. e83. https://doi.org/10.1016/j.cels.2020.09.006 PMID: 33053371

121. Fu YZ, Su S, Gao YQ, Wang PP, Huang ZF, Hu MM, et al. Human Cytomegalovirus Tegument Protein UL82 Inhibits STING-Mediated Signaling to Evade Antiviral Immunity. Cell Host Microbe. 2017; 21(2): 231-243. https://doi.org/10.1016/j.chom.2017.01.001 PMID: 28132838

122. Nguyen NV, Tran JT, Sanchez DJ. HIV blocks Type I IFN signaling through disruption of STAT1 phosphorylation. Innate Immun. 2018; 24(8): 490-500. https://doi.org/10.1177/1753425918803674 PMID: 30282499

123. Frias-Staheli N, Giannakopoulos NV, Kikkert M, Taylor SL, Bridgen A, Paragas J, et al. Ovarian tumor domain-containing viral proteases evade ubiquitin- and ISG15-dependent innate immune responses. Cell Host Microbe. 2007; 2(6): 404-416. https://doi.org/10.1016/j.chom.2007.09.014 PMID: 18078692

124. Aguirre S, Maestre AM, Pagni S, Patel JR, Savage T, Gutman D, et al. DENV inhibits type I IFN production in infected cells by cleaving human STING. PLoS Pathog. 2012; 8(10): e1002934. https://doi.org/10.1371/journal.ppat.1002934 PMID: 23055924

125. Grant A, Ponia SS, Tripathi S, Balasubramaniam V, Miorin L, Sourisseau M, et al. Zika Virus Targets Human STAT2 to Inhibit Type I Interferon Signaling. Cell Host

Microbe. 2016; 19(6): 882-890. https://doi.org/10.1016/j.chom.2016.05.009 PMID: 27212660

126. Fonseca GJ, Thillainadesan G, Yousef AF, Ablack JN, Mossman KL, Torchia J, et al. Adenovirus evasion of interferon-mediated innate immunity by direct antagonism of a cellular histone posttranslational modification. Cell Host Microbe. 2012; 11(6): 597-606. https://doi.org/10.1016/j.chom.2012.05.005 PMID: 22704620

127. von Kobbe C, van Deursen JM, Rodrigues JP, Sitterlin D, Bachi A, Wu X, et al. Vesicular stomatitis virus matrix protein inhibits host cell gene expression by targeting the nucleoporin Nup98. Mol Cell. 2000; 6(5): 1243-1252. https://doi.org/10.1016/s1097-2765(00)00120-9 PMID: 11106761

128. Garaigorta U, Chisari FV. Hepatitis C virus blocks interferon effector function by inducing protein kinase R phosphorylation. Cell Host Microbe. 2009; 6(6): 513-522. https://doi.org/10.1016/j.chom.2009.11.004 PMID: 20006840

129. Xu W, Edwards MR, Borek DM, Feagins AR, Mittal A, Alinger JB, et al. Ebola virus VP24 targets a unique NLS binding site on karyopherin alpha 5 to selectively compete with nuclear import of phosphorylated STAT1. Cell Host Microbe. 2014; 16(2): 187-200. https://doi.org/10.1016/j.chom.2014.07.008 PMID: 25121748

130. Lin JS, Lai EM. Protein-Protein Interactions: Co-Immunoprecipitation. Methods Mol Biol. 2017; 1615: 211-219. https://doi.org/10.1007/978-1-4939-7033-9_17 PMID: 28667615

131. Louche A, Salcedo SP, Bigot S. Protein-Protein Interactions: Pull-Down Assays. Methods Mol Biol. 2017; 1615: 247-255. https://doi.org/10.1007/978-1-4939-7033-9_20 PMID: 28667618

132. Kim B. Western Blot Techniques. Methods Mol Biol. 2017; 1606: 133-139. https://doi.org/10.1007/978-1-4939-6990-6_9 PMID: 28501998

133. Arora B, Tandon R, Attri P, Bhatia R. Chemical Crosslinking: Role in Protein and Peptide Science. Curr Protein Pept Sci. 2017; 18(9): 946-955. https://doi.org/10.2174/1389203717666160724202806 PMID: 27455969

134. Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E, et al. The IntAct database: efficient access to fine-grained molecular interaction data. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gkab1006 PMID: 34761267

135. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019; 47(D1): D529-d541. https://doi.org/10.1093/nar/gky1079 PMID: 30476227

136. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Res. 2015; 43(Database issue): D583-587. https://doi.org/10.1093/nar/gku1121 PMID: 25392406

137. Calderone A, Licata L, Cesareni G. VirusMentha: a new resource for virus-host protein interactions. Nucleic Acids Res. 2015; 43(Database issue): D588-592. https://doi.org/10.1093/nar/gku830 PMID: 25217587

138. Cook HV, Doncheva NT, Szklarczyk D, von Mering C, Jensen LJ. Viruses.STRING: A Virus-Host Protein-Protein Interaction Database. Viruses. 2018; 10(10). https://doi.org/10.3390/v10100519 PMID: 30249048

139. Tarca AL, Carey VJ, Chen X-w, Romero R, Drăghici S. Machine learning and its applications to biology. PLoS Comput Biol. 2007; 3(6): e116. https://doi.org/10.1371/journal.pcbi.0030116 PMID: 17604446

140. Halder AK, Dutta P, Kundu M, Basu S, Nasipuri M. Review of computational methods for virus–host protein interaction prediction: a case study on novel Ebola–human interactions. Briefings in functional genomics. 2018; 17(6): 381-391. https://doi.org/10.1093/bfgp/elx026 PMID: 29028879

141. Behzadi P, Ranjbar R. DNA microarray technology and bioinformatic web services. Acta Microbiol Immunol Hung. 2019; 66(1): 19-30. https://doi.org/10.1556/030.65.2018.028 PMID: 30010394

142. Waters DL, Shapter FM. The polymerase chain reaction (PCR): general methods. Methods Mol Biol. 2014; 1099: 65-75. https://doi.org/10.1007/978-1-62703-715-0_7 PMID: 24243196

143. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nature Reviews Genetics. 2019; 20(11): 631-656. https://doi.org/10.1038/s41576-019-0150-2 PMID: 31341269

144. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, et al. GenBank. Nucleic Acids Res. 2021; 49(D1): D92-d96. https://doi.org/10.1093/nar/gkaa1023 PMID: 33196830

145. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res. 2021; 49(D1): D884-d891. https://doi.org/10.1093/nar/gkaa942 PMID: 33137190

146. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res. 2021; 49(D1): D1020-D1028. https://doi.org/10.1093/nar/gkaa1105 PMID: 33270901

147. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. 2021; 49(D1): D437-d451. https://doi.org/10.1093/nar/gkaa1038 PMID: 33211854

148. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 2017; 22(13): 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 PMID: 28382917

149. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, et al. Virus Variation Resource–improved response to emergent viral outbreaks. Nucleic Acids Res. 2017; 45(D1): D482-D490. https://doi.org/10.1093/nar/gkw1065 PMID: 27899678

150. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020; 48(D1): D498-D503. https://doi.org/10.1093/nar/gkz1031 PMID: 31691815

151. Licata L, Lo Surdo P, Iannuccelli M, Palma A, Micarelli E, Perfetto L, et al. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. Nucleic Acids Res. 2020; 48(D1): D504-d510. https://doi.org/10.1093/nar/gkz949 PMID: 31665520

152. Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020; 369(6509): 1318-1330. https://doi.org/10.1126/science.aaz1776 PMID: 32913098

153. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020; 48(D1): D296-D306. https://doi.org/10.1093/nar/gkz1030 PMID: 31680160

154. Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, Aykac-Fas B, et al. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res. 2020; 48(D1): D269-d276. https://doi.org/10.1093/nar/gkz975 PMID: 31713636

155. Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 2020; 48(D1): D376-d382. https://doi.org/10.1093/nar/gkz1064 PMID: 31724711

156. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021; 49(D1): D412-D419. https://doi.org/10.1093/nar/gkaa913 PMID: 33125078

157. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020; 48(D1): D265-d268. https://doi.org/10.1093/nar/gkz991 PMID: 31777944

158. Mosca R, Céol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2014; 42(Database issue): D374-379. https://doi.org/10.1093/nar/gkt887 PMID: 24081580

159. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012; 40(Database issue): D857-861. https://doi.org/10.1093/nar/gkr930 PMID: 22096227

160. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021; 49(D1): D605-d612. https://doi.org/10.1093/nar/gkaa1074 PMID: 33237311

161. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res. 2016: gkw985. https://doi.org/10.1093/nar/gkw985 PMID: 27794551

162. Tranmer M, Elliot M. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR). 2008; 5(5): 1-5

163. Marill KA. Advanced statistics: linear regression, part II: multiple linear regression. Acad Emerg Med. 2004; 11(1): 94-102. https://doi.org/10.1197/j.aem.2003.09.006 PMID: 14709437

164. Rath S, Tripathy A, Tripathy AR. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2020; 14(5): 1467-1474. https://doi.org/10.1016/j.dsx.2020.07.045 PMID: 32771920

165. Ma R, Zheng X, Wang P, Liu H, Zhang C. The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method. Sci Rep. 2021; 11(1): 1-14. https://doi.org/10.1038/s41598-021-97037-5 PMID: 34465820

166. Piepho HP. A coefficient of determination (R2) for generalized linear mixed models. Biom J. 2019; 61(4): 860-872. https://doi.org/10.1002/bimj.201800270 PMID: 30957911

167. Zhou Z-H. A brief introduction to weakly supervised learning. National science review. 2018; 5(1): 44-53. https://doi.org/10.1093/nsr/nwx106

168. Hu G, Katuwawala A, Wang K, Wu Z, Ghadermarzi S, Gao J, et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Commun. 2021; 12(1): 1-8. https://doi.org/10.1038/s41467-021-24773-7 PMID: 34290238

169. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nature reviews molecular cell biology. 2007; 8(12): 995-1005. https://doi.org/10.1038/nrm2281 PMID: 18037900

170. Chen H, Li F, Wang L, Jin Y, Chi C-H, Kurgan L, et al. Systematic evaluation of machine learning methods for identifying human–pathogen protein–protein interactions. Brief Bioinform. 2021; 22(3): bbaa068. https://doi.org/10.1093/bib/bbaa068 PMID: 32459334

171. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals. 2020; 139: 110059. https://doi.org/10.1016/j.chaos.2020.110059 PMID: 32834612

172. Dyer MD, Murali T, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. Infect Genet Evol. 2011; 11(5): 917-923. https://doi.org/10.1016/j.meegid.2011.02.022 PMID: 21382517

173. Tastan O, Qi Y, Carbonell JG, Klein-Seetharaman J. Prediction of interactions between HIV-1 and human proteins by information integration. Biocomputing 2009: World Scientific; 2009. p. 516-527.

