UNDERSTANDING VIRUS-HOST INTERACTIONS USING ARTIFICIAL INTELLIGENCE

Dan Liu

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

College of Medical, Veterinary and Life Sciences
University of Glasgow

March 2025

Abstract

Viruses are associated with a wide range of hosts, encompassing all cellular life: humans, animals, plants and bacteria. Virus-host interactions are complex and diverse across species, mostly mediated by protein-protein interactions (PPIs). PPIs play important roles in essential biological activities, including forming protein complexes or more transient interactions in the context signalling pathways, regulatory networks etc., some of which are important for virus infection such as viral entry and replication. Understanding PPI mechanisms is helpful for revealing virus-host interactions, identifying PPIs associated with diseases and discovering potential therapeutic targets. However, the host specificity of most viruses remains unknown, and PPI networks remain sparse except for a few well-studied host species such as human. In this thesis, we developed computational approaches to predict host species of viruses and PPIs from genomes and corresponding protein sequences alone. Firstly, we introduce EvoMIL, a deep learning framework that leverages the protein language model (PLM) for viral protein representations and trains a multiple instance learning (MIL) model to predict prokaryotic and eukaryotic host species. We show that EvoMIL improves the accuracy of host species prediction and can identify key viral proteins that contribute to host specificity. Next, we introduce a deep-learning model, PLM-interact, jointly encoding protein pairs to learn protein interactions, analogous to the next-sentence prediction task in natural language processing (NLP). We show that PLM-interact improves PPI prediction in the intra-species benchmarking task and can identify mutational impacts of human PPIs. We show that PLM-interact can be implemented to predict virus-host PPIs. To enhance training datasets, we construct a dataset by integrating seven public virus-human PPI databases. We introduce three data-splitting strategies to create training, validation and test datasets where training and test sets have varying protein similarities, enabling comprehensive model evaluation. We discover that fine-tuning the human model on virus-human PPIs improves virus-human PPI prediction, offering the potential for developing a generalizable PPI model. In summary, this thesis aims to use deep learning techniques to predict the host specificity for viruses and identify PPIs within and between species using protein sequences. This broadens our view of virus-host interactions and provides insights into developing vaccines, drugs and therapies for human diseases.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank my PhD supervisors, David Robertson and Ke Yuan, for their guidance, insightful feedback and patience throughout my PhD journey. I feel incredibly fortunate to have the opportunity to work with you. Thank you both for your support, and I am deeply grateful for everything I have learned and gained under your mentorship.

I am extremely grateful to David for giving me the opportunity to work on this exciting PhD project funded by the Marie Sklodowska-Curie Actions Innovative Training Networks VIROINF. I extend my thanks to Craig Macdonald from the School of Computer Science for his insightful collaboration on language models, as well as to present and past members of the Robertson Lab: Francesca Young, Kieran Lamb, Spyros Lytras, Haiting Chai, Sejal Modha, Ewan Smith and Rubayet Alam.

I would like to thank everyone involved in the weekly regular virus-host interaction meeting with my supervisors, Kieran and Fran. I am grateful to Fran and Kieran for discussing my projects and providing valuable suggestions. Thanks to Fran for helping with my initial project about host prediction. It was nice to chat with Kieran at the Learning Hub and ARC, and I truly enjoyed discussing my projects with him. I greatly appreciate Adalberto Claudio Quiros's support in language model training for my PPI prediction project. Thanks to my reviewers, Joseph Hughes and Jake Lever, for their feedback during my three years of progress review meetings. I also thank the DiRAC Extreme Scaling service (Tursa) for providing computational resources that enabled my PPI prediction project. I greatly appreciate the technical meetings and journal clubs; I had valuable discussions with Kieran Lamb, Francesca Young, Alexandrina Pancheva, Adalberto Claudio Quiros, Crispin Miller, Craig Macdonald, David Robertson and Ke Yuan.

I am grateful to the VIROINF coordinators, Manja Marz and Winfried Göttsch, for organizing retreats and workshops in Germany and Belgium. I particularly cherish my time in Frankfurt, Leuven, and Hiddensee for research and presentation skills workshops. It was lovely to participate in these events alongside other PhD students involved in the VIROINF programme. These experiences were incredibly helpful to my PhD journey.

During my PhD, I have two replacements in Germany. I spent three months at Helmholtz Zentrum München and am grateful to Li Deng for hosting me. Thanks to Xue Peng for her support and insightful discussions about our projects during my replacement. My second replacement was with Bas E. Dutilh's group at Friedrich Schiller University Jena. I am grateful to Mikhail Fofanov for warmly welcoming me at the train station and providing continuous support throughout my three-month replacement. I would also like to thank Bas for discussing my host prediction project and participating in online meetings with David and Fran. Special thanks to Swapnil Doijad for preparing datasets for my model testing. Additionally, our lab hosts two visiting PhD students, I am grateful to Xue Peng for sharing many ideas and engaging in discussions about our host prediction projects. I also appreciate Jun Wang for his insightful conversations about language models.

Finally, I would like to express my deepest gratitude to my family. Thanks to my brother, who always supported me and listened to me. Thanks to my parents, who have encouraged and motivated me throughout this journey. I am also deeply grateful to my granny for teaching me to be courageous and ambitious. To my grandpa, I miss you dearly and hope I have made you proud.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Dan Liu

Jan 2025

Chapter 1

Introduction

In God we trust. All others must bring data.

W. Edwards Deming (1978)

Viruses enter host cells by binding with protein receptors and taking over the host cell's machinery for replication (Rampersad et al. 2018). This process involves hundreds of protein-protein interactions (PPIs) influencing viruses' fitness, host range and pathogenicity. PPIs mediate essential biological processes for viruses, including viral entry, replication and assembly; changes in binding affinity of protein interactions might lead to severe disease (Jemimah et al. 2020). Viruses are typically associated with a range of hosts (Bandín et al. 2011), and emerging viruses have influenced human health and global economics (Daszak et al. 2008). The COVID-19 pandemic caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) led to millions of deaths worldwide (Statistics 2025) and a global recession (Atzrodt et al. 2020). However, knowledge about hosts of novel viruses and virus-host PPIs remains limited. A comprehensive understanding of the complex virus-host associations is useful for predicting host ranges and host-switching events. Accurately identifying hosts of viruses provides strategies for early intervention and preventing zoonotic disease transmission. Understanding PPI mechanisms and identifying virus-human PPIs are essential for the discovery and design of antiviral drugs and vaccines and provide potential treatment strategies.

With the development of next-generation technologies (NGS), a growing number of new viruses have been discovered (Datta et al. 2015). However, the hosts of these viruses are typically unknown. Virus diversity and sparse virus-host association networks make identifying host specificity for novel viruses challenging. The PPI networks remain sparse as only a few species such as human and mouse are well-studied with comprehensive coverage (Kotlyar et al. 2015; Shin

et al. 2020). Conventional experimental approaches for identifying virus-host associations and their PPIs are time-consuming and costly, making it difficult to depend solely on experimental approaches for a large scale of unknown data. Additionally, experimental methods do not always provide accurate identification. For example, plaque assays are a sensitive approach for determining which phage infects which bacteria and infections are determined by host cell clearance, but this method is not applicable to all viruses (Iuchi et al. 2023). Viral tagging (VT) can quickly identify which host phage can bind and potentially infect but cannot confirm viral entry and the entire infectious cycle (Deng et al. 2014).

Artificial intelligence (AI) has developed rapidly over the past two decades and has been implemented in various fields, including computer science, biology and chemistry. Recently, deep learning and large language models (LLMs), originally used for natural language processing (NLP), have been successfully applied to virology, such as host prediction of viruses (Liu et al. 2024a), viral protein structure prediction (Yang et al. 2022a; Kim et al. 2025) and virus-host PPI prediction (Liu et al. 2024b; Sledzieski et al. 2023). Additionally, the number of protein and genomic sequences in public datasets has increased significantly. NCBI contains genomes and genes from a wide range of species, including viruses, bacteria, archaea, animals and plants. UniProt has collected roughly 120 million proteins from different species (The UniProt Consortium 2009). These extensive sequence resources provide training resources for developing machine-learning approaches to predict virus-host associations and PPIs

This thesis begins with an overview of virology, including virus diversity, viral life cycle, viral evolution and PPI concepts. Next, it outlines deep learning and LLM techniques. We will show how to train deep learning models to understand virus-host interactions and reveal the complexity of molecular interactions from protein sequences. This thesis will introduce a deep learning model using viral protein features represented by a pre-trained protein language model (PLM) to predict host specificity and retrain a PLM to predict PPIs within and between species by jointly encoding protein pairs.

1.1 Background to viruses and protein-protein interactions

An estimated 10^{31} virus particles exist on the earth (Mushegian 2020), they are highly diverse and can be classified into various categories. Viruses consist of a protein coat (capsid) encapsulating their genetic material and can infect all cellular life: animals, plants and microbiomes. Virus-host interactions involve hundreds of PPIs, and understanding PPI mechanisms is helpful for comprehensive analysis of the key biological activities during viral infections. Identifying virus-host PPIs offers the potential to discover drugs and design proteins targeting viral diseases (Idrees et al. 2024) or use phage therapy for infectious bacteria (Fabijan et al. 2019). However,

viral escape and host immune response create dynamics in virus-host interactions and coevolution, making it challenging to identify virus-host PPIs.

In 2019, the COVID-19 pandemic changed people's life worldwide. Viral variants such as Alpha and Omicron made the virus more infectious, spread more broadly and caused the lockdown in some areas (Aiano et al. 2022; Yang et al. 2022b). Identifying which human proteins interact with emerging viruses can accelerate vaccine development. Furthermore, identifying hosts for a novel virus helps determine its host range, which provides insights into potential host-switching and strategies for tracking and preventing virus spread across different species. Viral infections might lead to human and plant diseases. For example, cervical cancer is caused by the long-term infection of human papillomavirus (HPV) (Hausen 1996). Moreover, mutations that disrupt PPIs in humans are associated with diseases (Gao et al. 2015); identifying PPIs helps develop efficient antiviral drugs and offers potential therapeutic strategies.

1.1.1 Virus Baltimore classification

The Baltimore classification system (Baltimore 1971), introduced by David Baltimore in 1971, is used to classify viruses based on the different ways of mRNA production and replication. Viruses are highly diverse and have different classifications, including DNA viruses, RNA viruses and reverse-transcribing viruses. The Baltimore classification system was created based on the essential role of viral mRNAs in the processes of protein synthesis. Figure 1.1 shows that mRNA is in the centre and the pathways from DNA or RNA genomes to mRNA. RNA viruses generally have high error and mutation rates due to the lack of proofreading mechanisms, apart from rare instances like SARS-CoV-2, which increases their diversity and allows them to evolve and adapt to diverse environments (Szpara et al. 2021; Robson et al. 2020). In contrast, DNA viruses are more stable and can correct replication errors through DNA polymerases (Szpara et al. 2021).

The Baltimore classification includes seven classes of viral genomes, as shown in Figure 1.1. **Group I**: Double-stranded DNA (dsDNA) viruses, **Group II**: Single-stranded DNA (ssDNA) viruses, **Group III**: Double-stranded RNA (dsRNA) viruses, **Group IV**: Positive-sense single-stranded RNA (+ssRNA) viruses, **Group V**: Negative-sense single-stranded RNA (-ssRNA) viruses, **Group VI**: single-stranded RNA viruses with a DNA intermediate (ssRNA-RT), and **Group VII**: double-stranded DNA viruses with an RNA intermediate (dsDNA-RT). Most viruses associated with prokaryotes are dsDNA (double-stranded DNA) viruses belonging to **Group I**, such as Escherichia virus T4 (T4 phage). Viruses associated with eukaryotes are highly diverse RNA viruses, such as SARS-CoV-2 (**Group IV**) and HIV (Human Immunodeficiency Virus) (**Group VI**). Overall, the Baltimore classification system clarifies the genetic information flow

in viruses and different mechanisms for translating RNA to proteins. Understanding the various types of viruses associated with prokaryotes and eukaryotes is helpful for analysing virus evolution and virus-host interactions.

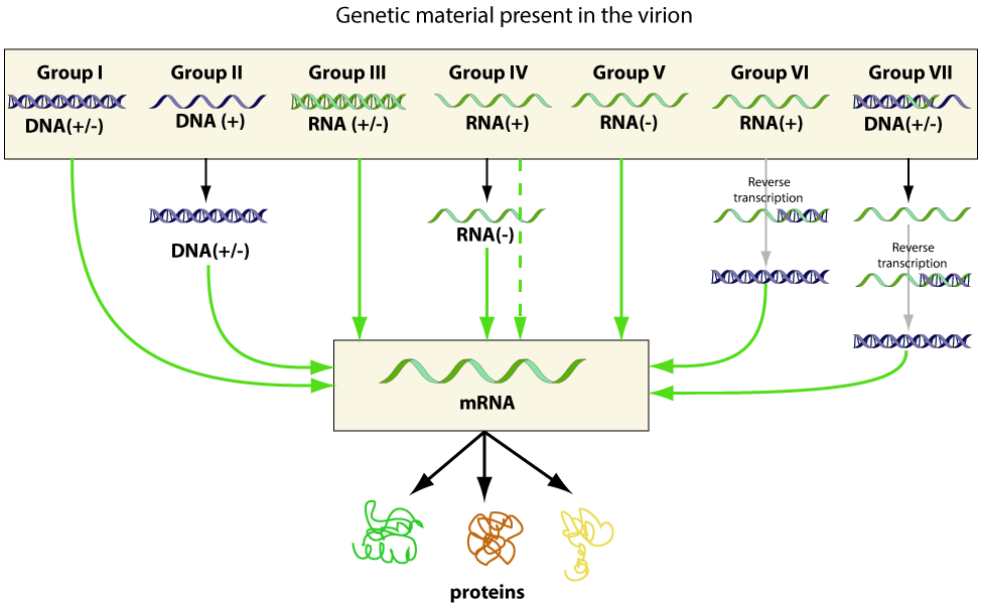

Figure 1.1. The Baltimore virus classification (Baltimore 1971). The Baltimore classification system classifies viruses into seven classes based on mRNA synthesis and the polarity of genomes. This figure is from Viralzone (https://viralzone.expasy.org/).

1.1.2 Virus Taxonomy

Virus taxonomy is essential to understanding the evolutionary history of viruses and analysing complicated evolutionary relationships among viruses (Simmonds et al. 2023). The naming and taxonomic classification guidelines of viruses are carried out by the International Committee on Taxonomy of Viruses (ICTV). ICTV published the four principles to create the universal virus taxonomy and reported 15 ranks from realm to species (Figure 1.2). The species is at the lowest taxonomic level, which includes closely related viruses that share genetic information. In contrast, the realm is at the highest taxonomic level, which consists of a wide range of virus species with diverse and complex evolutionary relationships.

In Figure 1.2, at lower taxonomies (species, genus and family), nucleotide and amino acid sequence similarities are used to identify evolutionary relationships. For higher taxonomy (realm, kingdom, and class), due to the sequence divergence of the higher taxonomy viruses, conserved virus features are required for virus phylogeny: (1) core genes, which are highly conserved and necessary for essential activities such as virus replication (Payne 2017); (2) viral motifs, which evolved slowly and used for host mimicry due to critical functions in viral life cycle (Shuler et al. 2022). The virus taxonomy tree is helpful to understand virus evolutionary history and iden-

tify potential hosts, e.g., Babayan et al. (2018) applied viral phylogenetic relatedness to predict reservoir hosts, as closely related viruses tend to share similar hosts (Kitchen et al. 2011).

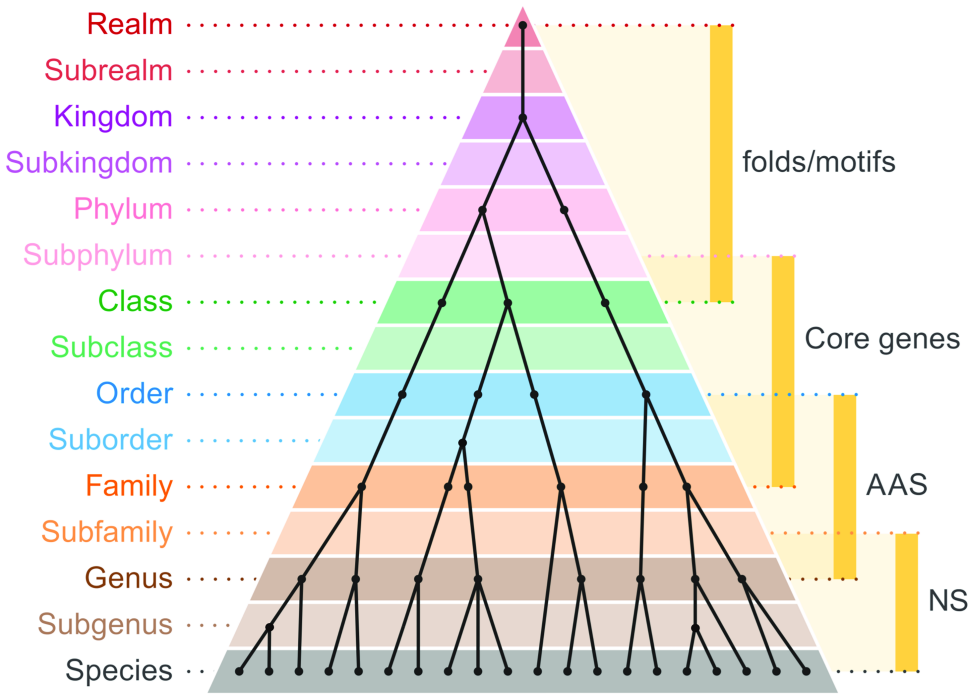

Figure 1.2. Ranks of virus taxonomy. The 15 ranks used by the ICTV are shown on the left, and the pyramid shape shows the number of taxa increasing from realm to species. The methodologies are on the right (NS, nucleotide sequence similarity; AAS, amino acid sequence similarity; Core genes; folds/motifs).This figure is taken from Simmonds et al. (2023).