174. Mei S. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. PLoS One. 2013; 8(11): e79606. https://doi.org/10.1371/journal.pone.0079606 PMID: 24260261

175. Dey L, Chakraborty S, Mukhopadhyay A. Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins. Biomedical journal. 2020; 43(5): 438-450. https://doi.org/10.1016/j.bj.2020.08.003 PMID: 33036956

176. Wu J, Zhang P, Zhang L, Meng W, Li J, Tong C, et al. Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. MedRxiv. 2020. https://doi.org/10.1101/2020.04.02.20051136

177. Sun L, Song F, Shi N, Liu F, Li S, Li P, et al. Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. J Clin Virol. 2020; 128: 104431. https://doi.org/10.1016/j.jcv.2020.104431 PMID: 32442756

178. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med. 2020; 121: 103795. https://doi.org/10.1016/j.compbiomed.2020.103795 PMID: 32568676

179. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med. 2020; 121: 103792. https://doi.org/10.1016/j.compbiomed.2020.103792 PMID: 32568675

180. Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China. PLoS Negl Trop Dis. 2017; 11(10): e0005973. https://doi.org/10.1371/journal.pntd.0005973 PMID: 29036169

181. Cui G, Fang C, Han K, editors. Prediction of protein-protein interactions between viruses and human by an SVM model. BMC Bioinformatics; 2012: Springer.

182. Li J, Zhang S, Li B, Hu Y, Kang X-P, Wu X-Y, et al. Machine learning methods for predicting human-adaptive influenza A viruses based on viral nucleotide compositions. Mol Biol Evol. 2020; 37(4): 1224-1236. https://doi.org/10.1093/molbev/msz276 PMID: 31750915

183. Jiang D, Hao M, Ding F, Fu J, Li M. Mapping the transmission risk of Zika virus using machine learning models. Acta Trop. 2018; 185: 391-399. https://doi.org/10.1016/j.actatropica.2018.06.021 PMID: 29932934

184. Barman RK, Saha S, Das S. Prediction of interactions between viral and host proteins using supervised machine learning methods. PLoS One. 2014; 9(11): e112034. https://doi.org/10.1371/journal.pone.0112034 PMID: 25375323

185. Nusinovici S, Tham YC, Yan MYC, Ting DSW, Li J, Sabanayagam C, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. J Clin Epidemiol. 2020; 122: 56-69. https://doi.org/10.1016/j.jclinepi.2020.03.002 PMID: 32169597

186. Jostins L, McVean G. Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. Bioinformatics. 2016; 32(12): 1898-1900. https://doi.org/10.1093/bioinformatics/btw075 PMID: 26873930

187. Bužić D, Dobša J, editors. Lyrics classification using naive bayes. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); 2018: IEEE.

188. Zhang H, Wei H, Tang Y, Pu Q, editors. Research on classification of scientific and technological documents based on Naive Bayes. Proceedings of the 2019 11th International Conference on Machine Learning and Computing; 2019.

189. Zhang S, Cheng D, Deng Z, Zong M, Deng X. A novel kNN algorithm with data-driven k parameter computation. Pattern Recognition Letters. 2018; 109: 44-54. https://doi.org/10.1016/j.patrec.2017.09.036

190. Hu L-Y, Huang M-W, Ke S-W, Tsai C-F. The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus. 2016; 5(1): 1-9. https://doi.org/10.1186/s40064-016-2941-7

191. Ooi H-L, Ng S-C, Lim E. Ano detection with k-nearest neighbor using minkowski distance. International Journal of Signal Processing Systems. 2013; 1(2): 208-211. https://doi.org/10.12720/ijsps.1.2.208-211

192. Hassanat AB, Abbadi MA, Altarawneh GA, Alhasanat AA. Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. arXiv preprint arXiv:14090919. 2014:

193. Song Y, Huang J, Zhou D, Zha H, Giles CL, editors. Iknn: Informative k-nearest neighbor pattern classification. European Conference on Principles of Data Mining and Knowledge Discovery; 2007: Springer.

194. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI). 2012; 9(5): 272

195. Farid DM, Zhang L, Rahman CM, Hossain MA, Strachan R. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert systems with applications. 2014; 41(4): 1937-1946. https://doi.org/10.1016/j.eswa.2013.08.089

196. Zhang J, Chai H, Gao B, Yang G, Ma Z. HEMEsPred: Structure-based ligand-specific heme binding residues prediction by using fast-adaptive ensemble learning scheme. IEEE/ACM Trans Comput Biol Bioinform. 2016; 15(1): 147-156. https://doi.org/10.1109/TCBB.2016.2615010 PMID: 28029626

197. Wahid MF, Tafreshi R, Langari R. A Multi-Window Majority Voting Strategy to Improve Hand Gesture Recognition Accuracies Using Electromyography Signal. IEEE Trans Neural Syst Rehabil Eng. 2020; 28(2): 427-436. https://doi.org/10.1109/tnsre.2019.2961706 PMID: 31870989

198. Prajapati GL, Patle A, editors. On performing classification using SVM with radial basis and polynomial kernel functions. 2010 3rd International Conference on Emerging Trends in Engineering and Technology; 2010: IEEE.

199. Zhang Y, Li J, Hong M, Man Y. Intelligent approaches to forecast the chemical property: Case study in papermaking process. Applications of Artificial Intelligence in Process Systems Engineering: Elsevier; 2021. p. 93-118.

200. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011; 2(3): 1-27. https://doi.org/10.1145/1961189.1961199

201. Chai H, Gu Q, Hughes J, Robertson DL. In silico prediction of HIV-1-host molecular interactions and their directionality. PLoS Comput Biol. 2022; 18(2): e1009720. https://doi.org/10.1371/journal.pcbi.1009720 PMID: 35134057

202. Wong T-T, Yeh P-Y. Reliable accuracy estimates from k-fold cross validation. IEEE Transactions on Knowledge and Data Engineering. 2019; 32(8): 1586-1594. https://doi.org/10.1109/TKDE.2019.2912815

203. Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. Bioinformatics. 2019; 35(14): i343-i353. https://doi.org/10.1093/bioinformatics/btz324 PMID: 31510679

204. Brandenberg OF, Magnus C, Regoes RR, Trkola A. The HIV-1 entry process: a stoichiometric view. Trends Microbiol. 2015; 23(12): 763-774. https://doi.org/10.1016/j.tim.2015.09.003 PMID: 26541228

205. Lusic M, Siliciano RF. Nuclear landscape of HIV-1 infection and integration. Nat Rev Microbiol. 2017; 15(2): 69-82. https://doi.org/10.1038/nrmicro.2016.162 PMID: 27941817

206. Deeks SG, Overbaugh J, Phillips A, Buchbinder S. HIV infection. Nature reviews Disease primers. 2015; 1(1): 1-22. https://doi.org/10.1038/nrdp.2015.35 PMID: 27188527

207. Molle D, Maiuri P, Boireau S, Bertrand E, Knezevich A, Marcello A, et al. A real-time view of the TAR: Tat: P-TEFb complex at HIV-1 transcription sites. Retrovirology. 2007; 4(1): 1-5. https://doi.org/10.1186/1742-4690-4-36 PMID: 17537237

208. Debaisieux S, Rayne F, Yezid H, Beaumelle B. The ins and outs of HIV-1 Tat. Traffic. 2012; 13(3): 355-363. https://doi.org/10.1111/j.1600-0854.2011.01286.x PMID: 21951552

209. Malim MH, Emerman M. HIV-1 accessory proteins—ensuring viral survival in a hostile environment. Cell Host Microbe. 2008; 3(6): 388-398. https://doi.org/10.1016/j.chom.2008.04.008 PMID: 18541215

210. Seelamgari A, Maddukuri A, Berro R, de la Fuente C, Kehn K, Deng L, et al. Role of viral regulatory and accessory proteins in HIV-1 replication. Front Biosci. 2004; 9(9): 2388-2413. https://doi.org/10.2741/1403 PMID: 15353294

211. Balachandran A, Wong R, Stoilov P, Pan S, Blencowe B, Cheung P, et al. Identification of small molecule modulators of HIV-1 Tat and Rev protein accumulation. Retrovirology. 2017; 14(1): 1-21. https://doi.org/10.1186/s12977-017-0330-0 PMID: 28122580

212. Meyerson NR, Rowley PA, Swan CH, Le DT, Wilkerson GK, Sawyer SL. Positive selection of primate genes that promote HIV-1 replication. Virology. 2014; 454: 291-298. https://doi.org/10.1016/j.virol.2014.02.029 PMID: 24725956

213. Towers GJ, Noursadeghi M. Interactions between HIV-1 and the cell-autonomous innate immune system. Cell Host Microbe. 2014; 16(1): 10-18. https://doi.org/10.1016/j.chom.2014.06.009 PMID: 25011104

214. Valera M-S, de Armas-Rillo L, Barroso-González J, Ziglio S, Batisse J, Dubois N, et al. The HDAC6/APOBEC3G complex regulates HIV-1 infectiveness by inducing Vif autophagic degradation. Retrovirology. 2015; 12(1): 1-26. https://doi.org/10.1186/s12977-015-0181-5 PMID: 26105074

215. Shoji-Kawata S, Zhong Q, Kameoka M, Iwabu Y, Sapsutthipas S, Luftig RB, et al. The RING finger ubiquitin ligase RNF125/TRAC-1 down-modulates HIV-1 replication in primary human peripheral blood mononuclear cells. Virology. 2007; 368(1): 191-204. https://doi.org/10.1016/j.virol.2007.06.028 PMID: 17643463

216. Doyle T, Goujon C, Malim MH. HIV-1 and interferons: who's interfering with whom? Nat Rev Microbiol. 2015; 13(7): 403-413. https://doi.org/10.1038/nrmicro3449 PMID: 25915633

217. MacPherson JI, Dickerson JE, Pinney JW, Robertson DL. Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. PLoS Comput Biol. 2010; 6(7): e1000863. https://doi.org/10.1371/journal.pcbi.1000863 PMID: 20686668

218. Engelman A, Cherepanov P. The structural biology of HIV-1: mechanistic and therapeutic insights. Nat Rev Microbiol. 2012; 10(4): 279-290. https://doi.org/10.1038/nrmicro2747 PMID: 22421880

219. Dorr P, Westby M, Dobbs S, Griffin P, Irvine B, Macartney M, et al. Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. Antimicrobial Agents and Chemotherapy. 2005; 49(11): 4721-4732. https://doi.org/10.1128/AAC.49.11.4721-4732.2005 PMID: 16251317

220. Pinney JW, Dickerson JE, Fu W, Sanders-Beer BE, Ptak RG, Robertson DL. HIV–host interactions: a map of viral perturbation of the host system. AIDS. 2009; 23(5): 549-554. https://doi.org/10.1097/QAD.0b013e328325a495 PMID: 19262354

221. Dickerson JE, Pinney JW, Robertson DL. The biological context of HIV-1 host interactions reveals subtle insights into a system hijack. BMC Syst Biol. 2010; 4(1): 1-13. https://doi.org/10.1186/1752-0509-4-80 PMID: 20529270

222. Chen K-C, Wang T-Y, Chan C-h. Associations between HIV and human pathways revealed by protein-protein interactions and correlated gene expression profiles. PLoS One. 2012; 7(3): e34240. https://doi.org/10.1371/journal.pone.0034240 PMID: 22479575

223. Chen L, Keppler OT, Schölz C. Post-translational modification-based regulation of HIV replication. Front Microbiol. 2018; 9: 2131. https://doi.org/10.3389/fmicb.2018.02131 PMID: 30254620

224. Qi Y, Tastan O, Carbonell JG, Klein-Seetharaman J, Weston J. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. Bioinformatics. 2010; 26(18): i645-i652. https://doi.org/10.1093/bioinformatics/btq394 PMID: 20823334

225. Mukhopadhyay A, Maulik U, Bandyopadhyay S. A novel biclustering approach to association rule mining for predicting HIV-1–human protein interactions. PLoS One. 2012; 7(4): e32289. https://doi.org/10.1371/journal.pone.0032289 PMID: 22539940

226. Mukhopadhyay A, Ray S, Maulik U. Incorporating the type and direction information in predicting novel regulatory interactions between HIV-1 and human proteins using a biclustering approach. BMC Bioinformatics. 2014; 15(1): 1-22. https://doi.org/10.1186/1471-2105-15-26 PMID: 24460683

227. Doolittle JM, Gomez SM. Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. Virol J. 2010; 7(1): 1-15. https://doi.org/10.1186/1743-422X-7-82 PMID: 20426868

228. Evans P, Dampier W, Ungar L, Tozeren A. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC Med Genomics. 2009; 2(1): 1-13. https://doi.org/10.1186/1755-8794-2-27 PMID: 19450270

229. Becerra A, Bucheli VA, Moreno PA. Prediction of virus-host protein-protein interactions mediated by short linear motifs. BMC Bioinformatics. 2017; 18(1): 1-11. https://doi.org/10.1186/s12859-017-1570-7 PMID: 28279163

230. Nourani E, Khunjush F, Durmuş S. Computational approaches for prediction of pathogen-host protein-protein interactions. Front Microbiol. 2015; 6: 94. https://doi.org/10.3389/fmicb.2015.00094 PMID: 25759684

231. Durmuş S, Çakır T, Özgür A, Guthke R. A review on computational systems biology of pathogen–host interactions. Front Microbiol. 2015; 6: 235. https://doi.org/10.3389/fmicb.2015.00235 PMID: 25914674

232. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely high mutation rate of HIV-1 in vivo. PLoS Biol. 2015; 13(9): e1002251. https://doi.org/10.1371/journal.pbio.1002251 PMID: 26375597

233. Gao G, Wu X, Zhou J, He M, He JJ, Guo D. Inhibition of HIV-1 transcription and replication by a newly identified cyclin T1 splice variant. J Biol Chem. 2013; 288(20): 14297-14309. https://doi.org/10.1074/jbc.M112.438465 PMID: 23569210

234. Okada H, Zhang X, Fofana IB, Nagai M, Suzuki H, Ohashi T, et al. Synergistic effect of human CycT1 and CRM1 on HIV-1 propagation in rat T cells and macrophages. Retrovirology. 2009; 6(1): 1-12. https://doi.org/10.1186/1742-4690-6-43 PMID: 19435492

235. Kwon Y, Kaake RM, Echeverria I, Suarez M, Shamsabadi MK, Stoneham C, et al. Structural basis of CD4 downregulation by HIV-1 Nef. Nat Struct Mol Biol. 2020; 27(9): 822-828. https://doi.org/10.1038/s41594-020-0463-z PMID: 32719457

236. Jette CA, Barnes CO, Kirk SM, Melillo B, Smith AB, Bjorkman PJ. Cryo-EM structures of HIV-1 trimer bound to CD4-mimetics BNM-III-170 and M48U1 adopt a CD4-bound open conformation. Nat Commun. 2021; 12(1): 1-10. https://doi.org/10.1038/s41467-021-21816-x PMID: 33782388

237. Singha S, Shenoy PP. An adaptive heuristic for feature selection based on complementarity. Machine Learning. 2018; 107(12): 2027-2071. https://doi.org/10.1007/s10994-018-5728-y

238. Yeom S, Giacomelli I, Fredrikson M, Jha S, editors. Privacy risk in machine learning: Analyzing the connection to overfitting. 2018 IEEE 31st Computer Security Foundations Symposium (CSF); 2018: IEEE.

239. Ying X, editor An overview of overfitting and its solutions. Journal of Physics: Conference Series; 2019: IOP Publishing.

240. Lever J, Krzywinski M, Altman N. Points of significance: model selection and overfitting. Nature Publishing Group; 2016.

241. Ako-Adjei D, Fu W, Wallin C, Katz KS, Song G, Darji D, et al. HIV-1, human interaction database: current status and new features. Nucleic Acids Res. 2015; 43(D1): D566-D570. https://doi.org/10.1093/nar/gku1126 PMID: 25378338

242. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames. org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2019; 47(D1): D786-D792. https://doi.org/10.1093/nar/gky930 PMID: 30304474

243. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44(D1): D733-D745. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804

244. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020; 48(D1): D682-D688. https://doi.org/10.1093/nar/gkz966 PMID: 31691826

245. Palasca O, Santos A, Stolte C, Gorodkin J, Jensen LJ. TISSUES 2.0: an integrative web resource on mammalian tissue expression. Database. 2018; 2018. https://doi.org/10.1093/database/bay028 PMID: 30403794

246. Consortium GO. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; 47(D1): D330-D338. https://doi.org/10.1093/nar/gky1055 PMID: 30395331

247. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28(23): 3150-3152. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610

248. Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. 2007; 8(10): 1-13. https://doi.org/10.1186/gb-2007-8-10-r209 PMID: 17916239

249. King CR, Mehle A. The later stages of viral infection: An undiscovered country of host dependency factors. PLoS Pathog. 2020; 16(8): e1008777. https://doi.org/10.1371/journal.ppat.1008777 PMID: 32841303

250. Martinelli E, Cicala C, Van Ryk D, Goode DJ, Macleod K, Arthos J, et al. HIV-1 gp120 inhibits TLR9-mediated activation and IFN-α secretion in plasmacytoid dendritic cells. Proceedings of the National Academy of Sciences. 2007; 104(9): 3396-3401. https://doi.org/10.1073/pnas.0611353104 PMID: 17360657

251. Taylor HE, Khatua AK, Popik W. The innate immune factor apolipoprotein L1 restricts HIV-1 infection. J Virol. 2014; 88(1): 592-603. https://doi.org/10.1128/JVI.02828-13 PMID: 24173214

252. Liu S, Wang Q, Yu X, Li Y, Guo Y, Liu Z, et al. HIV-1 inhibition in cells with CXCR4 mutant genome created by CRISPR-Cas9 and piggyBac recombinant technologies. Sci Rep. 2018; 8(1): 1-11. https://doi.org/10.1038/s41598-018-26894-4 PMID: 29872154

253. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML. APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Res. 2018; 46(D1): D213-D217. https://doi.org/10.1093/nar/gkx997 PMID: 29069475

254. Gordon DE, Watson A, Roguev A, Zheng S, Jang GM, Kane J, et al. A quantitative genetic interaction map of HIV infection. Mol Cell. 2020; 78(2): 197-209. https://doi.org/10.1016/j.molcel.2020.02.004 PMID: 32084337

255. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456(7221): 470-476. https://doi.org/10.1038/nature07509 PMID: 18978772

256. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews Genetics. 2010; 11(5): 345-355. https://doi.org/10.1038/nrg2776 PMID: 20376054

257. Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature. 2017; 550(7674): 124-127. https://doi.org/10.1038/nature24039 PMID: 28953888

258. Yu C-H, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol Cell. 2015; 59(5): 744-754. https://doi.org/10.1016/j.molcel.2015.07.018 PMID: 26321254

259. Guéguen L, Duret L. Unbiased estimate of synonymous and nonsynonymous substitution rates with nonstationary base composition. Mol Biol Evol. 2018; 35(3): 734-742. https://doi.org/10.1093/molbev/msx308 PMID: 29220511

260. Betts MJ, Russell RB. Amino acid properties and consequences of substitutions. Bioinformatics for geneticists. 2003; 317: 289. https://doi.org/10.1002/0470867302.ch14

261. Vens C, Rosso M-N, Danchin EG. Identifying discriminative classification-based motifs in biological sequences. Bioinformatics. 2011; 27(9): 1231-1238. https://doi.org/10.1093/bioinformatics/btr110 PMID: 21372086

262. Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, et al. ELM 2016—data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res. 2016; 44(D1): D294-D300. https://doi.org/10.1093/nar/gkv1291 PMID: 26615199

263. Noble WS. How does multiple testing correction work? Nat Biotechnol. 2009; 27(12): 1135-1137. https://doi.org/10.1038/nbt1209-1135 PMID: 20010596

264. Li A, Barber RF. Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2019; 81(1): 45-74. https://doi.org/10.1111/rssb.12298

265. Liu X-Y, Wu J, Zhou Z-H. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern. 2008; 39(2): 539-550. https://doi.org/10.1109/TSMCB.2008.2007853 PMID: 19095540

266. Walsh I, Martin AJ, Di Domenico T, Tosatto SC. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics. 2012; 28(4): 503-509. https://doi.org/10.1093/bioinformatics/btr682 PMID: 22190692

267. Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018; 46(W1): W329-W337. https://doi.org/10.1093/nar/gky384 PMID: 29860432

268. Xue B, Mizianty MJ, Kurgan L, Uversky VN. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol Life Sci. 2012; 69(8): 1211-1259. https://doi.org/10.1007/s00018-011-0859-3 PMID: 22033837

269. King DF, Siddiqui AA, Buffa V, Fischetti L, Gao Y, Stieh D, et al. Mucosal tissue tropism and dissemination of HIV-1 subtype B acute envelope-expressing chimeric

virus. J Virol. 2013; 87(2): 890-899. https://doi.org/10.1128/JVI.02216-12 PMID: 23135721

270. Bet A, Maze EA, Bansal A, Sterrett S, Gross A, Graff-Dubois S, et al. The HIV-1 antisense protein (ASP) induces CD8 T cell responses during chronic infection. Retrovirology. 2015; 12(1): 1-13. https://doi.org/10.1186/s12977-015-0135-y PMID: 25809376

271. Ahmed H, Howton T, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS. Network biology discovers pathogen contact points in host protein-protein interactomes. Nat Commun. 2018; 9(1): 1-13. https://doi.org/10.1038/s41467-018-04632-8 PMID: 29899369

272. Cafarelli T, Desbuleux A, Wang Y, Choi SG, De Ridder D, Vidal M. Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. Curr Opin Struct Biol. 2017; 44: 201-210. https://doi.org/10.1016/j.sbi.2017.05.003 PMID: 28575754

273. Doncheva NT, Assenov Y, Domingues FS, Albrecht M. Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc. 2012; 7(4): 670. https://doi.org/10.1038/nprot.2012.004 PMID: 22422314

274. Liu Z. A method of SVM with normalization in intrusion detection. Procedia Environmental Sciences. 2011; 11: 256-262. https://doi.org/10.1016/j.proenv.2011.12.040

275. Babbar R, Schölkopf B. Data scarcity, robustness and extreme multi-label classification. Machine Learning. 2019; 108(8): 1329-1351. https://doi.org/10.1007/s10994-019-05791-5

276. Cheng Q, Zhou H, Cheng J. The fisher-markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. IEEE Trans Pattern Anal Mach Intell. 2010; 33(6): 1217-1233. https://doi.org/10.1109/TPAMI.2010.195 PMID: 21493968

277. Chai H, Zhang J. Identification of mammalian enzymatic proteins based on sequence-derived features and species-specific scheme. IEEE Access. 2018; 6: 8452-8458. https://doi.org/10.1109/ACCESS.2018.2798284

278. Lee C-Y, Chen B-S. Mutually-exclusive-and-collectively-exhaustive feature selection scheme. Applied Soft Computing. 2018; 68: 961-971. https://doi.org/10.1016/j.asoc.2017.04.055

279. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020; 21(1): 1-13. https://doi.org/10.1186/s12864-019-6413-7 PMID: 31898477

280. Gautier VW, Gu L, O'Donoghue N, Pennington S, Sheehy N, Hall WW. In vitro nuclear interactome of the HIV-1 Tat protein. Retrovirology. 2009; 6(1): 1-18. https://doi.org/10.1186/1742-4690-6-47 PMID: 19454010

281. Fang M, Xu N, Shao X, Yang J, Wu N, Yao H. Inhibitory effects of human immunodeficiency virus gp120 and Tat on CpG-A-induced inflammatory cytokines in plasmacytoid dendritic cells. Acta Biochim Biophys Sin. 2012; 44(9): 797-804. https://doi.org/10.1093/abbs/gms062 PMID: 22814248

282. Barr SD, Smiley JR, Bushman FD. The interferon response inhibits HIV particle production by induction of TRIM22. PLoS Pathog. 2008; 4(2): e1000007. https://doi.org/10.1371/journal.ppat.1000007 PMID: 18389079

283. Gao D, Wu J, Wu Y-T, Du F, Aroh C, Yan N, et al. Cyclic GMP-AMP synthase is an innate immune sensor of HIV and other retroviruses. Science. 2013; 341(6148): 903-906. https://doi.org/10.1126/science.1240933 PMID: 23929945

284. Mangino G, Percario ZA, Fiorucci G, Vaccari G, Manrique S, Romeo G, et al. In vitro treatment of human monocytes/macrophages with myristoylated recombinant Nef of human immunodeficiency virus type 1 leads to the activation of mitogen-activated protein kinases, IκB kinases, and interferon regulatory factor 3 and to the release of beta interferon. J Virol. 2007; 81(6): 2777-2791. https://doi.org/10.1128/JVI.01640-06 PMID: 17182689

285. Harman AN, Nasr N, Feetham A, Galoyan A, Alshehri AA, Rambukwelle D, et al. HIV blocks interferon induction in human dendritic cells and macrophages by dysregulation of TBK1. J Virol. 2015; 89(13): 6575-6584. https://doi.org/10.1128/JVI.00889-15 PMID: 25855743

286. Bego MG, Côté É, Aschman N, Mercier J, Weissenhorn W, Cohen ÉA. Vpu exploits the cross-talk between BST2 and the ILT7 receptor to suppress anti-HIV-1 responses

by plasmacytoid dendritic cells. PLoS Pathog. 2015; 11(7): e1005024. https://doi.org/10.1371/journal.ppat.1005024 PMID: 26172439

287. MacPherson JI, Dickerson JE, Pinney JW, Robertson DL. Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. PLoS Comput Biol. 2010; 6(7): e1000863

288. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008; 40(12): 1413-1415. https://doi.org/10.1038/ng.259 PMID: 18978789

289. Hittinger CT, Carroll SB. Gene duplication and the adaptive evolution of a classic genetic switch. Nature. 2007; 449(7163): 677-681. https://doi.org/10.1038/nature06151 PMID: 17928853

290. Freilich S, Massingham T, Blanc E, Goldovsky L, Thornton JM. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. Genome Biol. 2006; 7(10): 1-14. https://doi.org/10.1186/gb-2006-7-10-r89 PMID: 17029626

291. Pearson WR. Finding protein and nucleotide similarities with FASTA. Curr Protoc Bioinformatics. 2016; 53(1): 3-9. https://doi.org/10.1002/0471250953.bi0309s04 PMID: 18428723

292. Davey NE, Cyert MS, Moses AM. Short linear motifs–ex nihilo evolution of protein regulation. Cell Commun Signal. 2015; 13(1): 1-15. https://doi.org/10.1186/s12964-015-0120-z PMID: 26589632

293. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015; 16(1): 18-29. https://doi.org/10.1038/nrm3920 PMID: 25531225

294. Maetschke SR, Simonsen M, Davis MJ, Ragan MA. Gene Ontology-driven inference of protein–protein interactions using inducers. Bioinformatics. 2012; 28(1): 69-75. https://doi.org/10.1093/bioinformatics/btr610 PMID: 22057159

295. Wang X, Ye L, Hou W, Zhou Y, Wang Y-J, Metzger DS, et al. Cellular microRNA expression correlates with susceptibility of monocytes/macrophages to HIV-1 infection. Blood. 2009; 113(3): 671-674. https://doi.org/10.1182/blood-2008-09-175000 PMID: 19015395

296. Persidsky Y, Gendelman HE. Mononuclear phagocyte immunity and the neuropathogenesis of HIV-1 infection. J Leukoc Biol. 2003; 74(5): 691-701. https://doi.org/10.1189/jlb.0503205 PMID: 14595004

297. Alexaki A, Wigdahl B. HIV-1 infection of bone marrow hematopoietic progenitor cells and their role in trafficking and viral dissemination. PLoS Pathog. 2008; 4(12): e1000215. https://doi.org/10.1371/journal.ppat.1000215 PMID: 19112504

298. Caby F. CD4+/CD8+ ratio restoration in long-term treated HIV-1-infected individuals. AIDS. 2017; 31(12): 1685-1695. https://doi.org/10.1097/QAD.0000000000001533 PMID: 28700392

299. Boettler T, Spangenberg HC, Neumann-Haefelin C, Panther E, Urbani S, Ferrari C, et al. T cells with a CD4+ CD25+ regulatory phenotype suppress in vitro proliferation of virus-specific CD8+ T cells during chronic hepatitis C virus infection. J Virol. 2005; 79(12): 7860-7867. https://doi.org/10.1128/JVI.79.12.7860-7867.2005 PMID: 15919940

300. Puntel M, Barrett R, Sanderson NS, Kroeger KM, Bondale N, Wibowo M, et al. Identification and visualization of CD8+ T cell mediated IFN-γ signaling in target cells during an antiviral immune response in the brain. PLoS One. 2011; 6(8): e23523. https://doi.org/10.1371/journal.pone.0023523 PMID: 21897844

301. Mohanapriya M, Lekha J, editors. Comparative study between decision tree and knn of data mining classification technique. Journal of Physics: Conference Series; 2018: IOP Publishing.