1.1.3 Viral life cycle

The viral life cycle comprises six key stages: attachment, viral entry, uncoating, replication, assembly and release (Gil-Farina et al. 2016) (see Figure 1.3). These stages involve hundreds of PPIs for essential biological activities, including viral entry and replication. Viruses attach to the host cell's surface and enter the host cell; viruses replicate using the host replication mechanism and then assemble and release their progeny to infect other cells. Although the viral life cycle has the same basic stages for viral replication, some processes might differ across viruses. For example, virulent phages hijack host cellular machinery to generate a certain number of progeny phages, then lyse hosts and release their progenies to infect new cells. In contrast, lysogenic phages insert genetic materials into the host cells' DNA sequence, allowing the virus to replicate alongside the host genome replication (Doss et al. 2017).

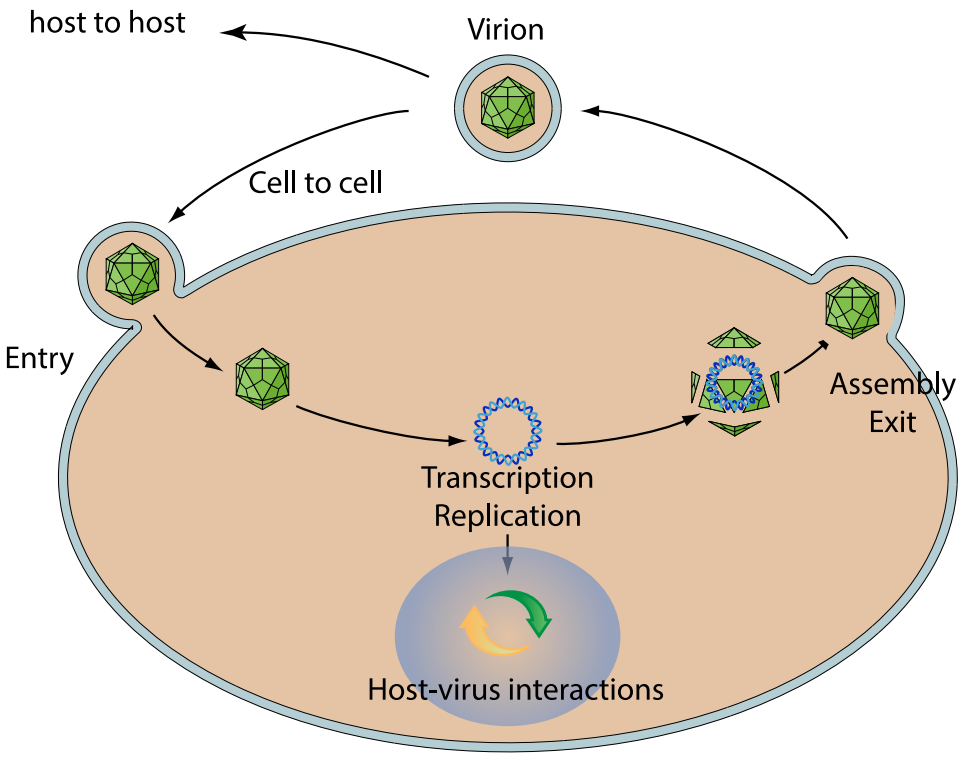

Figure 1.3. Viral life cycle. There are six main steps: binding, fusion, transcription, replication, assembly, and release. This figure is downloaded from Viralzone (https://viralzone.expasy.org/).

1.1.3.1 Virus entry

Viruses can infect different hosts including bacteria, archaea and eukaryotes and attach to the host surface by binding to specific host receptors. Viruses use a diverse set of viral proteins for viral entry (Dimitrov 2004). For example, bacteriophage T4 binds to the *E. coli* receptor surface consisting of the OmpC protein, lipopolysaccharide and the peptidoglycan layer (Yu et al. 1982), then injects dsDNA into *E. coli* cell by its tail fibres. The receptor-binding domain (RBD) of SARS-CoV-2 spike (S) glycoprotein is attached to the angiotensin-converting enzyme 2 (ACE2) receptor on the surface of human cells (Lan et al. 2020), enabling the fusion of human cells. Hundreds of viruses can infect humans and cause various diseases (Dimitrov 2004). Viral entry inhibitors prevent viruses from binding to host receptors, which have been used to develop new antiviral agents for HIV treatments (Moore et al. 2003; Melby et al. 2009). Understanding viral entry proteins and host-cell receptor interactions offers potential strategies for the development of novel antiviral therapies and vaccines (Tilton et al. 2010).

1.1.3.2 Virus replication

Viruses do not have cellular machinery for independent replication; rather, they hijack the host cell machinery to produce messenger mRNA (mRNA) for protein synthesis. After viral entry,

enzymes of viruses or hosts degrade the virus capsid to release viral genomic nucleic acids such as DNA and RNA. All cellular organisms have double-stranded DNA genomes, and the flow of genetic information follows the Central Dogma' of molecular biology: DNA is transcribed to mRNA and then translated to protein (Crick 1958). In contrast, viral genomes can be DNA or RNA with either single-stranded or double-stranded forms. Viral replication is diverse and depends on the viruses' genome types according to the Baltimore classification system mentioned in Section 1.1.1. For example, dsDNA and dsRNA viruses can be directly transcribed to mRNA, whereas ssDNA viruses first convert to dsDNA and then transcribe to mRNA.

1.1.4 Viral evolution

Viruses typically have high mutation rates, large population sizes and genetic exchange across diverse gene pools (Villarreal 2008), especially RNA viruses. While virus and host evolution follow the same Darwinian principles involving variation and natural selection, viruses evolve on shorter timescales since their generations are much shorter. High error rates of viruses indicate rapid evolution and genetic diversity, which benefit viruses in adapting to the host environment and escaping the host's immune response (Stern et al. 2016). This makes it challenging to control infection and design antiviral treatments and vaccines.

Viral evolution can be influenced by hosts and environmental factors through genetic drift and natural selection (LaTourrette et al. 2022). Genetic drift is the random change in allele frequency across generations. It can result in some alleles going extinct or becoming common (Masel 2011). Natural selection is the tendency of traits to increase or decrease in a population, which is the main driver of organisms for environmental adaptation and genetic diversity (Karlsson et al. 2014). Mutations that are associated with beneficial traits of increasing survival probabilities and reproducing are referred to as being selected by positive or adaptive selection (Gayà-Vidal et al. 2014). Mutations that remove detrimental alleles from the population are referred to as being under negative or purifying selection (Karlsson et al. 2014). Many mutations are neutral, having no selective advantage or disadvantage and vary in frequency. Viruses can also evolve through recombination and reassortment. Recombination is the exchange of genetic materials between different but related viruses that infect the same host cell, resulting in the generation of new variants. Reassortment is the swap of genome segments from segmented viruses during co-infection and can make new viral strains. Virus adaptive evolution refers to mutations that increase fitness, optimize replication and adapt to their host environments.

In influenza research, evolutionary mechanisms are referred to as antigenic drift and antigenic shift. Influenza viruses undergo antigenic drift and antigenic shift to evade host immune recognition and infect host cells (Shao et al. 2017). Antigenic drift happens occasionally and consists

of a small-scale change caused by the gradual accumulation of mutations. Antigenic shift is the larger-scale change due to the reassortment of gene segments between different viral strains and can shape a new variant. Additionally, Horizontal gene transfer (HGT) is a process of gene exchange between viruses and their hosts. Viral HGT leads to the acquisition of new genes from their hosts, which drives the evolution of viruses and their hosts (Szpara et al. 2021; Irwin et al. 2022).

1.1.4.1 Arms race

Viruses evolve to evade host defence responses, driving a rapid virus-host evolutionary arms race (Franzosa et al. 2011). The receptor of host cells evolves to escape virus infection, and virus binding proteins also evolve to make the virus more infectious and adapt to host evolution. For example, most mutations in SARS-CoV-2 occur within the Spike receptor-binding domain (RBD), which plays an essential role in binding with the ACE2 receptor of the human cell surface. The RBD is also the primary target for neutralizing antibodies (Graham et al. 2021). These mutations can increase the probability of antibody escape (Magazine et al. 2022) and alter the binding affinity to the ACE2 receptor, increasing the transmissibility of viruses (Wrobel 2023). Due to selective pressures from viruses' evolution, hosts evolved by different survival mechanisms, including resistance and tolerance that limit the influences of infection (Schneider et al. 2008), and defence system that limits viral replication and transmission (Duggal et al. 2012).

1.1.4.2 Viral mimicry

Viruses often imitate their hosts' nucleotide sequences and protein structures to facilitate entry, exploit host mimics to evade host immune defences and promote infection. This strategy allows viruses to successfully complete replication and assembly (Davey et al. 2011). Viruses primarily mimic their hosts through two mechanisms: nucleotide sequence mimicry and protein structure mimicry.

Nucleotide sequence mimicry includes dinucleotide and codon usage bias (CUB). Dinucleotides consist of two nucleic acids, such as CpG and UpA. The Zinc-finger Antiviral Protein (ZAP) in hosts binds CpG dinucleotides of viral RNA and can restrict viral replication (Ficarelli et al. 2021; Andrade et al. 2023), which protects cells from RNA viruses' infection (Meagher et al. 2019). RNA viruses of most vertebrates and plants tend to reduce the frequency of the dinucleotide CpG in order to mimic their hosts' mRNA for escaping host immunity (Greenbaum et al. 2008). For example, CpG dinucleotides are suppressed in HIV-1 to escape ZAP inhibition

(Kmiec et al. 2020). CUB is the difference of synonymous codon frequencies in a given gene (Mordstein et al. 2021), which can be driven by mutational and selection pressures, dinucleotide frequencies and GC content (Tyagi et al. 2022). CUB help viruses adapt to hosts' codon usage patterns and evade antiviral defences (Bahir et al. 2009; Butt et al. 2016).

Host protein structure mimicry is a mechanism employed by viruses to interact with hosts and evade the host's immune system (Maguire et al. 2024). This mimicry occurs through various mechanisms, including the imitation of short linear motifs (SLiMs) and the acquisition of host proteins. Viruses attach and enter host cells by mimicking short linear motifs (SLiMs) of hosts (Via et al. 2015; Hagai et al. 2014). SLiMs are short-peptide sequences and typically three to ten consecutive amino acids; they play an essential role in biological activities and mediate PPIs (Davey et al. 2012). Soorajkumar et al. (2022) reported four dominant SLiMs in the Omicron variant of SARS-CoV-2 and demonstrated that they are most likely to interact with immune functions. SLiM mimicry helps viral entry and survival, and identifying host proteins imitated by viruses provides the potential to develop effective vaccines and therapeutic targets for viruses (Mihali et al. 2023; Simonetti et al. 2023). In addition to SLiMs, viruses can acquire host proteins to assist viral function. The acquisition of host proteins helps viruses with assembly and budding (Tremblay et al. 1998; Lin et al. 2022). Identifying host proteins acquired by viruses can enhance our understanding of virus mimicry.

Identifying virus mimicry is essential to understand virus-host interactions. Using nucleotide sequence mimicry or molecular mimicry, viruses can avoid detection by immune cells which provides insight into the development of antibodies and viral vaccines. Recent computational approaches have been developed to identify viral structural mimicry. Honig et al. (2021) demonstrated that more than 70% of viral mimics cannot be identified by protein sequence homology alone. Using protein structure similarities between viruses and hosts, over 6,000,000 instances of structural mimicry are identified.

1.1.5 Protein-protein interactions

Proteins have various functions in living entities and mainly perform functions by interacting with one or more proteins. Protein-protein interactions (PPIs) and multi-protein complexes play essential roles in different biological processes within hosts and between virus and host, such as immune response, cellular signal transduction and regulating cell signalling pathways. Protein complexes regulate DNA replication, transcription and translation, which are usually controlled by PPIs (Lu et al. 2020). Because of their key role in biological function, abnormal PPIs are associated with cancer (White et al. 2008) and neurodegenerative diseases (Blazer et al. 2009).

Amino acid substitutions include neutral and disease-associated mutations. These mutations may change the physicochemical properties of amino acids and cause or disrupt PPIs (Kucukkal et al. 2015), which can significantly impact biological processes and lead to cancers and congenital heart diseases (Xiong et al. 2022). Therefore, understanding PPI mechanisms is an important strategy for new drug development and helpful for disease treatment (Lu et al. 2020). For example, the MDM2 protein has a strong affinity for the p53 tumour suppressor protein. Blocking the MDM2-p53 interaction can stabilise p53, providing a potential strategy for cancer treatment (Vassilev et al. 2004). Overall, constructing protein interaction networks helps PPI inference and understanding of human diseases associated with PPIs, offering crucial insights into developing disease therapies.

Experimental and computational approaches have been applied to identify PPIs. Typical experimental approaches include yeast two-hybrid system(Y2H) (Suter et al. 2008), co-immunoprecipitation (Co-IP) (Iqbal et al. 2018) and tandem affinity purification coupled to mass spectrometry (TAP-MS) (Ho et al. 2002). However, these methods are time-consuming and require laboratory equipment. Public PPI databases, such as STRING (Szklarczyk et al. 2019), IntAct (Kerrien et al. 2012) and BioGrid (Oughtred et al. 2019) have collected PPIs from different species, offering extensive training resources for the development of machine-learning approaches. Recent PPI models such as D-SCRIPT (Sledzieski et al. 2021) and TT3D (Sledzieski et al. 2023) are trained with human PPIs and then tested on PPIs of other species, providing an alternative approach to predict PPIs in less-studied organisms with limited training data. AlphaFold3 (Abramson et al. 2024) was developed for bimolecular interactions, including protein-ligand, protein-nucleic acid and antibody-antigen interactions. These computational methods greatly expedite the discovery of unknown PPIs and have significant potential for advancing drug design and discovery.

1.1.5.1 Host protein-protein interactions

Host protein-protein interactions (PPIs) are essential in cellular function and complex biological processes. Mutations in PPIs might disrupt PPIs and damage canonical protein functions, leading to severe human diseases. For example, mutations that disrupt or change protein interactions can lead to pan-cancer and Parkinson disease (Li et al. 2023; Muda et al. 2014). The human PPI networks provide insights into analysing molecule pathways of disease phenotypes and predicting disease-related genes (Safari-Alighiarloo et al. 2014). Furthermore, the construction of PPI networks can reveal fundamental activities in humans, which contributes to protein function discovery and the development of ideal drug targets and potential treatment strategies.

In current public PPI networks such as IntAct (Kerrien et al. 2012) and STRING (Szklarczyk et

al. 2019), human is well-studied with the largest number of PPIs. In IntAct, there are 1,630,482 binary interactions with 1,066,927 human PPIs. Luck et al. (2020) presented an interactome map with nearly 53,000 binary protein interactions in humans. PPIs from mouse, fly, worm, yeast and *E. coli* are also available in these public PPI databases but contain much fewer PPIs than human. As mentioned in Section 1.1.2, motifs and core genes are conserved between species. It indicates that there are similar protein patterns across species, especially in closely related species, such as human and mouse. Therefore, the large scale of human PPIs can be used as training resources for PPI inference in other species (Sledzieski et al. 2023; Liu et al. 2024b).

1.1.5.2 Virus-host protein-protein interactions

Virus-host protein-protein interactions (VH-PPIs) are essential in viral infection because viruses depend on hosts to survive. Viruses bind with host receptor proteins, inhibit the host response and escape the host's immune system. Various viruses, such as HIV, Influenza virus and SARS-CoV-2, can cause different human diseases: acquired immunodeficiency syndrome (AIDS), flu and COVID-19. For some viruses, infection can lead to cancer or tumours. For example, long-term infection of HPV can cause cervical cancer (Hausen 1996). Identification of VH-PPIs is important to understand viral pathogenesis, interaction mechanisms and host immune response for developing antiviral therapies and phage therapy strategies for antibiotic-resistant bacteria.

To replicate and spread to new hosts, viruses must evade or suppress interferons (IFNs)-mediated antiviral defences (García-Sastre 2017). IFNs are proteins produced by host cells in response to virus infection, acting as innate antiviral cytokines. They protect uninfected cells from viral invasion by inducing the transcription of hundreds of IFN-stimulated genes (ISGs) to trigger innate immune responses and inhibit viral replication and pathogenesis within host cells (Isaacs et al. 1957; Schneider et al. 2014). Viruses have evolved different strategies to counteract the antiviral effects of IFNs by disrupting their production or activity (García-Sastre 2017; Versteeg et al. 2010). Viruses evade these antiviral responses by encoding mechanisms that modulate IFN signalling, inhibit ISG activities and disrupt IFN-related cellular processes (Katze et al. 2002). The intricate interactions between host IFNs inhibiting viral replication and viral factors suppressing IFN activities form a complex dynamic network (García-Sastre 2017). Studies of IFNs are important for developing clinical treatments for virus-infectious diseases. For example, previous research demonstrated that treatment with IFNs can suppress HIV replication (Doyle et al. 2015) and IFN-λ is an antiviral therapeutic for influenza (Davidson et al. 2016). Overall, viruses must suppress IFN response to spread to new hosts. A comprehensive understanding of the modulation of the IFN system and viral evasion mechanisms is useful for developing effective methods to identify VH-PPIs.

In addition to IFNs, interactions between antigen-presenting cells (APCs) and T cells play an important role in the antiviral immune response. APCs can capture viral antigens and present them to T cells, initiating an adaptive immune response that helps identify and eliminate infected cells (Gaudino et al. 2019). When a virus infects a host, some T cells differentiate into memory T cells, enabling the body to quickly trigger innate immune defence mechanisms upon re-exposure to the same virus (Wang et al. 2021). This mechanism has been utilized in vaccine development, where inactivated vaccines are used to stimulate the production of memory T cells, providing immune protection. For example, current COVID-19 vaccines produce strong T cell responses that provide protection (Moss 2022). Overall, T cells play an essential role in the control of viruses and offer the potential for vaccine development.

1.2 Artificial Intelligence

Approaches based on Artificial intelligence (AI) have accelerated the discovery of novel viruses and protein-protein interactions (PPIs). Oncogenic mutations may disrupt or enable PPIs (Fu et al. 2025), and understanding PPI mechanisms can help develop potential cancer treatments (White et al. 2008). The development of sequencing technologies has advanced the construction of public databases, such as the protein database UniProt (The UniProt Consortium 2009) and the biological information database NCBI. These databases provide extensive training resources for Machine Learning (ML) models, enhancing the development of computational approaches for solving biological tasks. In this section, we will discuss some of the basic ML models, Deep Learning (DL) models, large language models (LLMs) and protein language models (PLMs). We will show how to encode biological sequences with PLMs and train DL-based models for specific tasks in virology.