302. Han S, Kim H, Lee Y-S. Double random forest. Machine Learning. 2020; 109(8): 1569-1586. https://doi.org/10.1007/s10994-020-05889-1

303. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009; 5(12): e1000605. https://doi.org/10.1371/journal.pcbi.1000605 PMID: 20011109

304. Churchill MJ, Deeks SG, Margolis DM, Siliciano RF, Swanstrom R. HIV reservoirs: what, where and how to target them. Nat Rev Microbiol. 2016; 14(1): 55-60. https://doi.org/10.1038/nrmicro.2015.5 PMID: 26616417

305. Harlalka GV, Baple EL, Cross H, Kühnle S, Cubillos-Rojas M, Matentzoglu K, et al. Mutation of HERC2 causes developmental delay with Angelman-like features. J

Med Genet. 2013; 50(2): 65-73. https://doi.org/10.1136/jmedgenet-2012-101367 PMID: 23243086

306. Sathasivam K, Neueder A, Gipson TA, Landles C, Benjamin AC, Bondulich MK, et al. Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. Proceedings of the National Academy of Sciences. 2013; 110(6): 2366-2370. https://doi.org/10.1073/pnas.1221891110 PMID: 23341618

307. Rose M, Schubert C, Dierichs L, Gaisa NT, Heer M, Heidenreich A, et al. OASIS/CREB3L1 is epigenetically silenced in human bladder cancer facilitating tumor cell spreading and migration in vitro. Epigenetics. 2014; 9(12): 1626-1640. https://doi.org/10.4161/15592294.2014.988052 PMID: 25625847

308. Khan HA, Margulies CE. The role of mammalian Creb3-like transcription factors in response to nutrients. Front Genet. 2019; 10: 591. https://doi.org/10.3389/fgene.2019.00591 PMID: 31293620

309. Qiu J, Liang T, Wu J, Yu F, He X, Tian Y, et al. N-Substituted Pyrrole Derivative 12m Inhibits HIV-1 Entry by Targeting Gp41 of HIV-1 Envelope Glycoprotein. Front Pharmacol. 2019; 10. https://doi.org/10.3389/fphar.2019.00859 PMID: 31427969

310. Jordan IK, Wolf YI, Koonin EV. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol. 2004; 4(1): 1-11. https://doi.org/10.1186/1471-2148-4-22 PMID: 15238160

311. Carmel L, Rogozin IB, Wolf YI, Koonin EV. Evolutionarily conserved genes preferentially accumulate introns. Genome Res. 2007; 17(7): 1045-1050. https://doi.org/10.1101/gr.5978207 PMID: 17495009

312. Wibmer CK, Gorman J, Ozorowski G, Bhiman JN, Sheward DJ, Elliott DH, et al. Structure and recognition of a novel HIV-1 gp120-gp41 interface antibody that caused MPER exposure through viral escape. PLoS Pathog. 2017; 13(1): e1006074. https://doi.org/10.1371/journal.ppat.1006074 PMID: 28076415

313. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45(D1): D353-D361. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662

314. Mostafavi S, Yoshida H, Moodley D, LeBoité H, Rothamel K, Raj T, et al. Parsing the interferon transcriptional network and its disease associations. Cell. 2016; 164(3): 564-578. https://doi.org/10.1016/j.cell.2015.12.032 PMID: 26824662

315. Shaw AE, Hughes J, Gu Q, Behdenna A, Singer JB, Dennis T, et al. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. PLoS Biol. 2017; 15(12): e2004086. https://doi.org/10.1371/journal.pbio.2004086 PMID: 29253856

316. Shalhoub S. Interferon beta-1b for COVID-19. The Lancet. 2020; 395(10238): 1670-1671. https://doi.org/10.1016/S0140-6736(20)31101-6 PMID: 32401712

317. Harris BD, Schreiter J, Chevrier M, Jordan JL, Walter MR. Human interferon-ϵ and interferon-κ exhibit low potency and low affinity for cell-surface IFNAR and the poxvirus antagonist B18R. J Biol Chem. 2018; 293(41): 16057-16068. https://doi.org/10.1074/jbc.RA118.003617 PMID: 30171073

318. Li S-f, Zhao F-r, Shao J-j, Xie Y-l, Chang H-y, Zhang Y-g. Interferon-omega: Current status in clinical applications. Int Immunopharmacol. 2017; 52: 253-260. https://doi.org/10.1016/j.intimp.2017.08.028 PMID: 28957693

319. Kak G, Raza M, Tiwari BK. Interferon-gamma (IFN-γ): exploring its implications in infectious diseases. Biomol Concepts. 2018; 9(1): 64-79. https://doi.org/10.1515/bmc-2018-0007 PMID: 29856726

320. Hemann EA, Gale Jr M, Savan R. Interferon lambda genetics and biology in regulation of viral control. Front Immunol. 2017; 8: 1707. https://doi.org/10.3389/fimmu.2017.01707 PMID: 29270173

321. Schneider WM, Chevillotte MD, Rice CM. Interferon-stimulated genes: a complex web of host defenses. Annu Rev Immunol. 2014; 32: 513-545. https://doi.org/10.1146/annurev-immunol-032713-120231 PMID: 24555472

322. Kotenko SV, Durbin JE. Contribution of type III interferons to antiviral immunity: location, location, location. J Biol Chem. 2017; 292(18): 7295-7303. https://doi.org/10.1074/jbc.R117.777102 PMID: 28289095

323. Lazear HM, Schoggins JW, Diamond MS. Shared and distinct functions of type I and type III interferons. Immunity. 2019; 50(4): 907-923. https://doi.org/10.1016/j.immuni.2019.03.025 PMID: 30995506

324. Takaoka A, Yanai H. Interferon signalling network in innate defence. Cell Microbiol. 2006; 8(6): 907-922. https://doi.org/10.1111/j.1462-5822.2006.00716.x PMID: 16681834

325. Stark GR, Darnell Jr JE. The JAK-STAT pathway at twenty. Immunity. 2012; 36(4): 503-514. https://doi.org/10.1016/j.immuni.2012.03.013 PMID: 22520844

326. Schoggins JW. Interferon-stimulated genes: what do they all do? Annu Rev Virol. 2019; 6: 567-584. https://doi.org/10.1146/annurev-virology-092818-015756 PMID: 31283436

327. Aso H, Ito J, Koyanagi Y, Sato K. Comparative description of the expression profile of interferon-stimulated genes in multiple cell lineages targeted by HIV-1 infection. Front Microbiol. 2019; 10: 429. https://doi.org/10.3389/fmicb.2019.00429 PMID: 30915053

328. Dang W, Xu L, Yin Y, Chen S, Wang W, Hakim MS, et al. IRF-1, RIG-I and MDA5 display potent antiviral activities against norovirus coordinately induced by different types of interferons. Antiviral Res. 2018; 155: 48-59. https://doi.org/10.1016/j.antiviral.2018.05.004 PMID: 29753657

329. Masola V, Bellin G, Gambaro G, Onisto M. Heparanase: A multitasking protein involved in extracellular matrix (ECM) remodeling and intracellular events. Cells. 2018; 7(12): 236. https://doi.org/10.3390/cells7120236 PMID: 30487472

330. Schoggins JW. Recent advances in antiviral interferon-stimulated gene biology. F1000Research. 2018; 7. https://doi.org/10.12688/f1000research.12450.1 PMID: 29568506

331. Spence JS, He R, Hoffmann H-H, Das T, Thinon E, Rice CM, et al. IFITM3 directly engages and shuttles incoming virus particles to lysosomes. Nat Chem Biol. 2019; 15(3): 259-268. https://doi.org/10.1038/s41589-018-0213-2 PMID: 30643282

332. Haller O, Staeheli P, Schwemmle M, Kochs G. Mx GTPases: dynamin-like antiviral machines of innate immunity. Trends Microbiol. 2015; 23(3): 154-163. https://doi.org/10.1016/j.tim.2014.12.003 PMID: 25572883

333. Giotis ES, Robey RC, Skinner NG, Tomlinson CD, Goodbourn S, Skinner MA. Chicken interferome: avian interferon-stimulated genes identified by microarray and RNA-seq of primary chick embryo fibroblasts treated with a chicken type I interferon

(IFN-α). Vet Res. 2016; 47(1): 1-12. https://doi.org/10.1186/s13567-016-0363-8 PMID: 27494935

334. Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, et al. Interferome v2. 0: an updated database of annotated interferon-regulated genes. Nucleic Acids Res. 2012; 41(D1): D1040-D1046. https://doi.org/10.1093/nar/gks1215 PMID: 23203888

335. OhAinle M, Helms L, Vermeire J, Roesch F, Humes D, Basom R, et al. A virus-packageable CRISPR screen identifies host factors mediating interferon inhibition of HIV. Elife. 2018; 7: e39823. https://doi.org/10.7554/eLife.39823 PMID: 30520725

336. Zhang Y, Burke CW, Ryman KD, Klimstra WB. Identification and characterization of interferon-induced proteins that inhibit alphavirus replication. J Virol. 2007; 81(20): 11246-11255. https://doi.org/10.1128/JVI.01282-07 PMID: 17686841

337. Pamela C, Kanchwala M, Liang H, Kumar A, Wang L-F, Xing C, et al. The IFN response in bats displays distinctive IFN-stimulated gene expression kinetics with atypical RNASEL induction. The Journal of Immunology. 2018; 200(1): 209-217. https://doi.org/10.4049/jimmunol.1701214 PMID: 29180486

338. Feld JJ, Nanda S, Huang Y, Chen W, Cam M, Pusek SN, et al. Hepatic gene expression during treatment with peginterferon and ribavirin: Identifying molecular pathways for treatment response. Hepatology. 2007; 46(5): 1548-1563. https://doi.org/10.1002/hep.21853 PMID: 17929300

339. Dalman MR, Deeter A, Nimishakavi G, Duan Z-H, editors. Fold change and p-value cutoffs significantly alter microarray interpretations. BMC Bioinformatics; 2012: BioMed Central.

340. Trilling M, Bellora N, Rutkowski AJ, de Graaf M, Dickinson P, Robertson K, et al. Deciphering the modulation of gene expression by type I and II interferons combining 4sU-tagging, translational arrest and in silico promoter analysis. Nucleic Acids Res. 2013; 41(17): 8107-8125. https://doi.org/10.1093/nar/gkt589 PMID: 23832230

341. Yu X, Liu H, Hamel KA, Morvan MG, Yu S, Leff J, et al. Dorsal root ganglion macrophages contribute to both the initiation and persistence of neuropathic pain. Nat Commun. 2020; 11(1): 1-12. https://doi.org/10.1038/s41467-019-13839-2 PMID: 31937758

342. Chen Y, Lun AT, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research. 2016; 5. https://doi.org/10.12688/f1000research.8987.2 PMID: 27508061

343. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl comparative genomics resources. Database. 2016; 2016: bav096. https://doi.org/10.1093/database/bav096 PMID: 26896847

344. Li HD, Menon R, Omenn GS, Guan Y. Revisiting the identification of canonical splice isoforms through integration of functional genomics and proteomics evidence. Proteomics. 2014; 14(23-24): 2709-2718. https://doi.org/10.1002/pmic.201400170 PMID: 25265570

345. Bragg JG, Potter S, Bi K, Moritz C. Exon capture phylogenomics: efficacy across scales of divergence. Mol Ecol Resour. 2016; 16(5): 1059-1068. https://doi.org/10.1111/1755-0998.12449 PMID: 26215687

346. Sieber P, Platzer M, Schuster S. The definition of open reading frame revisited. Trends Genet. 2018; 34(3): 167-170. https://doi.org/10.1016/j.tig.2017.12.009 PMID: 29366605

347. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. Selection in the evolution of gene duplications. Genome Biol. 2002; 3(2): 1-9. https://doi.org/10.1186/gb-2002-3-2-research0008 PMID: 11864370