1.2.1 Features representations of biological sequences

Protein and DNA sequences encode essential biological information in organisms. To enable the development of computational approaches, these sequences are converted into meaningful numerical vectors that encode biological information. ML-based approaches can leverage these features to train models on specific tasks. As mentioned in Section 1.1.5.2, understanding virus-host associations and their PPIs is important for developing drugs or vaccines. Here, we will discuss different methods of extracting feature representations from biological sequences for the prediction of virus-host associations and PPIs within and between species.

Conventional feature extraction methods of proteins and genomes for host prediction are based

on sequence alignment and composition. Alignment-based approaches search for homology between viruses and hosts: (1) exact matches between viruses and host genes (Fouts 2006); (2) viruses matched with host-encoded CRISPR spacers (Staals et al. 2013; Horvath et al. 2010). Alignment-free methods use sequence composition to measure virus-host sequence similarities, such as k-mers, which represent the frequencies of substrings of length k from DNA, amino acid sequences and physicochemical properties (Young et al. 2020). Computational approaches of predicting PPIs are mainly based on sequence homology and structural information. Shen et al. (2007) deployed protein sequence k-mers to train a ML-based model to predict PPIs. Structure-based features, including domain-domain interactions (DDIs) and domainmotif interactions (DMIs), are essential for understanding the structural and functional dynamics of PPIs and can be used for protein feature representation in PPI prediction models (Singhal et al. 2007; Pang et al. 2010). Overall, conventional computational approaches primarily encoded biological sequences using sequence alignment and composition.

With the growing number of protein and genome sequences in the public databases, PLMs (Rives et al. 2021; Elnaggar et al. 2021), DNA and RNA language models (Zhou et al. 2023; Chen et al. 2022) are developed to extract meaningful information from the large scale of biological sequences. These foundation models are implemented for downstream tasks, including protein and RNA structure prediction and generative design. Unlike the previous k-mer features with sequence composition information only, PLMs encode protein secondary structures and evolutionary information, providing biological features for protein sequences. Pre-trained PLMs have been successfully implemented to represent protein features, advancing state-of-the-art approaches for virus-host association and PPI prediction. EvoMIL (Liu et al. 2024a) applied a PLM ESM-1b (Rives et al. 2021) to extract viral protein features for host prediction. TT3D (Sledzieski et al. 2023) extracted protein features by concatenating embeddings from a pre-trained PLM and from a one-hot encoding of 3D interaction (3Di) structural sequence generated by Foldseek (Van Kempen et al. 2024) to predict PPIs.

1.2.2 Machine Learning Paradigms

Machine learning has been applied to different fields, including natural language processing (NLP) (Chowdhary 2020) and biology (Zitnik et al. 2019). The primary training methods in machine learning are supervised learning, unsupervised learning, semi-supervised learning and self-supervised learning. Supervised learning requires labelled samples to train, and each input has an output, such as regression and classification tasks. Unsupervised learning is typically used on unlabelled datasets, where it is trained to learn patterns or structures within the input data using algorithms, such as hierarchical clustering (Nielsen 2016) and K-means (Likas et al. 2003). Semi-supervised learning trains a model using a dataset composed of both labelled

and unlabelled samples, combining elements of supervised and unsupervised learning. Self-supervised learning is trained on unlabelled datasets and generates their labels from the data, such as LLMs and generative models. These training methods can be used to train different datasets based on specific task requirements.

1.2.3 Classification

Classification in machine learning aims to divide the data into different classes; the output is a discrete value, like true or false, based on the data features. It involves training a model on labelled datasets, with each training sample having a specific class, and then predicting labels for unseen data. Binary, multi-class, and multi-label classification are the primary classification categories in ML. In binary classification tasks, models are often trained on datasets with a single label and two possible outputs: 0 or 1. Multi-class classification tasks include more than two classes and have an output for each class. Multi-label classification can assign multiple labels to each sample. Classification evaluation metrics include accuracy, precision, recall, the area under the receiver operating characteristic curve (AUC), the area under the precision-recall curve (AUPR) and F1 score. Additionally, the receiver operating characteristic (ROC) curve (Bradley 1997) and precision-recall (PR) curve (Davis et al. 2006) are often used to show models' performance.

Binary classification is especially commonly applied to process tasks in virus-host interactions. Positive and negative samples are required to train binary classification models, allowing the model to learn features from each class and make predictions. Logistic regression and Support Vector Machines (SVM) are common binary classification methods. SVM has been used to train binary classification tasks to predict virus-host associations (Young et al. 2020), where the model outputs a probability score for classifying whether a query virus is associated with a host. DL algorithms like Convolutional Neural Networks (CNNs) (LeCun et al. 1998) and Recurrent Neural Networks (RNNs) (Rumelhart et al. 1986) are effective for handling more complex patterns on large dataset classification tasks. For example, D-SCRIPT (Sledzieski et al. 2021) was trained on roughly 421K human data using CNNs to predict PPIs.

1.2.4 Multiple instance learning

Multiple instance learning (MIL) is a variation of supervised learning that assigns labels to a collection of instances without individually classifying each instance (Maron et al. 1997). It was originally applied for drug activity prediction (Dietterich et al. 1997) and has been widely applied to image classification tasks (Quellec et al. 2017; Sudharshan et al. 2019). The training

samples in MIL are bags; each bag has one label and includes a group of instances. MIL is trained to predict a label for a bag of instances. In Chapter 2, we apply MIL to train binary and multi-class classification models based on viral protein representations for host prediction. Each virus contains a bag of protein sequences with one host label. In this case, only the label for one virus bag is available; the individual label for each protein sequence is unknown. MIL provides a solution for this host prediction task, as it can learn instance patterns from the bag labels. The bag view of MIL is shown in Figure 1.4. The bag is labelled as negative if all instances within it are negative, or positive if it contains one or more positive instances.

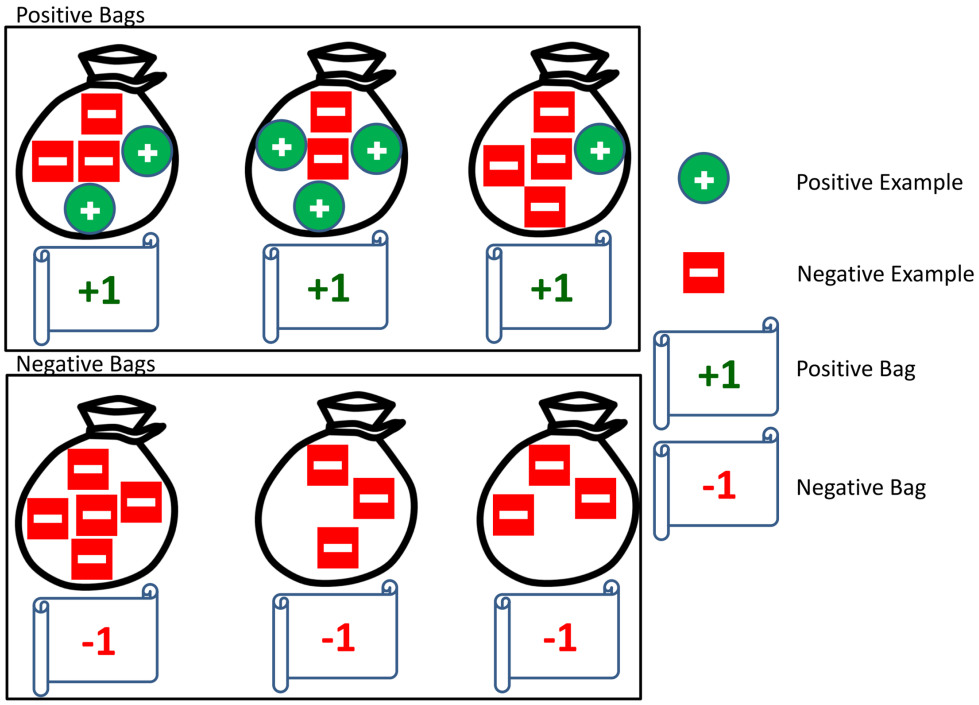

Figure 1.4. The diagrammatic representation of multiple instance learning taken from Afsar Minhas et al. (2017). The positive bag includes at least one positive instance, while the instances in the negative bag are all negative.

1.2.5 Deep learning

Deep learning is based on artificial neural networks. Rosenblatt (1958) proposed the perceptron in 1958, the first neural network that simulated the human brain. Neural networks are computational models like human neural networks designed to recognize patterns from the datasets. Recently, neural networks have been utilised across various fields, including drug discovery and PPI prediction. Different kinds of neural networks have different architectures for specific data and tasks. Convolutional Neural Networks (CNNs) (LeCun et al. 1998) used convolutional layers, pooling layers and fully connected layers to extract spatial features, which has been applied in intra-species PPI prediction (Sledzieski et al. 2023; Sledzieski et al. 2021). Graph Neural Networks (GNNs) (Scarselli et al. 2008) train and inference on graph data and have been used

for host prediction (Ma et al. 2025; Shang et al. 2022) and antibacterial discovery (Stokes et al. 2020). Long Short-Term Memory (LSTM) Network (Hochreiter 1997) was designed to capture long-range dependencies in time series or text. The Siamese neural network (SNN) is a dual architecture that shares the same weights between input vectors. LSTM and SNN have been successfully used to develop virus-human PPI models (Tsukiyama et al. 2021; Madan et al. 2022). Diffusion models are generative models that create data by adding noise and then reversing the process to reconstruct the original data (Sohl-Dickstein et al. 2015), and have been successfully implemented in protein design (Watson et al. 2023).

1.2.6 Large language models

The attention mechanism was first introduced by Bahdanau (2014) and applied in two machine translation tasks, which can learn long-range dependencies between words. The attention mechanism is the essential component in transformer architecture (Vaswani et al. 2017), it allows the model to compute weights of different parts of input sequences to capture contextual information of the entire sequence. Large language models (LLMs) based on the transformer architecture have been developed to process a variety of NLP tasks, such as generation, translation and summarization of text. In addition, LLMs have been successfully implemented in protein and genome sequences for developing protein and RNA language models, such as ESM-2 (Lin et al. 2023) and nucleotide transformer (Dalla-Torre et al. 2024).

1.2.6.1 Tokenisation and Embeddings

Tokenisation and embeddings are fundamental processes in LLMs that convert sequences into dense vectors, capture meaningful information, and enable AI models to learn and process different tasks. In NLP, tokenisation is the processing of breaking down text into tokens, words, or meaningful units for machines to learn human language. Embeddings are n-dimensional numerical vectors used to present each token in text, each component can be a number. Input sequences are tokenised to hot-encoded tokens, and embedded tokens are passed into decoder layers. Word2vec (Church 2024) is designed to train a word corpus to learn word features and make similar words have similar vector representations. Word2vec utilized tokenisation algorithms such as Continuous Bag of Words (CBOW) and the Skip-gram to obtain word embeddings. In the context of transformers, Byte-Pair Encoding (BPE) (Sennrich 2015), WordPiece (Schuster et al. 2012) and SentencePiece (Kudo 2018) are the main tokenizers algorithms. BPE was used in the transformer model GPT-2 (Radford et al. 2019), WordPiece was used for the pre-trained model Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018), SentencePiece was used in a lite BERT model ALBERT (Lan 2019) and text-to-text

model T5 (Raffel et al. 2020). Additionally, tokenisation techniques can encode biological sequences, including proteins and DNA. The pre-trained PLMs have been used to extract protein embeddings, which encode structural and evolutionary information of proteins.

1.2.6.2 Transformer

The transformer architecture was introduced by Vaswani et al. (2017). It is a model architecture for representation learning and generative modelling, and demonstrated a powerful performance in NLP tasks. Moreover, the transformer encoder model is based on position embeddings that encode the information about the order of input tokens. Due to the self-attention mechanism, all positions in the sequence can be represented by calculating interactions with any other positions simultaneously. Therefore, it can capture long-range dependencies of sequences.

Recently, many language models based on the transformer architecture have been developed. Generative models can generate text based on the input or prompts and predict missing words and the next sentence. OpenAI released Generative Pre-trained Transformer (GPT-3) (Brown et al. 2020) based on a decoder of the transformer. GPT-3 is unidirectional, meaning it generates text in a single direction and predicts each word based on all the previous words in the sequence. This unidirectional structure enables the generation of coherent and contextually relevant text. It is pre-trained on a large corpus and has strong text generation capabilities. Google developed an encoder-decoder model T5 (Raffel et al. 2020), which converted all tasks into a text-to-text format. T5 performs well on different tasks including translation and summarization.

Furthermore, transformer-based foundation models are released by training on a large scale of protein or nucleotide sequences. ESM-2 (Lin et al. 2023) is trained on 65 million unique protein sequences based on the mask language model and can be trained with a folding head to predict 3D protein structures. Nucleotide Transformer (Dalla-Torre et al. 2024) is trained on 3,202 human genomes and 850 multi-species genomes and can be fine-tuned to process sequence-related tasks.

1.2.6.3 Encoder and Decoder

The transformer architecture consists of two main components: Encoder and Decoder. The Encoder processes the input sequence to convert tokens (words) into a set of representations that capture the meaning and relationship of each word in the input sequence. The Decoder uses these representations to contextualize the input sequence and generate the output sequence. This architecture can handle dependency data and long sequences, it has been widely implemented in machine translation, text generation and question-answering. Figure 1.5 shows the transformer

architecture. Firstly, each input embedding is added with a positional embedding, ensuring that each token encodes the positional information. Token embeddings are then inputted into the N Encoder blocks. Next, the encoding information matrix, which contains the representations of all words in the sentence, is passed to the Decoder. The Decoder also has N identical layers to generate the next word using the output of the Encoder and its already generated outputs. This allows the model to consider both the input text and its own generated outputs, making the output context relevant.

The transformer model uses scaled dot-product attention, and the output matrix of the attention mechanism is (Vaswani et al. 2017):

$$Attention(Q(h),K(h),V(h)) = softmax(\frac{Q(h)K(h)^T}{\sqrt{d}})V(h) \tag{1.1}$$

Here, the input sequence vector $(h_1,...,h_n)$ is converted to a sequence of $(h'_1,...,h'_n)$ by computing interactions between all tokens in the sequences. The high-dimensional matrices of the input sequence are transformed to a smaller size Q/K/V matrix by multiplying with their learned weight matrices. The query Q, key K and value V are the projections of the input sequence and are used for attention score calculation. The query Q includes the representation vector of each position in the sequence that is used to match with any tokens in the sequence. The key K includes the representation vector of each token in the sequence that can be matched by the query token. The value V includes the token vectors that will be extracted and can be used for updating the token's final representation. The attention is computed by the dot product between query Q and key K, which is then divided by $\sqrt{d}$, where d is the dimension of the vector key K. Next, the attention score matrix is processed through the SoftMax function, determining which token should get more attention based on the query token. The final output of each token is the sum of the weighted value V vectors based on attention scores (Figure 1.6 (left)). In this way, the transformer allows each token in the sequence to be attended by the other, which can combine meaningful information from the entire input sequence and process long-range dependency data. Multi-head attention consists of multiple self-attention layers (Figure 1.6 (right)). The results from all self-attention layers are concatenated together and then as an input to a linear layer to get the final output.

1.2.6.4 Mask language model

Masked language modelling (MLM) is used to predict masking tokens in a sequence (Vaswani et al. 2017), making it particularly effective for tasks requiring a comprehensive contextual

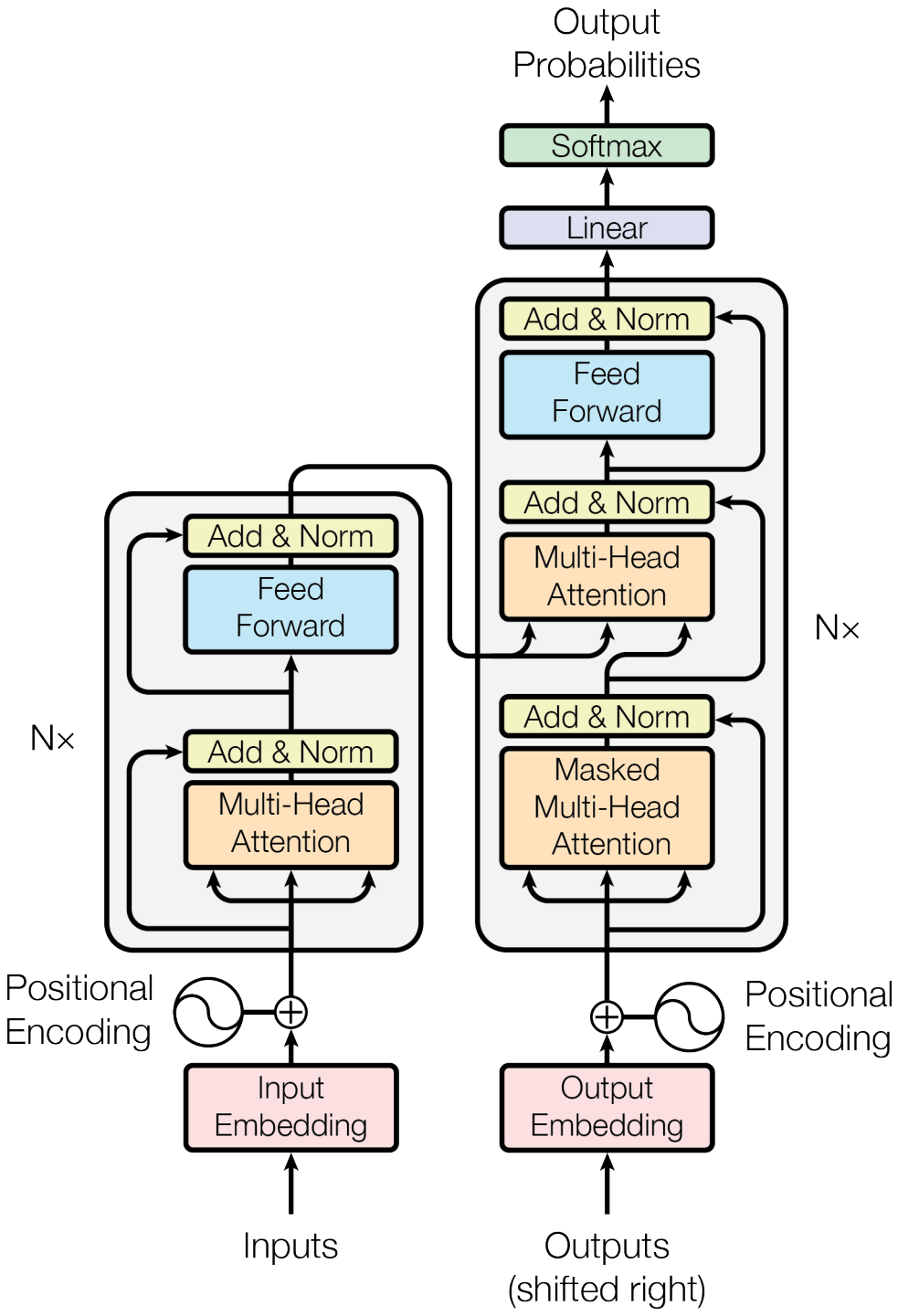

Figure 1.5. The Transformer model architecture (Vaswani et al. 2017). This is the transformer architecture using stacked self-attention. It shows the input embeddings, Encoder (left) and Decoder (right). This figure is taken from Vaswani et al. (2017).

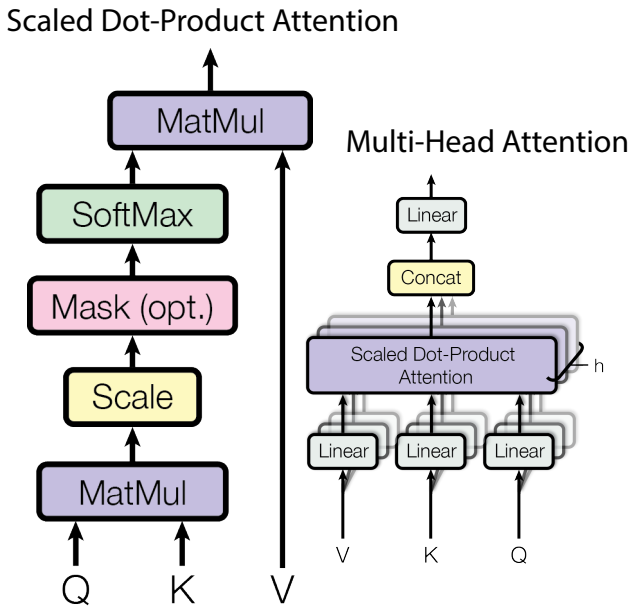

Figure 1.6. Scaled Dot-product Attention (left) and Multi-head Attention (right) (Vaswani et al. 2017). Multi-head attention is an important part of both the Encoder and Decoder. It shows the Self-Attention architecture (left) and multi-head attention (right). Query (Q), Key (K) and Value (V) are matrices for self-attention. This figure is taken from Vaswani et al. (2017).

understanding of the sequence. It is an important technique for training transformer models, including BERT (Devlin et al. 2018), GPT-2 (Radford et al. 2019) and RoBERTa (Liu 2019). In MLM, 15% of input sequences are typically randomly masked with a $< MASK >$ token during model training, and then these mask tokens are predicted based on the remaining context. Additionally, special tokens, such as $< CLS >$ (classification token) and $< EOS >$ (end-of-sequence token), are respectively used at the beginning and end of the input sequence, making the model learn contextualised information in the sequence.

Pre-trained model training using MLM can be fine-tuned to improve the performance of downstream tasks, such as next-sentence prediction. Google AI (Devlin et al. 2018) trained a BERT model with an MLM objective to learn representations that can be used to train state-of-the-art models for 11 downstream tasks. The pre-trained parameters initialise the BERT model and then are fine-tuned using datasets from the downstream tasks. It can be applied to sequence-to-sequence language generation tasks, such as question-answering, next-sentence prediction and abstract summarisation.

1.2.7 Protein language model

Language models are used to predict missing words or the next word in a sentence, and the semantics of words can be inferred based on the context of the sentence. These models can also

be used for protein sequences to understand the semantics of amino acids. Recently, protein foundation models, including ESM-1b (Rives et al. 2021), ESM-2 (Lin et al. 2023), ESM3 (Hayes et al. 2024) and ProtTrans (Elnaggar et al. 2021), have leveraged language models to train protein language models (PLMs) for predicting protein structures and capturing features encoded in protein sequences. ESM-1b was trained on a large-scale protein dataset to learn patterns in protein sequences, and it can be used to predict secondary structures, residue-residue contacts and mutational effects. ESM-2 is an improved version of the architectures and datasets of ESM-1b. It has been fine-tuned to create ESMFold, which is faster for protein structure prediction than AlphaFold2 (Jumper et al. 2021). The generative masked language model ESM3 can be applied to analyse protein sequences, structures and functions. ProtTrans leverages NLP architectures, including BERT and T5, to train protein sequences and can be used for secondary structure prediction.

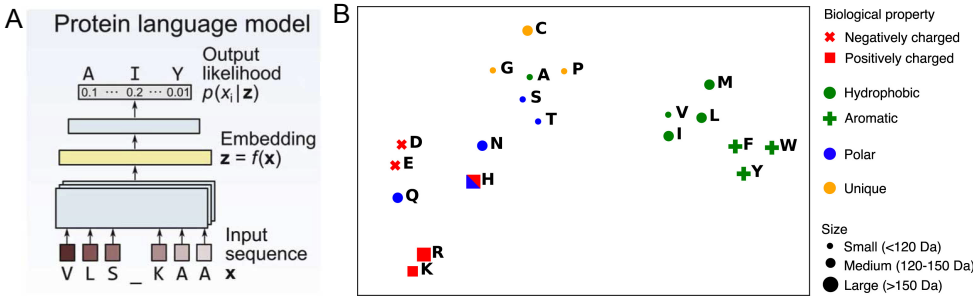

Figure 1.7. The flowchart of protein language model (A) (Hie et al. 2022) and t-SNE visualization of amino acid embeddings (B) (Rives et al. 2021). **A**. The flowchart illustrates how input amino acid sequences are processed and converted into likelihoods that represent the features of the input sequences. **B**. The t-SNE visualization depicts the output embeddings of the PLM ESM-1b, showing that amino acids with similar biological properties cluster together.

In Figure 1.7, the left panel shows the input and output of a PLM. The input is the protein sequence, followed by tokenization and the embedding for the input protein. Finally, the output represents the likelihood for each amino acid. In the right panel, t-SNE visualises the different groups of amino acids represented by the output embeddings of the pre-trained transformer model. These amino acids are clustered into different groups, indicating that the pre-trained PLM captures features of amino acid properties. Currently, pre-trained PLMs can effectively encode protein sequences and are widely implemented for protein representations. For each protein sequence, representative vectors are extracted by pre-trained PLMs and can be used as protein features for DL model training (Sledzieski et al. 2023). Moreover, the pre-trained PLM has been fine-tuned to jointly encode paired protein sequences to learn intra-species and inter-species PPIs (Liu et al. 2024b). In summary, the pre-trained PLMs can be applied to process downstream tasks or fine-tuned for specific tasks.

1.2.8 ESM-1b Transformer

In Chapter 2, we use the PLM ESM-1b (Rives et al. 2021) for protein feature representation. In Chapter 3 and 4, we retrain the PLM ESM-2 (Lin et al. 2023) on paired proteins to predict PPIs. ESM-1b is a deep Transformer encoder model trained on 250 million protein sequences from UniParc (Leinonen et al. 2004), consisting of 86 billion amino acids, with the model parameters ranging from 6 to 650 million. ESM-2 is an improved version of ESM-1b with a more efficient architecture, larger dataset and increased parameters ranging from 8 million parameters to 15 billion parameters. ESM-2 replaces the static sinusoidal positional encoding with a learned Rotary Position Embedding (RoPE), enabling inference beyond the context window during training. This section focuses on the architectures and training details of ESM-1b as a foundation for understanding ESM models.

The ESM-1b model utilizes the transformer architecture. It takes amino acid characters as input and processes them using alternating self-attention layers and fully connected feed-forward connections. The self-attention mechanisms enable the model to consider all other positional information of tokens when processing each position. In the network, complex representations integrating context from the entire sequence are built by self-attention. Each position in the sequence is represented by the sum of the weighted other positions. Therefore, each pair of interactions between all positions can be constructed based on self-attention architectures. As such, it can capture long-range dependencies within protein sequences and more effectively understand the protein properties (Khandelwal et al. 2018).

In the pre-training model, ESM-1b is trained with the MLM objective, following the BERT masking pattern with dynamic batching. For each protein sequence, a set of indices is randomly sampled for masking, where the mask token is used to replace the true token at each index to identify relationships between the masked indices and the unmasked parts of the sequence. During training, 15% of amino acids from the input sequence are randomly sampled, and these tokens will be replaced with special mask tokens (details see Section 1.2.6.4). Regarding masking of 15% amino acids, four masking strategies are investigated in ESM-1b: (1) all are replaced with a mask token; (2) all are replaced with a uniform random amino acid; (3) all are replaced with an amino acid according to their frequency; (4) 80% are changed to a mask token, 10% are replaced with a uniform random amino acid, and 10% remain unchanged. The fourth strategy achieves the best performance in the model training. This model is trained on a large scale of protein sequences, using masked amino acids as self-supervision to predict the masked tokens. A robust set of hyperparameters for ESM-1b was optimized on 100M parameter models. The hyperparameters were then scaled to train a model with 33 layers (650M parameters) on representative sequences from the UniParc clustering at 50% sequence identity (Rives et al. 2021).

1.2.9 Bi-Encoder and Cross-Encoder

Bi-Encoder and Cross-Encoder are two NLP techniques for training Transformer models. Transformer models can be trained or fine-tuned using either approach depending on the specific tasks. In Chapter 3 and 4, we use Cross-Encoder for PLM training. Figure 1.8 shows the architectures of Bi-Encoder and Cross-Encoder, respectively. The Bi-Encoder processes sentences separately and calculates cosine similarities between any pairs of sentences for sentence similarities. In contrast, the Cross-Encoder processes two sentences together within a single encoder and outputs a similarity score ranging from 0 to 1, it is computationally intensive as it needs to process all sentence pairs.

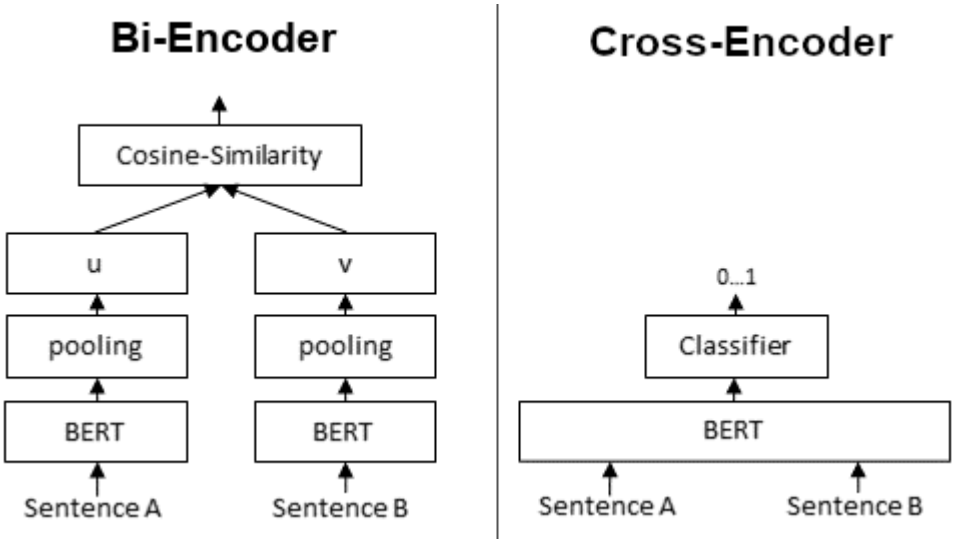

Figure 1.8. Bi-Encoder (left) and Cross-Encoder (right). This figure is taken from https://www.sbert.net/examples/applications/cross-encoder/README.html

The Bi-Encoder known as Sentence-Transformer framework (Reimers et al. 2019), is designed for efficient sentence training. It transforms sentences or text pairs into vector representations for various downstream tasks such as semantic search, text similarity and paraphrase mining. The Bi-Encoder is suitable for large-scale datasets such as information retrieval, where it avoids repeated sentence embedding calculations, offering a faster and more efficient architecture. The Cross-Encoder usually performs better than the Bi-Encoder as it jointly trains input pairs and learns contextually relevant features between inputs. Moreover, the Bi-Encoder and the Cross-Encoder can be combined for specific tasks such as information retrieval. Firstly, for a search query/question, the Bi-Encoder retrieves a large document collection to get potential relevance with the query. Next, a re-ranker based on the Cross-Encoder is used to retrieve highly relevant candidates. Finally, a list of ranked hits is obtained for the query.

1.3 Thesis Outline

My thesis is in the format of primary research articles:

1.3.1 Prediction of virus-host associations using protein language models and multiple instance learning

This chapter corresponds to my published paper: Dan Liu, Francesca Young, Kieran D. Lamb, David L. Robertson, Ke Yuan. Prediction of virus-host Associations using Protein Language Models and Multiple Instance Learning published in PLOS Computational Biology, November 19, 2024. doi: https://doi.org/10.1371/journal.pcbi.1012597. A preprint was made available on April 7, 2023:https://doi.org/10.1101/2023.04.07.536023. The code is available at GitHub Link: https://github.com/liudan111/EvoMIL

This chapter uses a protein language model to extract viral protein embeddings, then trains a multiple instance learning model to predict hosts for prokaryotic and eukaryotic viruses and identify important viral proteins contributing to virus-host associations. I collected datasets, wrote codes, trained models and conducted result analysis. I wrote the manuscript with assistance from Francesca Young. All authors contributed to discussions and reviewed the manuscript. David L. Robertson and Ke Yuan conceptualized the study, supervised the project and helped with editing.

1.3.2 PLM-interact: extending protein language models to prediction of protein-protein interactions

This chapter is taken from a preprint: Dan Liu, Francesca Young, Kieran D. Lamb, Adalberto Claudio Quiros, Alexandrina Pancheva, Crispin Miller, Craig Macdonald, David L. Robertson, and Ke Yuan. "PLM-interact: extending protein language models to predict protein-protein interactions." bioRxiv (2024): November 27, 2024. doi: https://doi.org/10.1101/2024.11.05.622169. It is under review. The code is available at GitHub Link: https://github.com/liudan111/PLM-interact

This chapter shows how a protein language model targeting a single protein is extended to jointly encode paired proteins for improving protein-protein interaction prediction in intra-species. The results demonstrated that our PPI model PLM-interact achieved the best performance on held-out species PPI datasets and can detect mutational impacts in human PPIs. I collected datasets,

wrote codes, trained models, conducted result analysis and wrote the manuscript. Craig Macdonald, David L. Robertson and Ke Yuan supervised the project. Francesca Young, Kieran D. Lamb, Adalberto Claudio Quiros, Alexandrina Pancheva and Craig Macdonald provided feedback on the manuscript. David L. Robertson and Ke Yuan conceptualized the study and helped with editing.

1.3.3 Predicting virus-host protein-protein interactions using PLM-interact

This chapter is to implement the PPI model PLM-interact to predict virus-human PPIs. Here, the virus-human PPI benchmarking results are taken from the preprint: Dan Liu, Francesca Young, Kieran D. Lamb, Adalberto Claudio Quiros, Alexandrina Pancheva, Crispin Miller, Craig Macdonald, David L. Robertson, and Ke Yuan. PLM-interact: extending protein language models to predict protein-protein interactions." bioRxiv (2024): November 27, 2024. doi: https://doi.org/10.1101/2024.11.05.622169. We intend to release the rest of this chapter as a preprint: Dan Liu, Francesca Young, Kieran D. Lamb, Adalberto Claudio Quiros, Alexandrina Pancheva, Crispin Miller, Craig Macdonald, Ke Yuan and David L. Robertson. Predicting virus-host protein-protein interactions using an interaction-aware protein language model.

We constructed a virus-human PPI dataset from seven public PPI databases. We used three strategies to split the PPI dataset and then train a PLM-interact-VH model for virus-human PPI prediction. Holding out virus families as test sets is the most challenging prediction task compared with random splitting and Park and Marcotte's C1/C2/C3. Fine-tuning the human PPI model with virus-human PPIs improves virus-human PPI prediction. This chapter investigated the impacts of three data-splitting strategies on the model's performance and discussed how to train a generalizable model using PPIs from multiple species. I wrote this chapter, collected data and conducted coding and result analysis. Francesca Young, Kieran Lamb, Craig Macdonald, Ke Yuan and David L. Robertson contributed to discussions and suggestions. Ke Yuan and David L. Robertson conceptualized the study, supervised the project and helped with editing.

Chapter 2

Prediction of virus-host associations

An inefficient virus kills its host. A clever virus stays with it.

James Lovelock

2.1 Abstract

Predicting virus-host associations is essential to determine the specific host species that viruses interact with, and discover if new viruses infect humans and animals. Currently, the host of the majority of viruses is unknown, particularly in microbiomes. To address this challenge, we introduce EvoMIL, a deep-learning method that predicts the host species for viruses from viral sequences only. It also identifies important viral proteins that significantly contribute to host prediction. The method combines a pre-trained large protein language model (ESM) and attention-based multiple instance learning to allow protein-orientated predictions. Our results show that protein embeddings capture stronger predictive signals than sequence composition features, including amino acids, physicochemical properties, and DNA k-mers. In multi-host prediction tasks, EvoMIL achieves median F1 score improvements of 10.8%, 16.2%, and 4.9% in prokaryotic hosts, and 1.7%, 6.6% and 11.5% in eukaryotic hosts. EvoMIL binary classifiers achieve impressive AUC over 0.95 for all prokaryotic hosts and range from roughly 0.8 to 0.9 for eukaryotic hosts. Furthermore, EvoMIL identifies important proteins in the prediction task, capturing key functions involved in virus-host specificity.

2.2 Introduction

Advances in sequencing technologies, particularly metagenomics, have resulted in the identification of many new viruses. However, more than 90% of the virus sequences held in publicly available databases are not annotated with any host information (Roux et al. 2019). Currently, there are no high-throughput experimental methods that can definitively assign a host to these uncultivated viruses. With a growing number of viruses being discovered, relying only on experiments to identify virus-host associations is a limiting step in this important challenge.