348. Esposito M, Moreno-Hagelsieb G. Non-synonymous to synonymous substitutions suggest that orthologs tend to keep their functions, while paralogs are a source of functional novelty. bioRxiv. 2018: 354704. https://doi.org/10.1101/354704

349. Sievers A, Bosiek K, Bisch M, Dreessen C, Riedel J, Froß P, et al. K-mer content, correlation, and position analysis of genome DNA sequences for the identification of function and evolutionary features. Genes. 2017; 8(4): 122. https://doi.org/10.3390/genes8040122 PMID: 28422050

350. Zhou Z, Dang Y, Zhou M, Li L, Yu C-h, Fu J, et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceedings of the National Academy of Sciences. 2016; 113(41): E6117-E6125. https://doi.org/10.1073/pnas.1606724113 PMID: 27671647

351. Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit. 2004; 17(1): 17-32. https://doi.org/10.1002/jmr.647 PMID: 14872534

352. Tufarelli C, Ahmad A, Strohbuecker S, Scotti C, Sottile V. In Silico Identification of SOX1 Post-Translational Modifications Highlights a Shared Protein Motif. 2020. https://doi.org/10.3390/cells9112471 PMID: 33202879

353. Yoon J, Blumer A, Lee K. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. Bioinformatics. 2006; 22(24): 3106-3108. https://doi.org/10.1093/bioinformatics/btl533 PMID: 17060356

354. Friedel CC, Zimmer R. Influence of degree correlations on network structure and stability in protein-protein interaction networks. BMC Bioinformatics. 2007; 8(1): 1-10. https://doi.org/10.1186/1471-2105-8-297 PMID: 17688687

355. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L. Hierarchical organization of modularity in metabolic networks. Science. 2002; 297(5586): 1551-1555. https://doi.org/10.1126/science.1073374 PMID: 12202830

356. Hagai T, Azia A, Babu MM, Andino R. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. Cell Rep. 2014; 7(5): 1729-1739. https://doi.org/10.1016/j.celrep.2014.04.052 PMID: 24882001

357. Jović A, Brkić K, Bogunović N, editors. A review of feature selection methods with applications. 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO); 2015: Ieee.

358. MacFarland TW, Yates JM. Mann–whitney u test. Introduction to nonparametric statistics for the biological sciences using R: Springer; 2016. p. 103-132.

359. Van den Eynden J, Larsson E. Mutational signatures are critical for proper estimation of purifying selection pressures in cancer somatic mutation data when using the dN/dS metric. Front Genet. 2017; 8: 74. https://doi.org/10.3389/fgene.2017.00074 PMID: 28642787

360. Song H, Bremer BJ, Hinds EC, Raskutti G, Romero PA. Inferring protein sequence-function relationships with large-scale positive-unlabeled learning. Cell Syst. 2020. https://doi.org/10.1016/j.cels.2020.10.007 PMID: 33212013

361. Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol Evol. 2012; 4(7): 675-682. https://doi.org/10.1093/gbe/evs052 PMID: 22628461

362. Lee NK, Li X, Wang D. A comprehensive survey on genetic algorithms for DNA motif prediction. Inf Sci. 2018; 466: 25-43. https://doi.org/10.1016/j.ins.2018.07.004

363. Di Rienzo L, Miotto M, Bò L, Ruocco G, Raimondo D, Milanetti E. Characterizing hydropathy of amino acid side chain in a protein environment by investigating the structural changes of water molecules network. Front Mol Biosci. 2021; 8. https://doi.org/10.3389/fmolb.2021.626837 PMID: 33718433

364. Bhadra P, Yan J, Li J, Fong S, Siu SW. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci Rep. 2018; 8(1): 1-10. https://doi.org/10.1038/s41598-018-19752-w PMID: 29374199

365. Pfleger CM, Kirschner MW. The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. Genes Dev. 2000; 14(6): 655-665 PMID: 10733526

366. Fehr AR, Yu D. Control the host cell cycle: viral regulation of the anaphase-promoting complex. J Virol. 2013; 87(16): 8818-8825. https://doi.org/10.1128/JVI.00088-13 PMID: 23760246

367. Bösl K, Ianevski A, Than TT, Andersen PI, Kuivanen S, Teppor M, et al. Common nodes of virus–host interaction revealed through an integrated network analysis. Front Immunol. 2019; 10: 2186. https://doi.org/10.3389/fimmu.2019.02186 PMID: 31636628

368. Michael S, Travé G, Ramu C, Chica C, Gibson TJ. Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. Bioinformatics. 2008; 24(4): 453-457. https://doi.org/10.1093/bioinformatics/btm624 PMID: 18184688

369. Abedi M, Gheisari Y. Nodes with high centrality in protein interaction networks are responsible for driving signaling pathways in diabetic nephropathy. PeerJ. 2015; 3: e1284. https://doi.org/10.7717/peerj.1284 PMID: 26557424

370. Ozato K, Shin D-M, Chang T-H, Morse HC. TRIM family proteins and their emerging roles in innate immunity. Nat Rev Immunol. 2008; 8(11): 849-860. https://doi.org/10.1038/nri2413

371. Shaw AE, Rihn SJ, Mollentze N, Wickenhagen A, Stewart DG, Orton RJ, et al. The antiviral state has shaped the CpG composition of the vertebrate interferome to avoid self-targeting. PLoS Biol. 2021; 19(9): e3001352. https://doi.org/10.1371/journal.pbio.3001352 PMID: 34491982

372. Zhang M-L, Zhou Z-H. ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition. 2007; 40(7): 2038-2048. https://doi.org/10.1016/j.patcog.2006.12.019

373. Cheng D, Zhang S, Deng Z, Zhu Y, Zong M, editors. kNN algorithm with data-driven k value. International Conference on Advanced Data Mining and Applications; 2014: Springer.

374. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nat Genet. 2013; 45(6): 580-585. https://doi.org/10.1038/ng.2653

375. Papatheodorou I, Moreno P, Manning J, Fuentes AM-P, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Res. 2020; 48(D1): D77-D83. https://doi.org/10.1093/nar/gkz947 PMID: 31665515

376. Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001; 411(6833): 41-42. https://doi.org/10.1038/35075138 PMID: 11333967

377. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 2005; 22(4): 803-806. https://doi.org/10.1093/molbev/msi072 PMID: 15616139

378. Batada NN, Hurst LD, Tyers M. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol. 2006; 2(7): e88. https://doi.org/10.1371/journal.pcbi.0020088 PMID: 16839197

379. Pérez-Martínez D. Innate immunity in vertebrates: an overview. Immunology. 2016; 148(2): 125-139. https://doi.org/10.1111/imm.12597 PMID: 26878338

380. Jopling CL. Mutations: Stop that nonsense! Elife. 2014; 3: e04300. https://doi.org/10.7554/eLife.04300

381. Zhu X, Pribis JP, Rodriguez PC, Morris Jr SM, Vodovotz Y, Billiar TR, et al. The central role of arginine catabolism in T-cell dysfunction and increased susceptibility

to infection after physical injury. Ann Surg. 2014; 259(1): 171-178. https://doi.org/10.1097/SLA.0b013e31828611f8 PMID: 23470573

382. Morris CR, Hamilton-Reeves J, Martindale RG, Sarav M, Ochoa Gautier JB. Acquired amino acid deficiencies: a focus on arginine and glutamine. Nutr Clin Pract. 2017; 32: 30S-47S. https://doi.org/10.1177/0884533617691250 PMID: 28388380

383. Levring TB, Hansen AK, Nielsen BL, Kongsbak M, Von Essen MR, Woetmann A, et al. Activated human CD4+ T cells express transporters for both cysteine and cystine. Sci Rep. 2012; 2(1): 1-6. https://doi.org/10.1038/srep00266 PMID: 22355778

384. Sikalidis AK. Amino acids and immune response: a role for cysteine, glutamine, phenylalanine, tryptophan and arginine in T-cell function and cancer? Pathol Oncol Res. 2015; 21(1): 9-17. https://doi.org/10.1007/s12253-014-9860-0 PMID: 25351939

385. Yin C, Zheng T, Chang X. Biosynthesis of S-Adenosylmethionine by magnetically immobilized Escherichia coli cells highly expressing a methionine adenosyltransferase variant. Molecules. 2017; 22(8): 1365. https://doi.org/10.3390/molecules22081365 PMID: 28820476

386. Feld JJ, Modi AA, El–Diwany R, Rotman Y, Thomas E, Ahlenstiel G, et al. S-adenosyl methionine improves early viral responses and interferon-stimulated gene induction in hepatitis C nonresponders. Gastroenterology. 2011; 140(3): 830-839. https://doi.org/10.1053/j.gastro.2010.09.010 PMID: 20854821

387. Li S-W, Lai C-C, Ping J-F, Tsai F-J, Wan L, Lin Y-J, et al. Severe acute respiratory syndrome coronavirus papain-like protease suppressed alpha interferon-induced responses through downregulation of extracellular signal-regulated kinase 1-mediated signalling pathways. J Gen Virol. 2011; 92(5): 1127-1140. https://doi.org/10.1099/vir.0.028936-0 PMID: 21270289

388. Flo TH, Smith KD, Sato S, Rodriguez DJ, Holmes MA, Strong RK, et al. Lipocalin 2 mediates an innate immune response to bacterial infection by sequestrating iron. Nature. 2004; 432(7019): 917-921. https://doi.org/10.1038/nature03104 PMID: 15531878

389. Tissot C, Rebouissou C, Klein B, Mechti N. Both human α/β and γ interferons upregulate the expression of CD48 cell surface molecules. J Interferon Cytokine Res. 1997; 17(1): 17-26. https://doi.org/10.1089/jir.1997.17.17 PMID: 9041467

390. Noçon AL, Ip JP, Terry R, Lim SL, Getts DR, Müller M, et al. The bacteriostatic protein lipocalin 2 is induced in the central nervous system of mice with West Nile virus encephalitis. J Virol. 2014; 88(1): 679-689. https://doi.org/10.1128/JVI.02094-13 PMID: 24173226

391. Zarama A, Perez-Carmona N, Farre D, Tomic A, Borst EM, Messerle M, et al. Cytomegalovirus m154 hinders CD48 cell-surface expression and promotes viral escape from host natural killer cell control. PLoS Pathog. 2014; 10(3): e1004000. https://doi.org/10.1371/journal.ppat.1004000 PMID: 24626474

392. Martínez-Vicente P, Farré D, Engel P, Angulo A. Divergent Traits and Ligand-Binding Properties of the Cytomegalovirus CD48 Gene Family. Viruses. 2020; 12(8): 813. https://doi.org/10.3390/v12080813 PMID: 32731344