A number of computational approaches have been developed to predict unknown virus-host species associations. The coevolution of a virus and its host leave signals in virus genomes arising from the virus-host interaction. These signals have been exploited for in silico prediction of virus-host associations from virus genomes alone and fall into two broad types: 1) alignment-based approaches that search for homology such as prophage (Roux et al. 2015), CRISPR-cas spacers (Staals et al. 2013; Horvath et al. 2010); 2) Alignment-free methods that use features such as k-mer composition, codon usage, CpG content etc. to measure the similarity between viral and host sequences or to other viruses with a known host(Greenbaum et al. 2008). To date, no computational approaches consider the structure of proteins from viruses for host species prediction purposes.

Here, we present a virus-host prediction model combining protein language models (PLMs) and multiple instance learning (MIL). Transformers are self-supervised deep learning models (Vaswani et al. 2017) that learn the relationships among words within a sentence, and now dominate the field of natural language processing. More recently, the same architecture has been applied in biology, where words are replaced by amino acids and sentences by protein sequences. These transformer-based protein language models generate protein embeddings that encode structural features inferred from amino acid sequences based on large-scale protein databases (Rives et al. 2021). Protein language models are trained on publicly available protein sequence archives and learn biological information from physicochemical properties of the individual amino acids to structural and functional information about proteins. Multiple instance learning (MIL) is a form of supervised learning that was developed for image processing tasks (Maron et al. 1997). Instead of using individually labelled instances for classification, multiple instances are arranged together in a bag with a single label and classified together. We use attention-based MIL (Ilse et al. 2018), which has the additional advantage of weighting instances in a bag, thereby indicating the importance of each instance in prediction.

The combination of the two approaches is particularly suited for virus-host prediction, as virus proteins collectively contribute to the association with a host. Instead of relying on predefined features, protein language models provide automatically learned features, free from design bi-

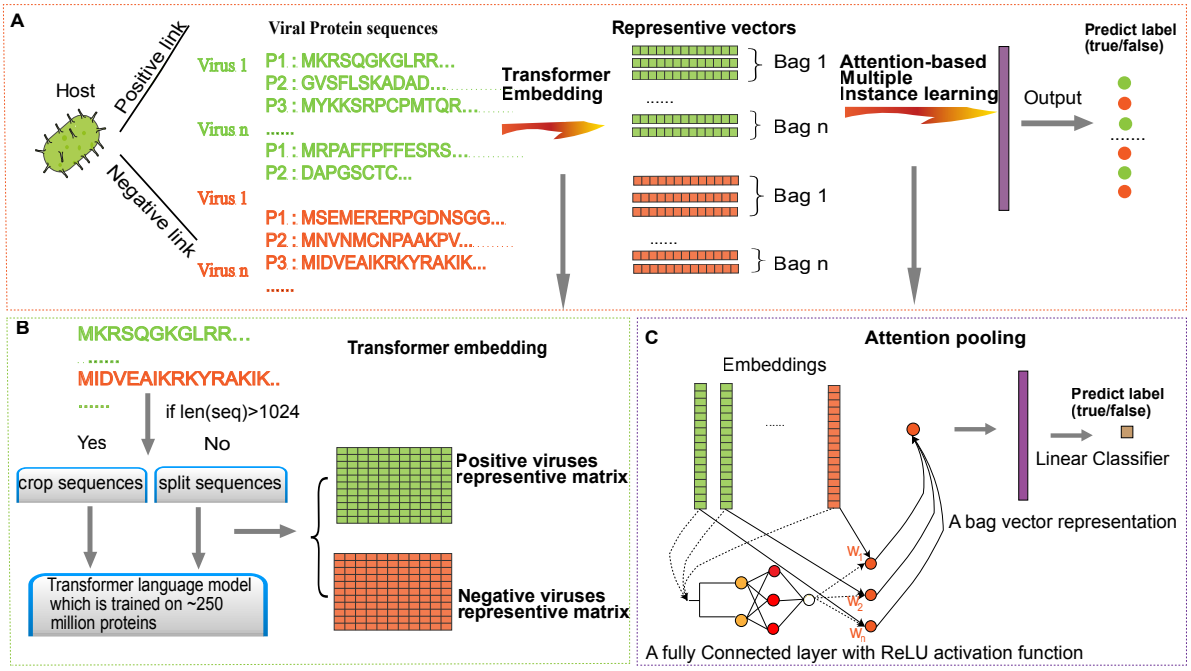

Figure 2.1. A diagrammatic representation of the EvoMIL method. (A) Protein sequences of viruses and virus-host associations are collected from the VHDB (Mihara et al. 2016). For each host, we collect the same number of positive and negative viruses, and then embeddings of protein sequences from viruses are obtained by the pre-trained transformer model (Rives et al. 2021), which are features for host predictions based on attention-based MIL; (B) Protein sequences of viruses are split to sub-sequences, which are used as input to the pre-trained transformer model to obtain the corresponding embeddings; (C) There is a host label for a set of protein sequences on each virus, and attention-based MIL is applied to train the model for each host dataset by protein embeddings of viruses. Finally, we predict the host label for each virus and assign an instance weight that represents the importance of each protein for the virus.

ases and limitations of the previous approaches. The ability to measure similarity and differences between protein sequences further boosts prediction performance through multiple instance learning, where viral proteins enabling interaction with hosts are highlighted through unbiased weight estimation.

In this chapter, we introduce EvoMIL, a method for predicting virus-host associations by combining the (Evo)lutionary Scale Modeling with (M)multiple (I)instance (L)earning (Figure 2.1). EvoMIL uses the model ESM-1b (Rives et al. 2021) to transform viral protein sequences into embeddings (i.e., numerical vectors) that are then used as features for virus-host classification. Multiple-instance learning allows us to consider each virus as a bag of proteins. We demonstrate that the embeddings capture the host signal from the viral sequences, achieving high prediction scores for both prokaryotic and eukaryotic hosts at the species level. Furthermore, attention-based MIL enables us to identify which proteins are highly important in driving prediction and by implication are important to virus-host specificity.

2.3 Materials and Methods

2.3.1 Data

The current study uses datasets collected from the Virus-Host database (VHDB), (https://www.genome.jp/virushostdb/) (Mihara et al. 2016). The VHDB contains a manually curated set of known species-level virus-host associations collated from a variety of sources, including public databases such as RefSeq, GenBank, UniProt, and ViralZone and evidence from the literature surveys. For each known interaction, this database provides NCBI taxonomic ID for the virus and host species and the Refseq IDs for the virus genomes. We downloaded datasets on 20-10-2021, along with the associated FASTA files containing the raw viral genomes and the FAA file with translated CDS for each viral protein. At this point, the VHDB contained 17,733 associations between 12,650 viruses and 3740 hosts that were used to construct binary datasets for both prokaryotic and eukaryotic hosts.

2.3.2 Constructing balanced binary dataset

In this study, we retrieve the reference genome from the Refseq genome (Tatusova et al. 2014) for each virus. The aim is to reduce the amount of redundant or similar sequences in the datasets. We obtain known virus-host associations for prokaryotic and eukaryotic hosts from the Virus-Host database VHDB (Mihara et al. 2016), and remove viruses whose protein sequences do not exist in the Refseq database.

Balanced binary datasets were constructed for each host with an equal number of positive and negative virus sets to obtain a balanced data set for binary classification tasks. For both the prokaryotic and eukaryotic datasets, we collected 4696 associations between 4696 viruses and 498 prokaryotic hosts at the species level; 9595 positive associations from 9595 viruses and 1665 eukaryotic hosts at the species level.

For each binary virus-host association data set, viruses can be represented by $V = \{V_1, V_2, ..V_V\}$, and the host is represented by H. The positive labels consisted of known associations from VHDB (Mihara et al. 2016), and the set of positive viruses can be represented by $V_{pos} = \{V_1, V_2, ..V_P\}$, where $P = \frac{V}{2}$. Most phages are likely to infect a range of hosts belonging to the same taxonomy (Flores et al. 2011; Ben-Hur et al. 2006), therefore, to mitigate possible errors, some researchers (Leite et al. 2018; López et al. 2019) construct negative virus-host associations by selecting those viruses from the remaining viruses that do not infect the given host, instead of relating to hosts with different taxa from the given host. Using this method, we construct

putative negative virus-host associations by collecting viruses from a specific range. As for a given host H in the prokaryotic datasets, all viruses are represented by $V_{pro} = \{V_1, V_2, ..V_T\}$, where T is the total number of viruses associated with prokaryotic hosts, we collect negative viruses $V_{neg} = V_{pro} - V_{pos}$, which are not related with the given host; and host species taxonomy of V_{neg} are different from host H. On top of that, in order to get the balanced datasets for model training and testing, the size of datasets V_{neg} is equal to V_{pos}, we randomly choose viruses from putative negative virus set V_{neg}, finally we get the same number of positive and negative virus datasets, that can be represented by $V_{neg} = \{V_{P+1}, V_{P+2}, ..V_V\}$. Similarly, we can create a balanced binary dataset for each eukaryotic host. The number of viruses related to the given host is at least 50 and 125 for the prokaryotic and eukaryotic datasets. In addition, there exist some segmented and non-segmented viruses; segmented viruses have multiple RefSeq sequences and need to be combined to represent complete viral sequences. After setting the threshold for the number of viruses at the species level, combining segmented viruses and removing redundant non-segmented viruses, we have 15 prokaryotic datasets and 5 eukaryotic datasets.

2.3.3 Feature extraction

2.3.3.1 ESM-1b

The Pre-trained ESM-1b model is used to transform protein sequences into fixed-length embedding vectors that are used as features for downstream tasks such as binary and multi-class classification. For n protein sequences $(h_1, ..., h_n)$ input into ESM-1b (Rives et al. 2021), we obtain n embedding vectors $\mathbf{X} = (\mathbf{x_1}, ...\mathbf{x_n})$, each with a dimension of 1280. In the pre-training process, the representation is projected to log probabilities $(y_1, ..., y_n)$, the model posterior of amino acid at position i is represented by a softmax over y_i, the output embeddings $(h_1, ..., h_n)$ are applied in downstream tasks, (the parameters are –repr_layers 33 –include mean per_tok). Here, we adapt virus embedding vectors to the supervised attention-based MIL learning tasks (EvoMIL), and the input feature dimension of ESM-1b is 1280 generated by esm1b_t33_650M_UR50S. EvoMIL is based on a fully connected layer with an input size of 1280 (ESM-1b embedding size) and an output size of 800, and then using a linear classifier and the output size is the number of prediction classes.

There are two special cases of sequences that need to be pre-processed to be suitable as input for ESM-1b:

1. In NCBI, some protein sequences include the amino acid J, which is used to refer to unresolved leucine (L) or isoleucine (I) residues. However, the ESM-1b model does not include the token 'J' and they must be removed before processing. In order to remove J

from NCBI sequences, we randomly replace J with either Leucine(L) or Isoleucine(I).

2. The ESM-1b model can only process sequences with a maximum of 1024 tokens including the beginning-of-sequence (BOS) token and the ending-of-sequence (EOS) token, meaning that the maximum length of the protein sequence is 1022 amino acids. The parameter '–truncation' can be used to crop a longer sequence, but this will result in a loss of some part of the sequence information. In order to include information from the entire protein sequence, we split longer sequences and process the sections simultaneously. For protein sequences longer than 1022, we split the sequence into lengths of 1022 amino acids. If the final section is shorter than 25 amino acids, it is discarded as it is considered too short to include meaningful information. So that for a given protein sequence H, of length $len(H)$:

   if $len(H) < 1022 + 25$ then truncate the sequence

   else if $len(H) \geq 1022 + 25$ split the sequence.

   For example, if we have a sequence of length 2049 ($2 \times 1022 + 25$), it will be cut into two sub-sequences with the length of 1022, and one sub-sequence with the length of 25. Resulting in three embedding vectors for the single protein sequence that are all assigned to the same label as the parent protein and will be considered instances in the attention-based MIL.

2.3.3.2 K-mers

K-mer composition vectors were generated for each of the CDS regions of a virus and used as an alignment-free representation of a sequence. A k-mer is a sub-sequence of length k and is generated for each position in a sequence. These features are obtained by calculating the frequency of each of the possible k-mers in a sequence. To minimise the effect of sequence length, the resulting vector is normalized so that its sum is equal to 1. Here, we extract k-mer features from sequences corresponding to DNA, amino acids (AA) and their physico-chemical properties (PC) (Young et al. 2020). DNA and amino acid sequences of the CDS regions were downloaded from the NCBI database, and used directly to extract AA and DNA k-mers. To extract PC k-mers from protein sequences we first re-label each amino acid as one of seven groups based on its physico-chemical properties, as described by Shen et al. (2007): AGV, C, FILP, MSTY, HNQW, DE, and KR, and then extract k-mers using the seven group labels as the alphabet. To generate k-mer composition vectors of reasonable length for MIL, we use k-mer lengths of 5, 2 and 3, respectively, for the DNA, AA and PC sequences. This results in DNA_5, AA_2, PC_3 feature sets of dimensions of 1024, 400, and 343, respectively.

2.3.4 Attention-based Multiple instance learning

We represent a virus as a set of protein embedding instances $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_M\}$. In a binary classification setting, the label of the virus is $Y \in [0,1]$. If $Y = 1$, the virus is known to be associated with the host, otherwise, the virus is not associated with the host, but the instance label $\{y_1, y_2, y_3, ..., y_M\}$ is unknown. Here, the instance represents the protein of a virus. We only have the host label for each virus, but which proteins of a virus lead to virus-host associations are unknown. Multiple instance learning (MIL) is used to predict a label for a bag with a set of instances, so it was used to predict a host label for a set of proteins from a virus. The label Y can be represented as follows:

$$Y = \begin{cases} 0, & \text{iff} \sum_m y_m = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{2.1}$$

The MIL model can be interpreted as a probabilistic model in which the label of the bag is distributed according to a Bernoulli distribution with the parameter $\theta(X) \in \{0,1\}$, which is the probability of a virus being associated with a host, i.e., Y = 1.

$$p(Y|X) = \theta(X)^Y (1 - \theta(X))^{1-Y}, \tag{2.2}$$

The model can be trained by maximising the Bernoulli likelihood function, which is equivalent to minimising the negative entropy function.

Given a bag of instance X, the scoring function $\theta(X)$ can be written as Eq (2.3) (Salakhutdinov et al. 2017):

$$\theta(X) = g(b^T \sum_{x_m \in X} a_m[\varphi(\mathbf{x}_m)]), \tag{2.3}$$

where g is the sigmoid function. φ is a neural network that transforms protein embedding to a lower dimensional space. φ is implemented as a fully connected layer with ReLU activation functions. It has an input size matching the embedding dimension from ESM-1b and k-mers, and an output size universally set to 800. $b \in \mathbb{R}^{800 \times 1}$ is the weight of the binary classifier.

a_m are the attention weights of instance m computed as

$$a_m = \frac{\exp\{\boldsymbol{w}^\top \tanh(\boldsymbol{V}\boldsymbol{u}_m^\top)\}}{\sum_{j=1}^{M} \exp\{\boldsymbol{w}^\top \tanh(\boldsymbol{V}\boldsymbol{u}_j^\top)\}}, \tag{2.4}$$

where $\boldsymbol{u}_m = \boldsymbol{\varphi}(\mathbf{x}_m)$. $V \in \mathbb{R}^{L\times M}$ and $\boldsymbol{w} \in \mathbb{R}^{L\times 1}$ are learnable parameters. $\tanh(.)$ is the hyperbolic tangent function. We used these weights to quantify the importance of proteins to host prediction.

In the multi-class classification case, g is the softmax function and $b \in \mathbb{R}^{800\times K}$, with K being the number of hosts.

2.3.5 Gene Ontology

The Gene Ontology (GO) is used to describe gene function from different organisms, which includes three GO domains: biological process, cellular component and molecular function (Consortium 2004). GO annotations are links between gene products and GO terms, and the structure of GO can be represented by a graph, where each node is a GO term and nodes are connected by edges. InterProScan (Jones et al. 2014) is a tool to obtain GO terms by classifying query proteins to protein families within InterPro's database. To analyze the protein functions of important virus proteins identified by EvoMIL, we used InterProScan to obtain GO annotations of virus proteins.

2.3.6 Taxonomic tree

We used the NCBI interface https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi to generate taxonomic trees for 22 prokaryotes and 36 eukaryotes when given taxonomy IDs, and then applied the R package ggtree (Yu et al. 2017) to obtain the trees in Figure 2.7, Figure 2.9 and Figure 2.10.

2.3.7 Experimental settings of prokaryotic host prediction approaches in benchmarking

We used iPHoP v1.3.3 (Nov 2023 version, (Roux et al. 2022)) and recommended database Sept_21_pub_rw for the test virus genomes. BLASTn and CRISPR predictions were based on BLASTn (v2.12.0+) between the test virus genomes and the iPHoP_db_Sept21 BLAST and

spacer database, respectively. BLASTp (v2.12.0+) was used to calculate the identity between the test virus and the prokaryotic host protein sequences from our training set, retrieved from NCBI. A maximum of 200 protein sequences per host were downloaded, and the host with the top-ranking match was selected as the predicted host. SpacePHARER predictions were obtained by running the 'predictmatch' function from SpacePHARER v5.c2e680a (Zhang et al. 2021) with default parameters. For WIsH (v1.0, (Galiez et al. 2017)) predictions, virus genomes were compared to the iPHoP_db_Sept21 WIsH database with a maximum p-value of 0.2. For PHP (July 2021 version, (Lu et al. 2021)), using iPHoP_db_Sept21 PHP database to predict hosts for virus genomes. VirHostMatcher (Apr 2018 version, (Ahlgren et al. 2017)) was used to get the s2* similarity score between each pair of virus and host similarities. For the prediction of VirHostMatcher-Net, we used VirHostMatcher-Net(July 2021 version, (Wang et al. 2020)) with complete genome mode and default parameters. vHULK (v1.0.0, (Amgarten et al. 2020)) host species predictions were obtained by using default parameters.