393. Ricquier D. UCP1, the mitochondrial uncoupling protein of brown adipocyte: a personal contribution and a historical perspective. Biochimie. 2017; 134: 3-8. https://doi.org/10.1016/j.biochi.2016.10.018 PMID: 27916641

394. Alexandersen S, Chamings A, Bhatta TR. SARS-CoV-2 genomic and subgenomic RNAs in diagnostic samples are not an indicator of active replication. Nat Commun. 2020; 11(1): 1-13. https://doi.org/10.1038/s41467-020-19883-7 PMID: 33247099

395. Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, et al. The origins of SARS-CoV-2: a critical review. Cell. 2021. https://doi.org/10.1016/j.cell.2021.08.017 PMID: 34480864

396. Meng B, Kemp SA, Papa G, Datir R, Ferreira IA, Marelli S, et al. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B. 1.1. 7. Cell Rep. 2021; 35(13): 109292. https://doi.org/10.1016/j.celrep.2021.109292 PMID: 34166617

397. Gidari A, Sabbatini S, Bastianelli S, Pierucci S, Busti C, Monari C, et al. Cross-neutralization of SARS-CoV-2 B. 1.1. 7 and P. 1 variants in vaccinated, convalescent and P. 1 infected. J Infect. 2021; 83(4): 467-472. https://doi.org/10.1016/j.jinf.2021.07.019 PMID: 34320390

398. Mlcochova P, Kemp SA, Dhar MS, Papa G, Meng B, Ferreira IA, et al. SARS-CoV-2 B. 1.617. 2 Delta variant replication and immune evasion. Nature. 2021: 1-6. https://doi.org/10.1038/s41586-021-03944-y PMID: 34488225

399. Kroidl I, Mecklenburg I, Schneiderat P, Müller K, Girl P, Wölfel R, et al. Vaccine breakthrough infection and onward transmission of SARS-CoV-2 Beta (B. 1.351) variant, Bavaria, Germany, February to March 2021. Eurosurveillance. 2021; 26(30): 2100673. https://doi.org/10.2807/1560-7917.ES.2021.26.30.2100673 PMID: 34328074

400. Chakraborty C, Sharma AR, Bhattacharya M, Agoramoorthy G, Lee S-S. Evolution, mode of transmission, and mutational landscape of newly emerging SARS-CoV-2 variants. Mbio. 2021; 12(4): e01140-01121. https://doi.org/10.1128/mBio.01140-21 PMID: 34465019

401. Zhao X, Zheng A, Li D, Zhang R, Sun H, Wang Q, et al. Neutralisation of ZF2001-elicited antisera to SARS-CoV-2 variants. The Lancet Microbe. 2021; 2(10): e494. https://doi.org/10.1016/S2666-5247(21)00217-2 PMID: 34458880

402. Wink PL, Volpato FCZ, Monteiro FL, Willig JB, Zavascki AP, Barth AL, et al. First identification of SARS-CoV-2 Lambda (C. 37) variant in southern Brazil. Infect Control Hosp Epidemiol. 2021: 1-2. https://doi.org/10.1017/ice.2021.390 PMID: 34470685

403. Messali S, Bertelli A, Campisi G, Zani A, Ciccozzi M, Caruso A, et al. A cluster of the new SARS-CoV-2 B. 1.621 lineage in Italy and sensitivity of the viral isolate to the BNT162b2 vaccine. J Med Virol. 2021. https://doi.org/10.1002/jmv.27247 PMID: 34329486

404. Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. The Lancet. 2021. https://doi.org/10.1016/S0140-6736(21)02758-67 PMID: 34871545

405. Chen J, Wang R, Wang M, Wei G-W. Mutations strengthened SARS-CoV-2 infectivity. J Mol Biol. 2020; 432(19): 5212-5226. https://doi.org/10.1016/j.jmb.2020.07.009 PMID: 32710986

406. Arya R, Kumari S, Pandey B, Mistry H, Bihani SC, Das A, et al. Structural insights into SARS-CoV-2 proteins. J Mol Biol. 2021; 433(2): 166725. https://doi.org/10.1016/j.jmb.2020.11.024 PMID: 33245961

407. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020; 583(7816): 459-468. https://doi.org/10.1038/s41586-020-2286-9 PMID: 32353859

408. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, et al. The coding capacity of SARS-CoV-2. Nature. 2021; 589(7840): 125-130. https://doi.org/10.1038/s41586-020-2739-1 PMID: 32906143

409. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. Cell. 2020; 181(4): 914-921. e910. https://doi.org/10.1016/j.cell.2020.04.011 PMID: 32330414

410. Suryawanshi RK, Koganti R, Agelidis A, Patil CD, Shukla D. Dysregulation of cell signaling by SARS-CoV-2. Trends Microbiol. 2020. https://doi.org/10.1016/j.tim.2020.12.007 PMID: 33451855

411. Xia H, Cao Z, Xie X, Zhang X, Chen JY-C, Wang H, et al. Evasion of type I interferon by SARS-CoV-2. Cell Rep. 2020; 33(1): 108234. https://doi.org/10.1016/j.celrep.2020.108234 PMID: 32979938

412. Lei X, Dong X, Ma R, Wang W, Xiao X, Tian Z, et al. Activation and evasion of type I interferon responses by SARS-CoV-2. Nat Commun. 2020; 11(1): 1-12. https://doi.org/10.1038/s41467-020-17665-9 PMID: 32733001

413. Anand P, Puranik A, Aravamudan M, Venkatakrishnan A, Soundararajan V. SARS-CoV-2 strategically mimics proteolytic activation of human ENaC. Elife. 2020; 9: e58603. https://doi.org/10.7554/eLife.58603 PMID: 32452762

414. Tompa P, Davey NE, Gibson TJ, Babu MM. A million peptide motifs for the molecular biologist. Mol Cell. 2014; 55(2): 161-169. https://doi.org/10.1016/j.molcel.2014.05.032 PMID: 25038412

415. Uversky VN. The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins. Intrinsically disordered proteins. 2013; 1(1): e24684. https://doi.org/10.4161/idp.24684 PMID: 28516010

416. Uversky VN. The intrinsic disorder alphabet. III. Dual personality of serine. Intrinsically disordered proteins. 2015; 3(1): e1027032. https://doi.org/10.1080/21690707.2015.1027032 PMID: 28232888

417. Neduva V, Russell RB. Linear motifs: evolutionary interaction switches. FEBS Lett. 2005; 579(15): 3342-3345. https://doi.org/10.1016/j.febslet.2005.04.005 PMID: 15943979

418. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, et al. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci. 2008; 13(6580): 603. https://doi.org/10.2741/3175 PMID: 18508681

419. Kadaveru K, Vyas J, Schiller MR. Viral infection and human disease-insights from minimotifs. Frontiers in bioscience: a journal and virtual library. 2008; 13: 6455. https://doi.org/10.2741/3166 PMID: 18508672

420. Duffy S. Why are RNA virus mutation rates so damn high? PLoS Biol. 2018; 16(8): e3000003. https://doi.org/10.1371/journal.pbio.3000003 PMID: 30102691

421. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology. 2021; 19(7): 409-424. https://doi.org/10.1038/s41579-021-00573-0 PMID: 34075212

422. Lu R, Yang P, Padmakumar S, Misra V. The herpesvirus transactivator VP16 mimics a human basic domain leucine zipper protein, luman, in its interaction with HCF. J Virol. 1998; 72(8): 6291-6297. https://doi.org/10.1128/JVI.72.8.6291-6297.1998 PMID: 9658067

423. Akhova O, Bainbridge M, Misra V. The neuronal host cell factor-binding protein Zhangfei inhibits herpes simplex virus replication. J Virol. 2005; 79(23): 14708-14718. https://doi.org/10.1128/JVI.79.23.14708-14718.2005 PMID: 16282471

424. Sobolev B, Poroikov V, Olenina L, Kolesanova E, Archakov A. Comparative analysis of amino acid sequences from envelope proteins isolated from different hepatitis C virus variants: possible role of conservative and variable regions. J Viral Hepat. 2000; 7(5): 368-374. https://doi.org/10.1046/j.1365-2893.2000.00242.x PMID: 10971825

425. Han Z, Sagum CA, Bedford MT, Sidhu SS, Sudol M, Harty RN. ITCH E3 ubiquitin ligase interacts with Ebola virus VP40 to regulate budding. J Virol. 2016; 90(20): 9163-9171. https://doi.org/10.1128/JVI.01078-16 PMID: 27489272

426. Min J-Y, Li S, Sen GC, Krug RM. A site on the influenza A virus NS1 protein mediates both inhibition of PKR activation and temporal regulation of viral RNA

synthesis. Virology. 2007; 363(1): 236-243. https://doi.org/10.1016/j.virol.2007.01.038 PMID: 17320139

427. Lee S, Joshi A, Nagashima K, Freed EO, Hurley JH. Structural basis for viral late-domain binding to Alix. Nat Struct Mol Biol. 2007; 14(3): 194-199. https://doi.org/10.1038/nsmb1203 PMID: 17277784

428. Zhai Q, Fisher RD, Chung H-Y, Myszka DG, Sundquist WI, Hill CP. Structural and functional studies of ALIX interactions with YPX n L late domains of HIV-1 and EIAV. Nat Struct Mol Biol. 2008; 15(1): 43-49. https://doi.org/10.1038/nsmb1319 PMID: 18066081

429. Sámano-Sánchez H, Gibson TJ. Mimicry of short linear motifs by bacterial pathogens: a drugging opportunity. Trends in biochemical sciences. 2020; 45(6): 526-544. https://doi.org/10.1016/j.tibs.2020.03.003 PMID: 32413327

430. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873): 583-589. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

431. Harcourt J, Tamin A, Lu X, Kamili S, Sakthivel SK, Murray J, et al. Severe acute respiratory syndrome coronavirus 2 from patient with coronavirus disease, United States. Emerg Infect Dis. 2020; 26(6): 1266. https://doi.org/10.3201/eid2606.200516 PMID: 32160149

432. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology. 2020; 5(11): 1403-1407. https://doi.org/10.1038/s41564-020-0770-5 PMID: 32669681

433. Gadhave K, Kumar P, Kumar A, Bhardwaj T, Garg N, Giri R. Conformational dynamics of 13 amino acids long NSP11 of SARS-CoV-2 under membrane mimetics and different solvent conditions. Microb Pathog. 2021: 105041. https://doi.org/10.1016/j.micpath.2021.105041 PMID: 34119626