2.3.8 Evaluation

To evaluate our classifiers, we use six evaluation metrics, including AUC, accuracy, F1 score, sensitivity, specificity and precision as defined below. The two main evaluation metrics used in our analysis and explanation are AUC and accuracy. The area under the receiver operating characteristic curve (AUC) is used to evaluate machine learning performance; accuracy is the ratio of the number of correctly predicted samples to the total number of samples.

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the parameters used to calculate specificity, sensitivity, and accuracy. True positive(TP): predicted label and known label are positive. True Negative(TN): the predicted label and known label are negative. False Positive (FP): The predicted label is positive, but the actual label is negative. False Negative(FN): The predicted label is negative, but the actual label is positive. Sensitivity is the percentage of positive samples that are predicted correctly; specificity is the percentage of negative samples that are predicted as negative. The F1 score is the harmonic mean of precision and recall. Here, we set average='macro' to calculate the F1 score and precision for each label and get their unweighted mean. The formula of the evaluation indices is as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{2.5}$$

$$Sensitivity = \frac{TP}{TP+FN}, \tag{2.6}$$

$$Specificity = \frac{TN}{TN+FP}, \tag{2.7}$$

$$Precision = \frac{TP}{TP+FP}, \tag{2.8}$$

$$Recall = \frac{TP}{TP+FN}, \tag{2.9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{2.10}$$

2.4 Results

2.4.1 Dataset for predicting virus-host association

Balanced binary datasets were generated from known virus-host associations documented in the Virus-Host database (VHDB) (Mihara et al. 2016) for all hosts with a minimum threshold number of associations. These datasets consist of either all prokaryotic or all eukaryotic viruses. 'Positive' viruses are those that are reported to be associated with the given host species. A matching number of 'negative' viruses are randomly sampled from all other prokaryotic or eukaryotic viruses. The prokaryote datasets consist of nearly all dsDNA (double-stranded DNA) viruses which have 45 to 212 proteins coded in their genomes (Table 2.1), while the eukaryotic datasets include many RNA viruses that contain fewer proteins, ranging from 2 to 23 protein sequences (Table 2.2). The performance of MIL improves with higher numbers of instances in each bag, therefore we need to increase the threshold of the number of viruses in the eukaryotic training datasets to achieve similar performance with MIL. Accordingly, we set a threshold for the minimum positive dataset size to 50 and 125 viruses for constructing prokaryotic and eukaryotic binary datasets, respectively. The aim of setting the threshold is to generate a sufficient number of training samples for MIL training on prokaryotic and eukaryotic hosts, respectively. Finally, we generated 15 prokaryotic host datasets and 5 eukaryotic host datasets for the binary classification tasks.

Table 2.1. Prokaryotic hosts: hostname and the number of viruses associated with the host

Index	Host name	Number of viruses	Mean_protein	Min_protein	Max_protein
1	Mycolicibacterium smegmatis	838	105	58	258
2	Escherichia coli	474	134	4	611
3	Salmonella enterica	231	117	15	276
4	Lactococcus lactis	185	56	22	194
5	Pseudomonas aeruginosa	182	91	3	375
6	Gordonia terrae	176	96	57	233
7	Staphylococcus aureus	152	86	18	233
8	Klebsiella pneumoniae	121	95	26	299
9	Cutibacterium acnes	81	45	42	48
10	Enterococcus faecalis	80	85	23	224
11	Bacillus thuringiensis	62	163	30	304
12	Acinetobacter baumannii	60	106	29	326
13	Vibrio cholerae	59	77	6	230
14	Arthrobacter sp. ATCC 21022	55	76	26	114
15	Bacillus cereus	54	127	25	292
16	Microbacterium foliorum	39	60	25	155
17	Erwinia amylovora	39	212	49	345
18	Cellulophaga baltica	36	87	13	198
19	Ralstonia solanacearum	35	56	3	343
20	Synechococcus sp. WH 7803	33	200	51	292
21	Listeria monocytogenes	32	122	38	196
22	Paenibacillus larvae	31	72	56	92

The table of 22 prokaryotic datasets shows the host species name, the number of viruses associated with each host, and the average, minimal and maximal number of protein sequences of viruses associated with each host.

Table 2.2. Eukaryotic hosts: hostname and the number of viruses associated with the host

Index	Host name	Number of viruses	Mean_protein	Min_protein	Max_protein
1	Homo sapiens	1321	8	1	262
2	Solanum lycopersicum	277	4	1	13
3	Sus scrofa	219	10	1	184
4	Bos taurus	202	14	1	236
5	Mus musculus	174	18	1	233
6	Nicotiana benthamiana	123	4	1	13
7	Gallus gallus	109	15	1	261
8	Chlorocebus aethiops	108	23	1	220
9	Nicotiana tabacum	81	4	1	13
10	Macaca mulatta	73	18	1	223
11	Canis lupus	72	10	1	233
12	Phaseolus vulgaris	69	4	1	13
13	Chenopodium quinoa	66	3	1	13
14	Felis catus	57	12	1	233
15	Capsicum annuum	55	3	1	8
16	Equus caballus	54	22	1	236
17	Zea mays	53	5	1	14
18	Ovis aries	51	13	1	150
19	Vitis vinifera	49	4	1	12
20	Rattus norvegicus	48	12	1	233
21	Abelmoschus esculentus	47	3	1	9

Continued on next page

Table 2.2 – continued from previous page

Index	Host name	Number of viruses	Mean_protein	Min_protein	Max_protein
22	Solanum tuberosum	46	3	1	10
23	Pan troglodytes	45	10	1	169
24	Triticum aestivum	43	5	1	14
25	Ipomoea batatas	42	5	1	11
26	Mesocricetus auratus	42	5	1	11
27	Glycine max	41	4	1	14
28	Anas platyrhynchos	40	3	1	29
29	Aedes albopictus	36	2	1	11
30	Vicia faba	34	4	1	9
31	Pisum sativum	33	3	1	8
32	Nicotiana clevelandii	32	4	1	13
33	Cucumis sativus	32	4	1	12
34	Vigna unguiculata	32	4	1	13
35	Beta vulgaris	32	5	1	10
36	Capra hircus	31	16	1	150

The table of 36 eukaryotic datasets shows the host species name, the number of viruses associated with each host, and the average, minimal and maximal number of protein sequences of viruses associated with each host.

To evaluate the performance of binary models, we created a balanced set of positive and negative samples. For negative samples, we used two different strategies to sample the negative viruses from those with no known association with each host identified above to create balanced binary datasets. Given that the actual association is unknown, this is susceptible to false negative labels. **Strategy 1** was used to establish the concept of EvoMIL. We sampled the negative viruses from all viruses that are in different genera than viruses in the positive dataset, with the aim of minimising the false negative viruses in the dataset. **Strategy 2** aimed to make the task progressively more challenging using the fact that as a result of coevolution and co-speciation similar viruses tend to infect similar hosts. Here we selected the negative viruses from those that infect hosts in the same taxonomic rank as the positive host, from phylum to genus, thereby meaning the classifier had to distinguish between more and more similar viruses. For **Strategy 2** the negative samples and positive samples are more likely to share proteins exhibiting structural mimicry(Honig et al. 2021), so it will be challenging to train classifier models and make the binary models sufficiently sensitive to capture the difference between positive and negative samples. The number of viruses related to each host is shown in Table 2.1. The largest prokaryotic dataset is *Mycolicibacterium smegmatis* with 838 known viruses, followed by *Escherichia coli* with approximately half the number of viruses. For the eukaryotic datasets, *Homo sapiens* have by far the largest number of known virus species (1321) with the next highest being the tomato (*Solanum lycopersicum*) at 277, see Table 2.2. The distribution of the top 10 virus families can be found in Figure 2.2. Approximately 60% of viruses associated with prokaryotes belong to the Siphoviridae family (see Figure 2.2 A), whereas the Geminiviridae, Picornaviridae and Papillomaviridae families are the top three families in eukaryotic hosts, each accounting for roughly 18% of viruses associated with eukaryotic hosts (see Figure 2.2 B). Note that viruses associated with eukaryotes are more diverse than those associated with prokaryotes.

2.4.2 EvoMIL achieves high performance for binary virus-host prediction

Embedding vectors for each of the proteins of a virus generated with the protein language model, ESM1b, were used as an instance in a virus bag for MIL. These labelled bags were used to train the MIL model using 5-fold cross-validation on 80% of the datasets, then each 5-fold model's performance was evaluated on the remaining 20% of the datasets. Each model is evaluated with a range of metrics: AUC, accuracy, F1 score, sensitivity, specificity, and precision. We evaluated the predictive performance of EvoMIL for binary classification using the datasets generated with both **Strategy 1** and **Strategy 2** above, training a prediction model for each host.

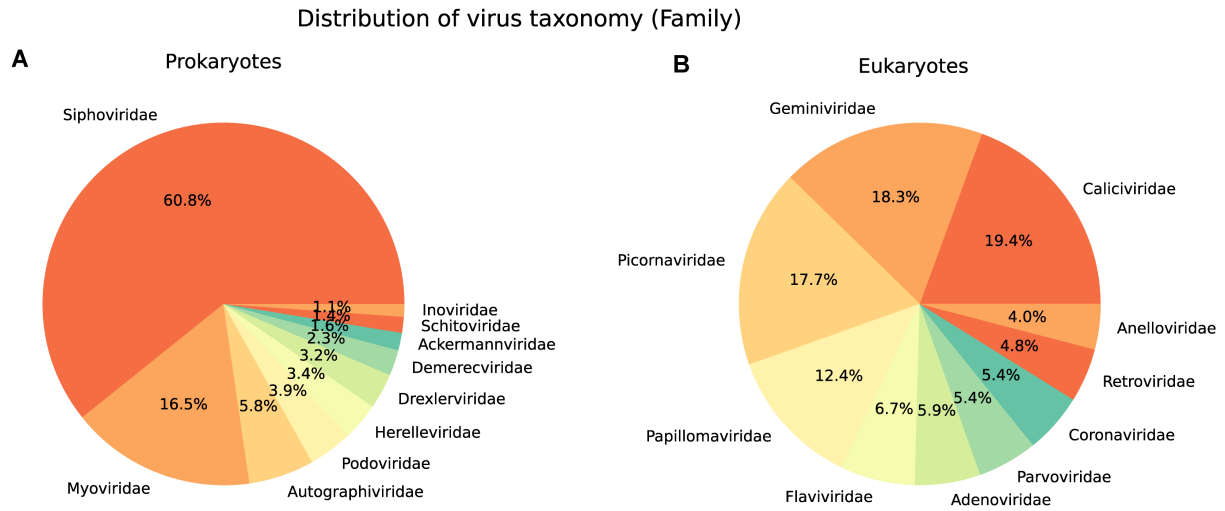

Figure 2.2. Virus taxonomy (family) distribution on prokaryotic (A) and eukaryotic (B) hosts. To illustrate the virus taxonomy distribution, we plot pie charts to show the distribution of virus families in prokaryotic and eukaryotic hosts. Viruses associated with prokaryotes are dominated by the Siphoviridae family, which constitutes approximately 60% (A), whereas the Geminiviridae, Picornaviridae and Papillomaviridae families are the top three ranked families, each accounting for roughly 18% (B).

2.4.2.1 Prokaryotic and Eukaryotic host performance

The heatmaps of evaluation indices for the prokaryotic and eukaryotic host classifiers are presented in Figure 2.3 A and C. Here, evaluation indices are calculated based on the best-performing model with the highest AUC chosen in 5-fold cross-validation. In Figure 2.3 A, the accuracy is higher than 0.9 except for two hosts, which are 0.86. The ROC curves in Figure 2.3 B show that all prokaryotic classifiers perform very strongly with each host achieving an AUC greater than 0.95 and 8 achieving an AUC of 1.

We also obtained the mean and standard deviation of each host, by testing 5-fold cross-validation models on the host test dataset (see Table 2.3). EvoMIL shows good performance, with 14/15 hosts achieving a mean AUC greater than 0.9. Overall, our results demonstrated that EvoMIL shows an impressive performance in the binary classification tasks of viruses associated with prokaryotic hosts. More evaluation metrics are included in Table 2.3.

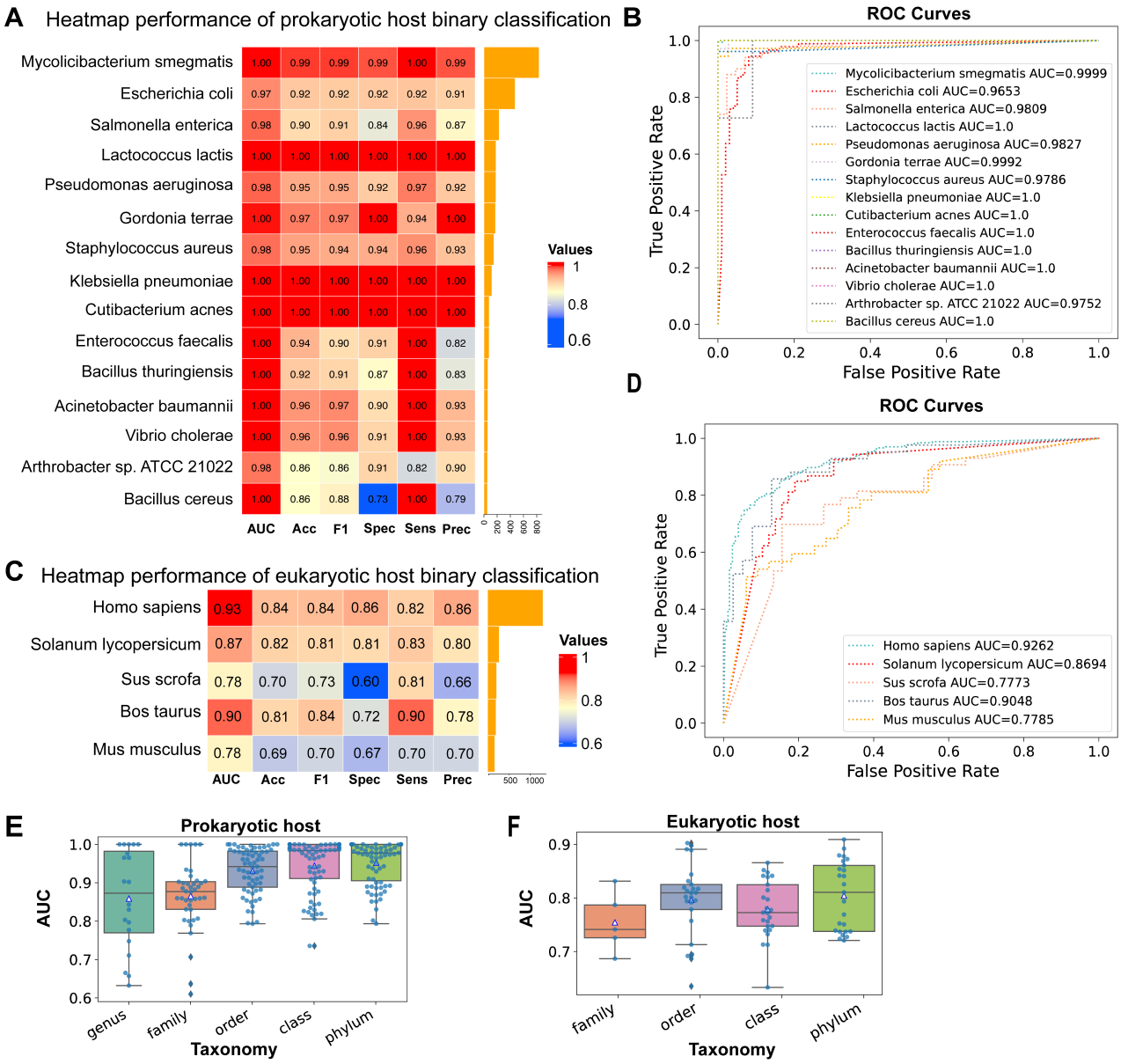

Figure 2.3. Performance of binary classification tasks. This figure separately shows the heatmap of AUC, accuracy, F1 score, sensitivity, specificity, and precision on 15 prokaryotic (A) and 5 eukaryotic host binary classifiers (C), negative samples are selected by **strategy 1**; ROC curves of 15 prokaryotic hosts (B) and 5 eukaryotic hosts (D) corresponding with heatmap plots A and C; AUC values of different taxonomies on prokaryotic (E) and eukaryotic hosts (F) where negative samples are selected using **strategy 2**.

Table 2.3. Results for prokaryotic hosts

Host name	AUC	Accuracy	F1	specificity	sensitivity	precision
Mycolicibacterium smegmatis	0.999 ± 0.0	0.993 ± 0.0	0.993 ± 0.0	0.998 ± 0.0	0.99 ± 0.0	0.993 ± 0.0
Escherichia coli	0.947 ± 0.01	0.893 ± 0.02	0.892 ± 0.02	0.868 ± 0.07	0.915 ± 0.03	0.896 ± 0.02
Salmonella enterica	0.965 ± 0.03	0.899 ± 0.02	0.897 ± 0.02	0.94 ± 0.02	0.851 ± 0.06	0.904 ± 0.01
Lactococcus lactis	0.999 ± 0.0	0.995 ± 0.01	0.995 ± 0.01	1.0 ± 0.0	0.988 ± 0.02	0.995 ± 0.01
Pseudomonas aeruginosa	0.978 ± 0.01	0.953 ± 0.01	0.953 ± 0.01	0.956 ± 0.02	0.951 ± 0.02	0.954 ± 0.01
Gordonia terrae	0.998 ± 0.0	0.975 ± 0.01	0.975 ± 0.01	0.966 ± 0.01	0.983 ± 0.02	0.975 ± 0.01
Staphylococcus aureus	0.974 ± 0.01	0.964 ± 0.01	0.963 ± 0.01	0.962 ± 0.0	0.966 ± 0.02	0.963 ± 0.02
Klebsiella pneumoniae	0.986 ± 0.02	0.963 ± 0.03	0.963 ± 0.03	0.978 ± 0.05	0.945 ± 0.04	0.966 ± 0.03
Cutibacterium acnes	0.68 ± 0.44	0.842 ± 0.22	0.751 ± 0.34	0.6 ± 0.55	1.0 ± 0.0	0.721 ± 0.38
Enterococcus faecalis	0.995 ± 0.0	0.944 ± 0.01	0.934 ± 0.02	1.0 ± 0.0	0.922 ± 0.02	0.917 ± 0.02
Bacillus thuringiensis	0.992 ± 0.01	0.928 ± 0.02	0.927 ± 0.02	1.0 ± 0.0	0.88 ± 0.03	0.924 ± 0.02
Acinetobacter baumannii	0.984 ± 0.02	0.933 ± 0.02	0.929 ± 0.02	1.0 ± 0.0	0.84 ± 0.05	0.949 ± 0.02
Vibrio cholerae	0.957 ± 0.03	0.925 ± 0.03	0.924 ± 0.04	0.969 ± 0.04	0.873 ± 0.05	0.931 ± 0.04
Arthrobacter sp. ATCC 21022	0.953 ± 0.02	0.873 ± 0.05	0.873 ± 0.05	0.873 ± 0.08	0.873 ± 0.05	0.875 ± 0.05
Bacillus cereus	0.977 ± 0.03	0.864 ± 0.03	0.862 ± 0.03	0.982 ± 0.04	0.745 ± 0.04	0.886 ± 0.03

Results of binary classifiers in prokaryotic hosts based on negative sampling strategy 1. Evaluation indices are obtained by testing 5-fold cross-validation models on each host, and then the mean and standard deviation of each evaluation metric can be obtained. Evaluation metrics include AUC, accuracy, F1, specificity, sensitivity, and precision.

The accuracy of each eukaryotic host classifier is shown in Figure 2.3 C, it is clear that all hosts perform with an accuracy higher than 0.8 except for two hosts which are roughly 0.7. *H. sapiens* obtained the highest accuracy with 0.84; *Mus musculus* has the lowest accuracy with 0.69. The ROC curves of the 5 eukaryotic host classifiers are presented in Figure 2.3 D. Although the eukaryotic classifiers achieve good performance with AUCs above 0.77, they perform less well than the prokaryotic classifiers, with only 3/5 datasets scoring an AUC above 0.85. There may be several explanations for the lower performance. Firstly, the average number of proteins per virus is much lower, resulting in small bags for MIL. Secondly, there is a much higher diversity of virus types in the datasets of the eukaryotic hosts, often containing viruses from multiple Baltimore classes. The virus of these different classes is polyphyletic meaning they will have no common ancestor and therefore have no shared genes and interact with different host pathways.

The mean and standard deviation of each host are obtained by testing five trained cross-validation models on the test data set (see Table 2.4). Here, the mean AUC is higher than 0.85 except for the classifiers of two hosts that perform less well, with AUC scores of 0.761 ± 0.01 for *Sus scrofa* and 0.762 ± 0.02 for *M. musculus*. Overall, our results demonstrate that EvoMIL performs well in binary classification tasks of viruses associated with eukaryotic hosts.

Table 2.4. Results for eukaryotic hosts

Host name	AUC	Accuracy	F1	specificity	sensitivity	precision
Homo sapiens	0.918 ± 0.01	0.84 ± 0.01	0.84 ± 0.01	0.856 ± 0.03	0.823 ± 0.04	0.842 ± 0.01
Solanum lycopersicum	0.851 ± 0.01	0.8 ± 0.03	0.8 ± 0.03	0.808 ± 0.06	0.793 ± 0.03	0.801 ± 0.03
Sus scrofa	0.761 ± 0.01	0.727 ± 0.02	0.726 ± 0.02	0.791 ± 0.04	0.667 ± 0.04	0.733 ± 0.02
Bos taurus	0.898 ± 0.0	0.837 ± 0.02	0.835 ± 0.02	0.895 ± 0.03	0.774 ± 0.07	0.844 ± 0.01
Mus musculus	0.762 ± 0.02	0.697 ± 0.01	0.696 ± 0.01	0.703 ± 0.02	0.691 ± 0.03	0.697 ± 0.01

Results of binary classifiers in eukaryotic hosts based on negative sampling strategy 1. Evaluation indices are obtained by testing 5-fold cross-validation models on each host, and then the mean and standard deviation of each evaluation metric can be obtained. Evaluation metrics include AUC, accuracy, F1, specificity, sensitivity, and precision.

2.4.2.2 Sampling negative samples from similar viruses makes binary host classification more challenging

Next, we test our model with more challenging tasks. Using the second strategy of selecting negative viruses that are associated with hosts sharing the same taxonomic rankings as the hosts associated with positive viruses, we observe that the classification task becomes increasingly challenging as we move from the phylum to the genus level. Results show that our EvoMIL models achieve high AUC scores but that distinguishing between viruses of similar hosts is more difficult with a noticeable drop in performance at family and genus levels. In Figure 2.3, the box plots show AUC values of prokaryotic (E) and eukaryotic (F) hosts based on negative selection **strategy 2** with five taxonomies genus/family/order/class/phylum. Phylum level (lime colour) presented significant improvement compared with lower taxonomies, especially the genus level. Note, at the lower taxonomic ranks there are only sufficient numbers of negative viruses to meet our threshold of 50 for 4 hosts at the genus level, 8 hosts at the family level and 13 hosts at the order level.

To quantify the difficulty of the task, we computed the sequence similarity scores between all pairs of positive and negative sets on **strategy 2** using MMSeq2 (Steinegger et al. 2017) (see Figure 2.4 and Figure 2.5). Most of the scores are above 0.6. Looking at the scores across taxonomy levels, the phylum (purple) level tends to have lower similarities, while the family (orange) and order (green) levels tend to exhibit higher identity scores. These suggest a good degree of sequence similarity between positive and negative viruses, therefore a challenging classification task.

2.4.3 Embedding features outperform protein and DNA k-mer features on multi-class classification tasks

To demonstrate that the ESM embeddings are encoding more predictive information than conventional features, we generated feature sets from the k-mer composition of the nucleic acid, amino-acid and physico-chemical sequences (Young et al. 2020), and evaluated ESM-1b and k-mer features with a multi-class classification task. Here, we performed multi-class classification extending attention-based MIL by modelling the joint multinomial distribution of bag labels. Both prokaryotic and eukaryotic multi-class datasets were constructed using hosts infected by at least 30 viruses. This resulted in 22 classes for the prokaryotes and 36 classes for the eukaryotes. Again, we applied 5-fold cross-validation on training datasets, then separately tested trained models on the testing dataset. These models are named ESM-1b, AA_2, PC_3 and DNA_5, according to different features.

The results for each model are presented in Table 2.5, we obtain AUC, accuracy, and F1 scores by evaluating each of the 5-fold cross-validation models with the test dataset. Comparing ESM-1b with AA_2, PC_3, and DNA_5 for both prokaryotic and eukaryotic hosts, the ESM-1b features have the strongest prediction signal since during testing on prokaryotic hosts, the mean F1 score of EvoMIL is 0.88, achieving improvements of 10.8%, 16.2%, and 4.9% compared with AA_2, PC_3 and DNA_5, respectively; while the counterpart of EvoMIL in eukaryotic hosts is 0.292, achieving improvements of 1.7%, 6.6% and 11.5% compared with AA_2, PC_3 and DNA_5 (Table 2.5). Additionally, ESM-1b demonstrates better performance in terms of AUC and accuracy compared to k-mers, except for eukaryotes, the accuracy of ESM-1b is comparable to that of AA_2.

Table 2.5. The AUC, Accuracy and F1 score of multi-class MIL by using ESM-1b and k-mer features. Here is a comparison of AUC, accuracy and F1 score between ESM-1b and k-mer feature sets (AA_2, PC_3 and DNA_5). For each feature, the evaluation was conducted by training multi-class classification models on 22 prokaryotic hosts and 36 eukaryotic hosts, and the mean and standard deviation of the evaluation metrics (AUC, accuracy, and F1 score) were obtained using 5-fold cross-validation. The accuracy of a random classifier is 0.045 for 22 classes and 0.028 for 36 classes.

Host type	Methods	AUC	Accuracy	F1 score
Prokaryotes	ESM-1b	**0.992±0.0**	**0.909± 0.0**	**0.88±0.01**
	AA_2	0.979 ± 0.0	0.856 ± 0.01	0.794 ± 0.02
	PC_3	0.969 ± 0.01	0.843 ± 0.01	0.757 ± 0.02
	DNA_5	0.987 ± 0.0	0.882 ± 0.01	0.839 ± 0.01
Eukaryotes	ESM-1b	**0.831±0.01**	**0.494 ± 0.01**	**0.292 ± 0.01**
	AA_2	**0.829±0.03**	**0.494 ± 0.01**	0.287 ± 0.01
	PC_3	0.821 ± 0.01	0.479 ± 0.01	0.274 ± 0.02
	DNA_5	0.801 ± 0.02	0.466 ± 0.02	0.262 ± 0.01

Figure 2.6 A and B, respectively, show AUC and accuracy across ESM-1b and k-mer features in both prokaryotes and eukaryotes, and AUC and accuracy are equivalent with those presented in Table 2.5. In Figure 2.6 A, ESM-1b has the highest AUC and smallest standard deviation among prokaryotic hosts; for eukaryotic hosts, ESM-1b shows the smallest standard deviation and the highest mean AUC compared with k-mer features. Here, AA_2 has the largest variation, despite obtaining the highest AUC. In Figure 2.6 B, ESM-1b presents the highest accuracy and the smallest standard deviation among prokaryotic hosts; for eukaryotic hosts, ESM-1b includes the highest accuracy. Additionally, although AA_2 presents the smallest standard deviation in accuracy, ESM-1b obtained the best mean AUC and F1 score (see Table 2.5).

To further evaluate the MIL models, we test the models using viruses of hosts that are not in the multi-class models trained above, only selecting hosts with between 5 and 30 associated viruses in the VHDB. These viruses were used to test all 5-fold models generated above, comparing all features. Figure 2.6 C and D show a comparison of the accuracy of ESM-1b and k-mer features for all cross-validation MIL multi-class models. To calculate accuracy, we determined

the percentage of correct predictions on all test samples. Here, a correct prediction is defined as one in which the predicted host rank is the same as that of the true host label for the taxonomic ranks of phylum, class, order, family, and genus.

In prokaryotic hosts (Figure 2.6 C), accuracy is roughly between 1% and 10% at genus, family and order levels, while accuracy is between 20% and 90% at class, phylum, and kingdom levels. These results indicate that prediction is more challenging at lower taxonomic ranks. ESM-1b performs best at the genus level with the highest mean accuracy, 1.8%, while the mean accuracy for AA_2, PC_3, DNA_5 are 1.09%, 1.18% and 1.23%, respectively (see Table 2.6). Furthermore, across each taxonomy level, the standard deviation of ESM-1b is the smallest compared with k-mer features, although ESM-1b does not perform with the highest accuracy for higher taxonomic ranks. Overall, ESM-1b performs best at the genus level and shows stable accuracy results across all taxonomy levels on prokaryotic hosts.

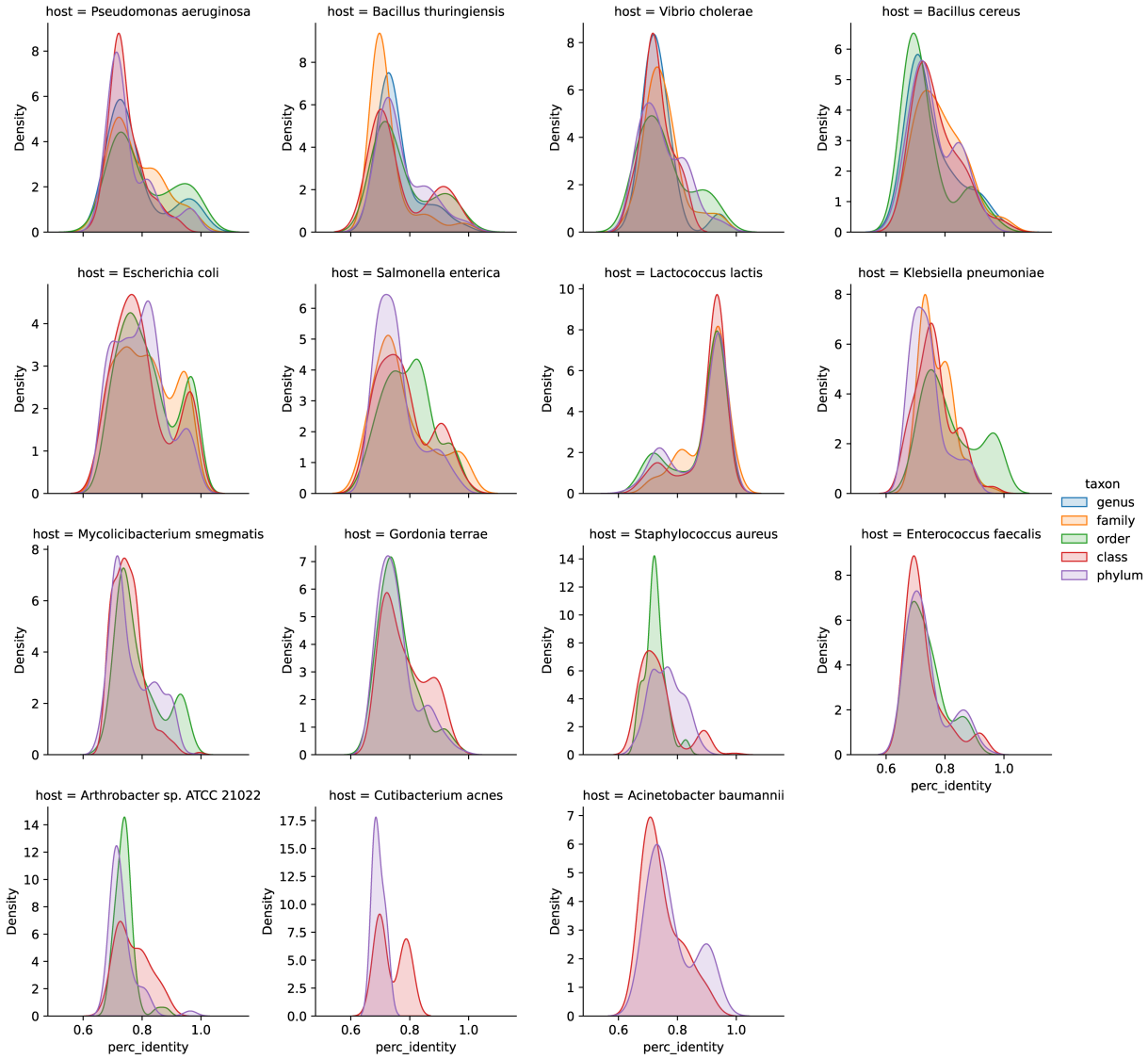

Figure 2.4. The genome sequence similarities between positive and negative viruses from different taxonomies on each prokaryotic host. This figure presents the distribution of genome similarities between positive and negative samples on each prokaryotic host, where the negative viruses are chosen based on the same taxonomy (genus, family, order, class, phylum) as the positive viruses.

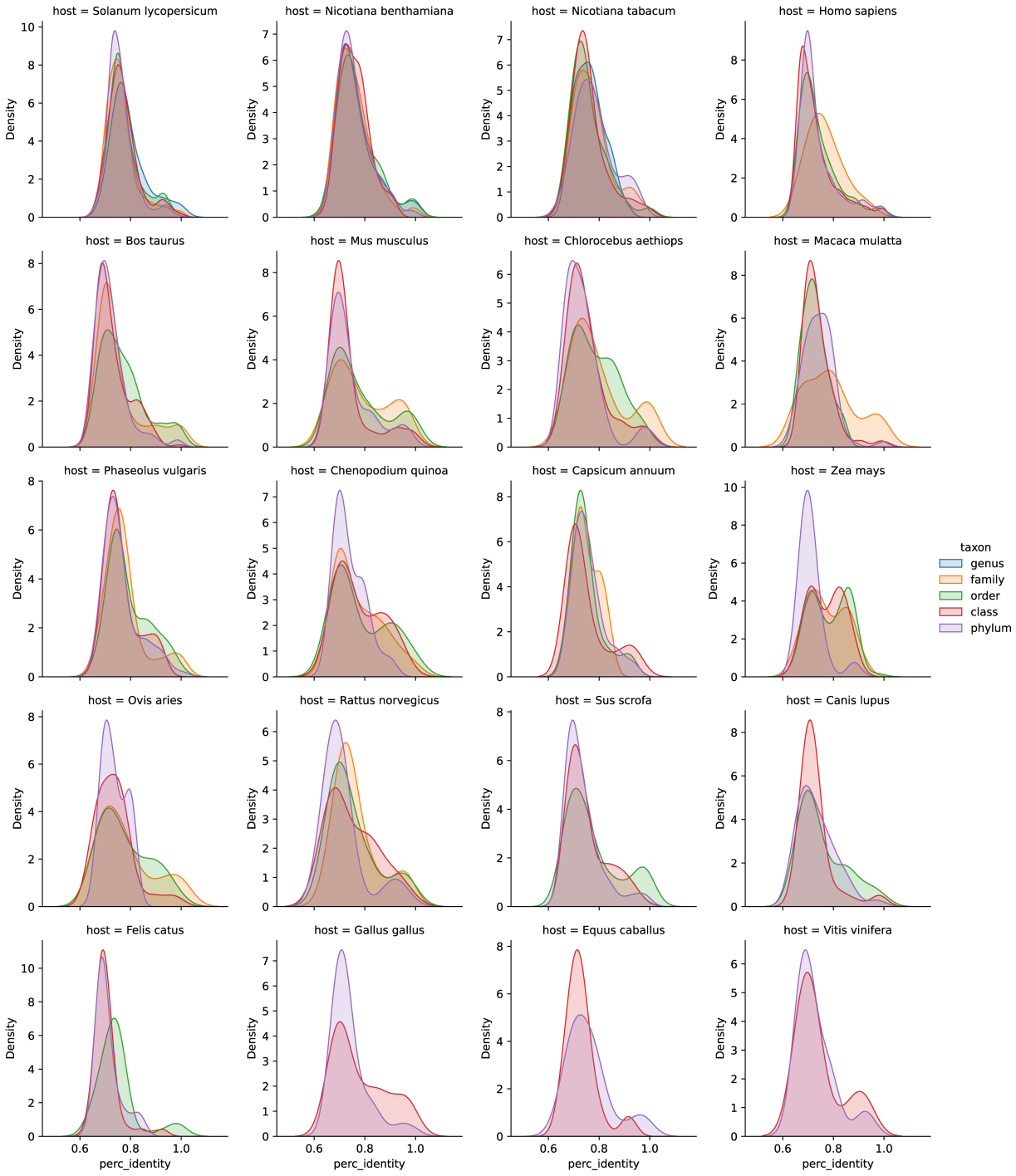

Figure 2.5. The genome sequence similarities between positive and negative viruses from different taxonomies on each eukaryotic host. This figure presents the distribution of genome similarities between positive and negative samples on each eukaryotic host, where the negative viruses are chosen based on the same taxonomy (genus, family, order, class, phylum) as the positive viruses.