434. Chen B, Tian E-K, He B, Tian L, Han R, Wang S, et al. Overview of lethal human coronaviruses. Signal transduction and targeted therapy. 2020; 5(1): 1-16. https://doi.org/10.1038/s41392-020-0190-2 PMID: 32533062

435. Tzou PL, Tao K, Nouhin J, Rhee S-Y, Hu BD, Pai S, et al. Coronavirus antiviral research database (CoV-RDB): an online database designed to facilitate comparisons

between candidate anti-coronavirus Compounds. Viruses. 2020; 12(9): 1006. https://doi.org/10.3390/v12091006 PMID: 32916958

436. Ziebuhr J, Thiel V, Gorbalenya AE. The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond. J Biol Chem. 2001; 276(35): 33220-33232. https://doi.org/10.1074/jbc.M104097200 PMID: 11431476

437. Sanches PR, Charlie-Silva I, Braz HL, Bittar C, Calmon M, Rahal P, et al. Recent advances in SARS-CoV-2 Spike protein and RBD mutations comparison between new variants Alpha (B. 1.1. 7, United Kingdom), Beta (B. 1.351, South Africa), Gamma (P. 1, Brazil) and Delta (B. 1.617. 2, India). Journal of Virus Eradication. 2021: 100054. https://doi.org/10.1016/j.jve.2021.100054 PMID: 34548928

438. Van Der Hoek L, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, Wolthers KC, et al. Identification of a new human coronavirus. Nat Med. 2004; 10(4): 368-373. https://doi.org/10.1038/nm1024 PMID: 15034574

439. St-Jean JR, Jacomy H, Desforges M, Vabret A, Freymuth F, Talbot PJ. Human respiratory coronavirus OC43: genetic stability and neuroinvasion. J Virol. 2004; 78(16): 8824-8834. https://doi.org/10.1128/JVI.78.16.8824-8834.2004 PMID: 15280490

440. Woo PC, Lau SK, Chu C-m, Chan K-h, Tsoi H-w, Huang Y, et al. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J Virol. 2005; 79(2): 884-895. https://doi.org/10.1128/JVI.79.2.884-895.2005 PMID: 15613317

441. Zaki AM, Van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012; 367(19): 1814-1820. https://doi.org/10.1056/NEJMoa1211721 PMID: 23075143

442. Tao Y, Tong S. Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from Kenyan bats. Microbiology resource announcements. 2019; 8(28): e00548-00519. https://doi.org/10.1128/MRA.00548-19 PMID: 31296683

443. McCallum M, Walls AC, Sprouse KR, Bowen JE, Rosen LE, Dang HV, et al. Molecular basis of immune evasion by the delta and kappa SARS-CoV-2 variants.

Science. 2021: eabl8506. https://doi.org/10.1101/2021.08.11.455956 PMID: 34401880

444. Romero PE, Dávila-Barclay A, Salvatierra G, González L, Cuicapuza D, Solís L, et al. The emergence of SARS-CoV-2 variant lambda (C. 37) in South America. Microbiology spectrum. 2021; 9(2): e00789-00721. https://doi.org/10.1128/Spectrum.00789-21 PMID: 34704780

445. Teo G, Liu G, Zhang J, Nesvizhskii AI, Gingras A-C, Choi H. SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. J Proteomics. 2014; 100: 37-43. https://doi.org/10.1016/j.jprot.2013.10.023 PMID: 24513533

446. Jaeger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, et al. Global landscape of HIV–human protein complexes. Nature. 2012; 481(7381): 365-370. https://doi.org/10.1038/nature10719 PMID: 22190034

447. Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014; 114(13): 6589-6631. https://doi.org/10.1021/cr400525m PMID: 24773235

448. Mihel J, Šikić M, Tomić S, Jeren B, Vlahoviček K. PSAIA–protein structure and interaction analyzer. BMC Struct Biol. 2008; 8(1): 1-11. https://doi.org/10.1186/1472-6807-8-21 PMID: 18400099

449. Zhang J, Cruz-Cosme R, Zhuang M-W, Liu D, Liu Y, Teng S, et al. A systemic and molecular study of subcellular localization of SARS-CoV-2 proteins. Signal transduction and targeted therapy. 2020; 5(1): 1-3. https://doi.org/10.1038/s41392-020-00372-8 PMID: 33203855

450. Thul PJ, Lindskog C. The human protein atlas: a spatial map of the human proteome. Protein Sci. 2018; 27(1): 233-244. https://doi.org/10.1002/pro.3307 PMID: 28940711

451. Liu J, Li Y, Liu Q, Yao Q, Wang X, Zhang H, et al. SARS-CoV-2 cell tropism and multiorgan infection. Cell discovery. 2021; 7(1): 1-4. https://doi.org/10.1038/s41421-021-00249-2 PMID: 33758165

452. Yoshimoto FK. The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. The protein journal. 2020; 39: 198-216. https://doi.org/10.1007/s10930-020-09901-4 PMID: 32447571

453. Cheng A, Grant CE, Noble WS, Bailey TL. MoMo: discovery of statistically significant post-translational modification motifs. Bioinformatics. 2019; 35(16): 2774-2782. https://doi.org/10.1093/bioinformatics/bty1058 PMID: 30596994

454. Kumar M, Michael S, Alvarado-Valverde J, Mészáros B, Sámano-Sánchez H, Zeke A, et al. The Eukaryotic Linear Motif resource: 2022 release. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gkab975 PMID: 34718738

455. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annual review of biochemistry. 2014; 83: 553-584. https://doi.org/10.1146/annurev-biochem-072711-164947 PMID: 24606139

456. Uversky VN. Paradoxes and wonders of intrinsic disorder: Stability of instability. Intrinsically disordered proteins. 2017; 5(1): e1327757. https://doi.org/10.1080/21690707.2017.1327757 PMID: 30250771

457. Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform. 2009; 10(3): 205-216. https://doi.org/10.1093/bib/bbn057 PMID: 19151098

458. Erdős G, Pajkos M, Dosztányi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gkab408 PMID: 34048569

459. Liu Y, Wang X, Liu B. RFPR-IDP: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins. Brief Bioinform. 2021; 22(2): 2000-2011. https://doi.org/10.1093/bib/bbaa018 PMID: 32112084

460. Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates III JR, et al. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol Cell. 2009; 33(3): 365-376. https://doi.org/10.1016/j.molcel.2008.12.028 PMID: 19217410

461. Kanhoush R, Beenders B, Perrin C, Moreau J, Bellini M, Penrad-Mobayed M. Novel domains in the hnRNP G/RBMX protein with distinct roles in RNA binding and targeting nascent transcripts. Nucleus (Calcutta). 2010; 1(1): 109-122. https://doi.org/10.4161/nucl.1.1.10857 PMID: 21327109

462. Horiuchi K, Kawamura T, Iwanari H, Ohashi R, Naito M, Kodama T, et al. Identification of Wilms' tumor 1-associating protein complex and its role in alternative splicing and the cell cycle. J Biol Chem. 2013; 288(46): 33292-33302. https://doi.org/10.1074/jbc.M113.500397 PMID: 24100041

463. Sillibourne JE, Delaval B, Redick S, Sinha M, Doxsey SJ. Chromatin remodeling proteins interact with pericentrin to regulate centrosome integrity. Mol Biol Cell. 2007; 18(9): 3667-3680. https://doi.org/10.1091/mbc.e06-07-0604 PMID: 17626165

464. Okuwaki M, Saito S, Hirawake-Mogi H, Nagata K. The interaction between nucleophosmin/NPM1 and the large ribosomal subunit precursors contribute to maintaining the nucleolar structure. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research. 2021; 1868(1): 118879. https://doi.org/10.1016/j.bbamcr.2020.118879 PMID: 33039556

465. Nikolakaki E, Giannakouros T. SR/RS motifs as critical determinants of coronavirus life cycle. Front Mol Biosci. 2020; 7. https://doi.org/10.3389/fmolb.2020.00219 PMID: 32974389

466. Wang M, Li M, Ren R, Li L, Chen E-Q, Li W, et al. International expansion of a novel SARS-CoV-2 mutant. J Virol. 2020; 94(12): e00567-00520. https://doi.org/10.1128/JVI.00567-20 PMID: 32269121

467. Wu H, Xing N, Meng K, Fu B, Xue W, Dong P, et al. Nucleocapsid mutations R203K/G204R increase the infectivity, fitness and virulence of SARS-CoV-2. Cell Host Microbe. 2021. https://doi.org/10.1016/j.chom.2021.11.005 PMID: 34822776

468. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Science Advances. 2020; 6(27): eabb9153. https://doi.org/10.1101/2020.03.20.000885 PMID: 32511348

469. Bergsneider B, Bailey E, Ahmed Y, Gogineni N, Huntley D, Montano X. Analysis of SARS-CoV-2 infection associated cell entry proteins ACE2, CD147, PPIA, and PPIB in datasets from non SARS-CoV-2 infected neuroblastoma patients, as potential prognostic and infection biomarkers in neuroblastoma. Biochemistry and biophysics reports. 2021; 27: 101081. https://doi.org/10.1016/j.bbrep.2021.101081 PMID: 34307909

470. Kumar A, Ishida R, Strilets T, Cole J, Lopez-Orozco J, Fayad N, et al. SARS-CoV-2 non-structural protein 1 inhibits the interferon response by causing depletion of key host signaling factors. J Virol. 2021: JVI. 00266-00221. https://doi.org/10.1128/JVI.00266-21 PMID: 34110264

471. Kanduc D, Shoenfeld Y. Molecular mimicry between SARS-CoV-2 spike glycoprotein and mammalian proteomes: implications for the vaccine. Immunol Res. 2020; 68(5): 310-313. https://doi.org/10.1007/s12026-020-09152-6 PMID: 32946016

472. Oldstone MB. Molecular mimicry: its evolution from concept to mechanism as a cause of autoimmune diseases. Monoclonal antibodies in immunodiagnosis and immunotherapy. 2014; 33(3): 158-165. https://doi.org/10.1089/mab.2013.0090 PMID: 24694269

473. Cortés MJ, Wong-Staal F, Lama J. Cell surface CD4 interferes with the infectivity of HIV-1 particles released from T cells. J Biol Chem. 2002; 277(3): 1770-1779. https://doi.org/10.1074/jbc.M109807200 PMID: 11704677