Table 2.6. The accuracy (%) of multi-class MIL by using ESM-1b and k-mer features on tiny host datasets

Host type	Features	Genus	Family	Order	Class	Phylum	Kingdom
Prokaryotes	ESM-1b	**1.8±0.332**	4.386±**0.446**	9.412±**0.61**	22.928±**1.61**	31.872±**0.939**	97.41±0.0
	AA_2	1.088±0.529	5.216±0.894	12.22±2.223	29.334±2.671	39.122±3.93	97.41±0.0
	PC_3	1.178±0.258	4.556±1.063	11.12±2.445	25.646±2.7	32.868±4.513	97.41±0.0
	DNA_5	1.23±0.326	4.028±0.479	9.542±1.067	21.308±1.665	25.644±2.253	97.41±0.0
Eukaryotes	ESM-1b	**0.908±0.143**	**3.842±0.628**	**5.498±0.589**	**36.868±1.54**	**39.992±1.446**	**47.416±1.41**
	AA_2	0.328 ± 0.142	2.086±0.225	4.8±0.31	24.006±1.761	29.844±1.248	43.222±0.555
	PC_3	0.334±0.063	2.19±0.174	4.826±0.602	27.374±1.492	32.374±1.372	44.344±0.915
	DNA_5	0.42±0.178	2.102±0.319	4.366±0.644	25.144±3.519	31.792±2.37	43.908±1.225

Comparison of mean accuracy and standard deviation between ESM-1b and k-mer feature sets: ESM-1b, AA_2, PC_3 and DNA_5. For each feature, training multi-class classification models on 22 prokaryotic hosts and 36 eukaryotic hosts based on 5-fold cross-validation, then the mean and standard deviation of accuracy are obtained by testing the trained model on host datasets that are associated with fewer viruses, ranging from 5 to 30 (tiny host datasets).

As for eukaryotic hosts (Figure 2.6 D), accuracy is roughly between 1% and 6% at genus, family and order levels, while accuracy is between 25% and 50% at class, phylum, and kingdom levels. Across all taxonomic ranks, ESM-1b consistently outperforms DNA and protein k-mer features in terms of accuracy. For example, the mean accuracy of ESM-1b at the class level is 36.87%, which is higher by 12.86%, 9.49%, and 11.72% than AA_2, PC_3 and DNA_5, respectively (see Table 2.6). In summary, the multi-class MIL model based on ESM-1b features shows potential in predicting hosts that are associated with fewer than 30 viruses in eukaryotic datasets.

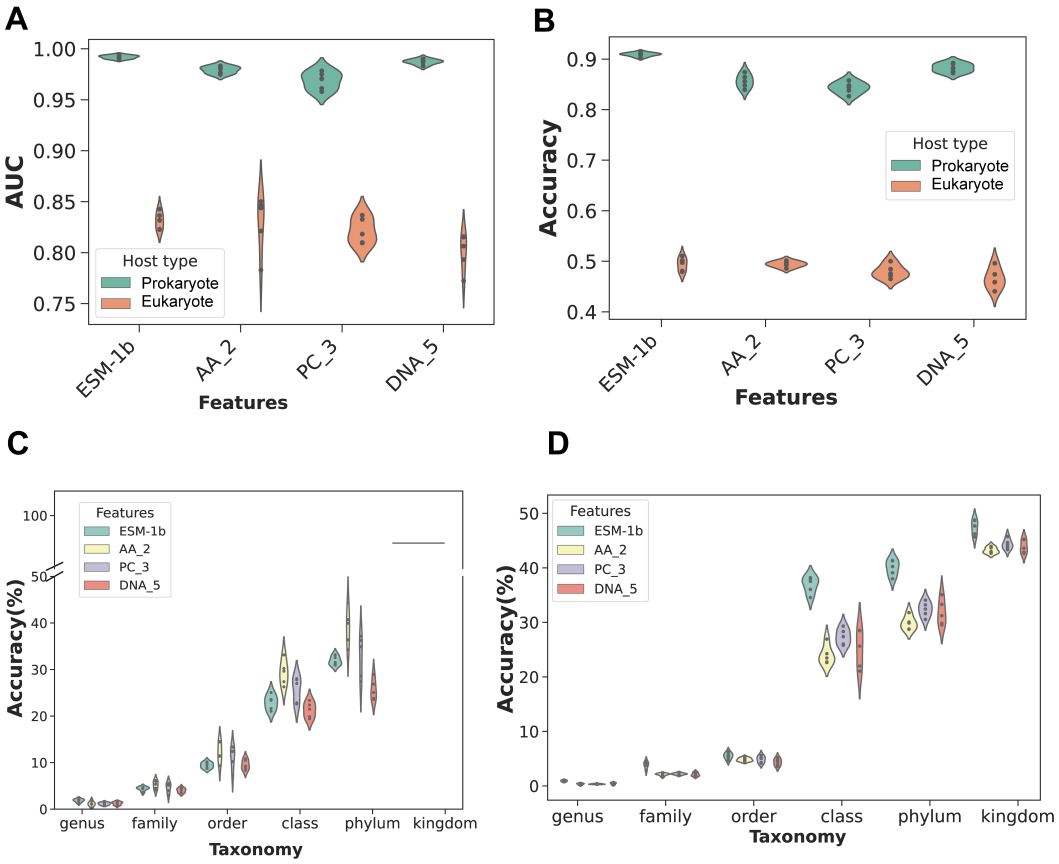

Figure 2.6. Performance of multi-class classifications on ESM-1b and k-mer features. A and B represent the AUC and accuracy, respectively, for prokaryotic and eukaryotic hosts using four feature sets (ESM-1b, AA_2, PC_3 and DNA_5), AUC and accuracy are equivalent to those presented in Table 2.5. C and D indicate the results of testing the trained models on prokaryotic and eukaryotic hosts associated with 5 to 30 viruses, using the four feature sets described above.

Overall, ESM-1b demonstrated superior performance to the k-mer features, through 5-fold cross-validation on both prokaryotic and eukaryotic hosts. Furthermore, we compare the accuracy of each host to evaluate the prediction performance of ESM-1b and k-mer features on multi-class classification tasks.

Figure 2.7 A and 2.7 B, respectively, presented the taxonomic tree of prokaryotic and eukaryotic and the mean Log2 accuracy ratio between ESM-1b and K-mers (AA_2, PC_3 and DNA_5) and standard deviation of 5-fold cross-validation for each host. Results show that ESM-1b achieved the highest mean accuracy in 17 out of 22 prokaryotic hosts compared with protein and DNA

k-mer features. These findings indicate that ESM-1b outperforms k-mer feature sets in multi-classification tasks for prokaryotic hosts. For eukaryotic hosts, the performance of ESM-1b is comparable to k-mer features in each eukaryotic host except *H. sapiens*. Notably, ESM-1b demonstrated a significant improvement over k-mer in *H. sapiens*, which has the largest number of training samples, indicating that EvoMIL performs better with a larger number of virus training samples for eukaryotic hosts.

To understand if the host phylogeny explains the prediction performance of multi-class classification, we visualise the taxonomic tree of hosts next to the heatmap showing the number of predicted false hosts for each host (Figure 2.9). The objective is to determine if false predicted host classes of viruses are more likely to share common parent nodes with true hosts in the taxonomic tree. In Figure 2.9, two heatmaps respectively present the number of false predicted hosts which belong to the same taxonomy as the true host in prokaryotic and eukaryotic hosts.

In the prokaryotic host heatmap (Figure 2.9 A), the highest number of falsely predicted hosts is observed at the family level. Notably, *E. coli* accounts for the largest number of false predicted hosts. Furthermore, an interesting observation is that *Bacillus cereus* and *Bacillus thuringiensis* host models predict each other as the host label (see Figure 2.10 A), and they belong to the same genus, indicating that predicted hosts have a tendency to be hosts sharing the same taxonomy levels with the true host label. In the eukaryotic host heatmap (Figure 2.9 B), the largest number of false predicted hosts belong to the class level. *H. sapiens* and *Bos taurus* have the highest number of false predicted hosts, and they predict each other as the host label (see Figure 2.10 B). Similarly, a similar situation can be observed between *S. scrofa* and *M. musculus*, which further reinforces the finding that closely related hosts within the taxonomic tree are more likely to be predicted as each other's labels.

To understand the relationship between host phylogeny and predicted results, we use *E. coli* as an example in prokaryotes. In Figure 2.10 A, based on ESM-1b, the numbers of false host labels predicted are: 7% for *S. enterica*, 3% labels for *Erwinia amylovora*, 2% labels for *K. pneumoniae*, and 1% label for *Ralstonia solanacearum*, where *S. enterica* and *K. pneumoniae* belong to the same family as *E. coli*, *E. amylovora* belong to the same order, and *R. solanacearum* belongs to the same phylum as *E. coli*. In Figure 2.7 B, it is clear that it is more challenging to predict eukaryotic hosts compared with prokaryotic hosts. ESM-1b demonstrates significantly superior performance compared to k-mer features in *Nicotiana clevelandii* and *Cucumis sativus*, while its performance on other hosts within the eukaryotic taxonomic tree is comparable to that of the k-mer features. In the case of *H. sapiens* as an example, the false positive host labels based on ESM-1b for *H. sapiens* primarily consist of 4% *Pan troglodytes* labels, 3% *Chlorocebus aethiops* labels, 3% *Macaca mulatta* labels, and 3% *S. scrofa* labels. Here, *Pan troglodytes* belongs to the same family as *H. sapiens*; *Chlorocebus aethiops* and *Macaca mulatta* belong to the same order

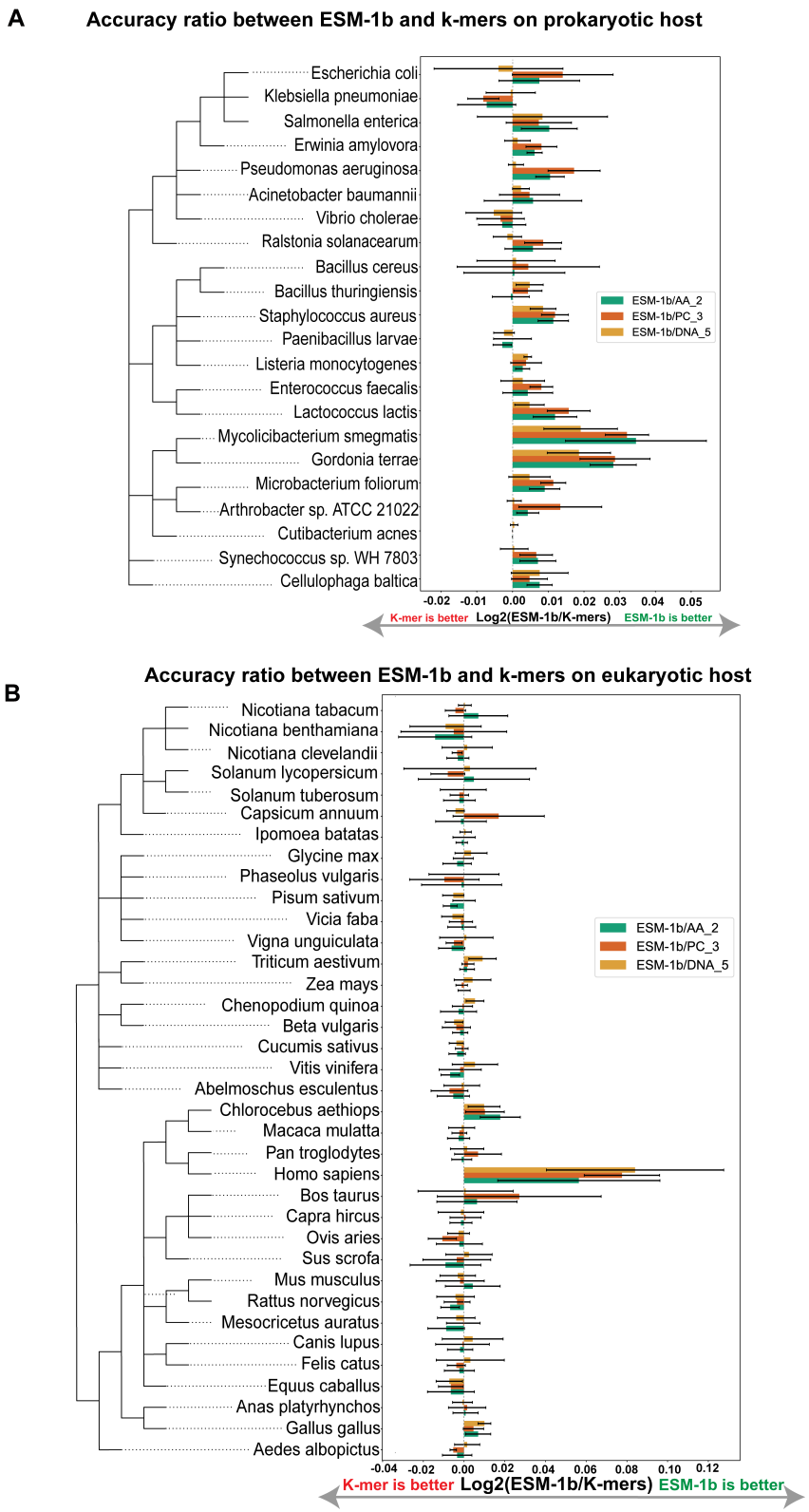

Figure 2.7. The taxonomic tree, aligning with Log2 of ratio accuracy between ESM-1b and k-mers. The figure shows the taxonomic tree of 22 prokaryotic (A) and 36 eukaryotic (B) hosts. Each host is aligned with a bar plot showing the accuracy ratio and standard deviation of 5-fold cross-validation between ESM-1b and AA_2, PC_3, and DNA_5, respectively. The taxonomic tree aligning with the accuracy between ESM-1b and k-mers is shown in Figure 2.8.

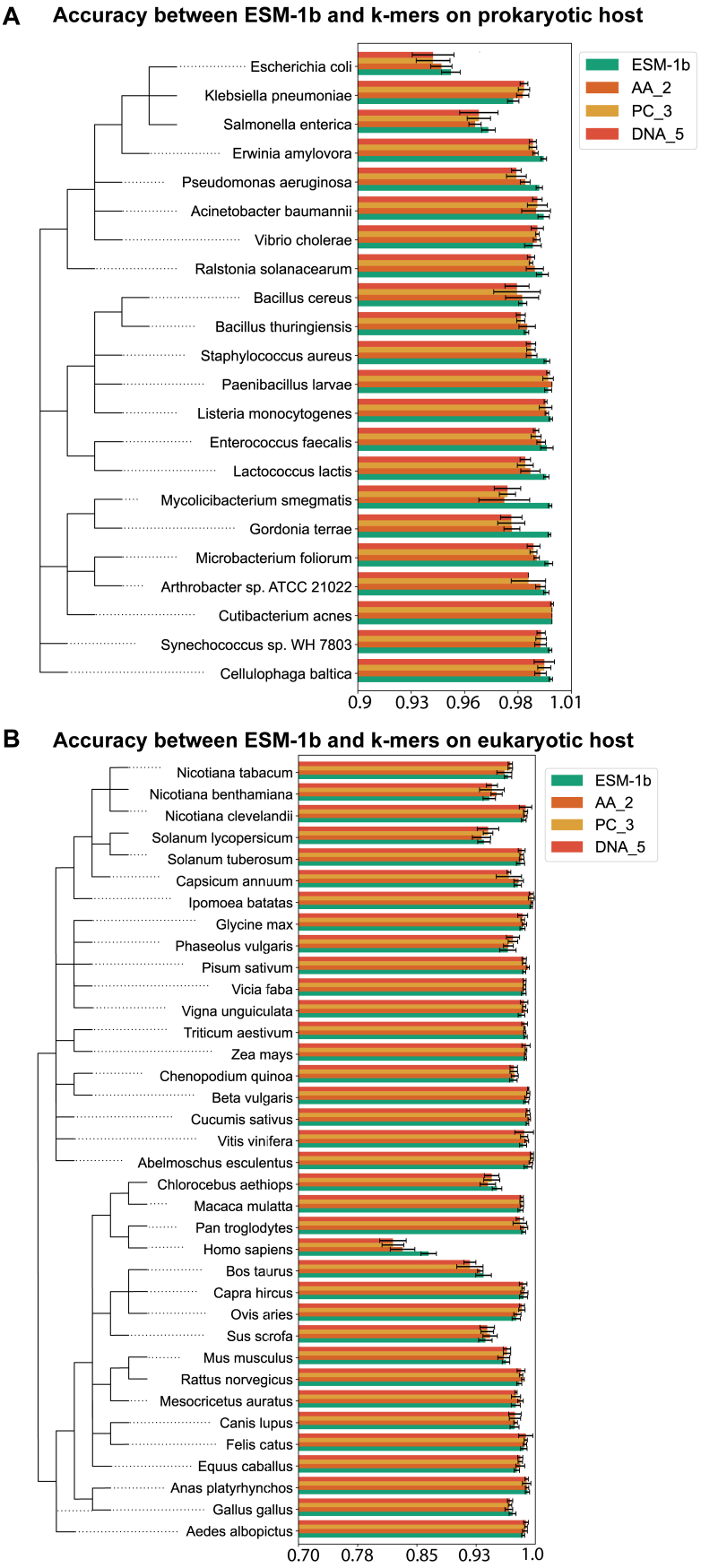

Figure 2.8. The taxonomic tree, aligning with accuracy values between ESM-1b and k-mers. The figure shows the taxonomic tree of 22 prokaryotic hosts (A) and 36 eukaryotic (B) hosts. Each host is aligned with a bar plot showing the accuracy and standard deviation of 5-fold cross-validation between ESM-1b and AA_2, PC_3, and DNA_5, respectively.

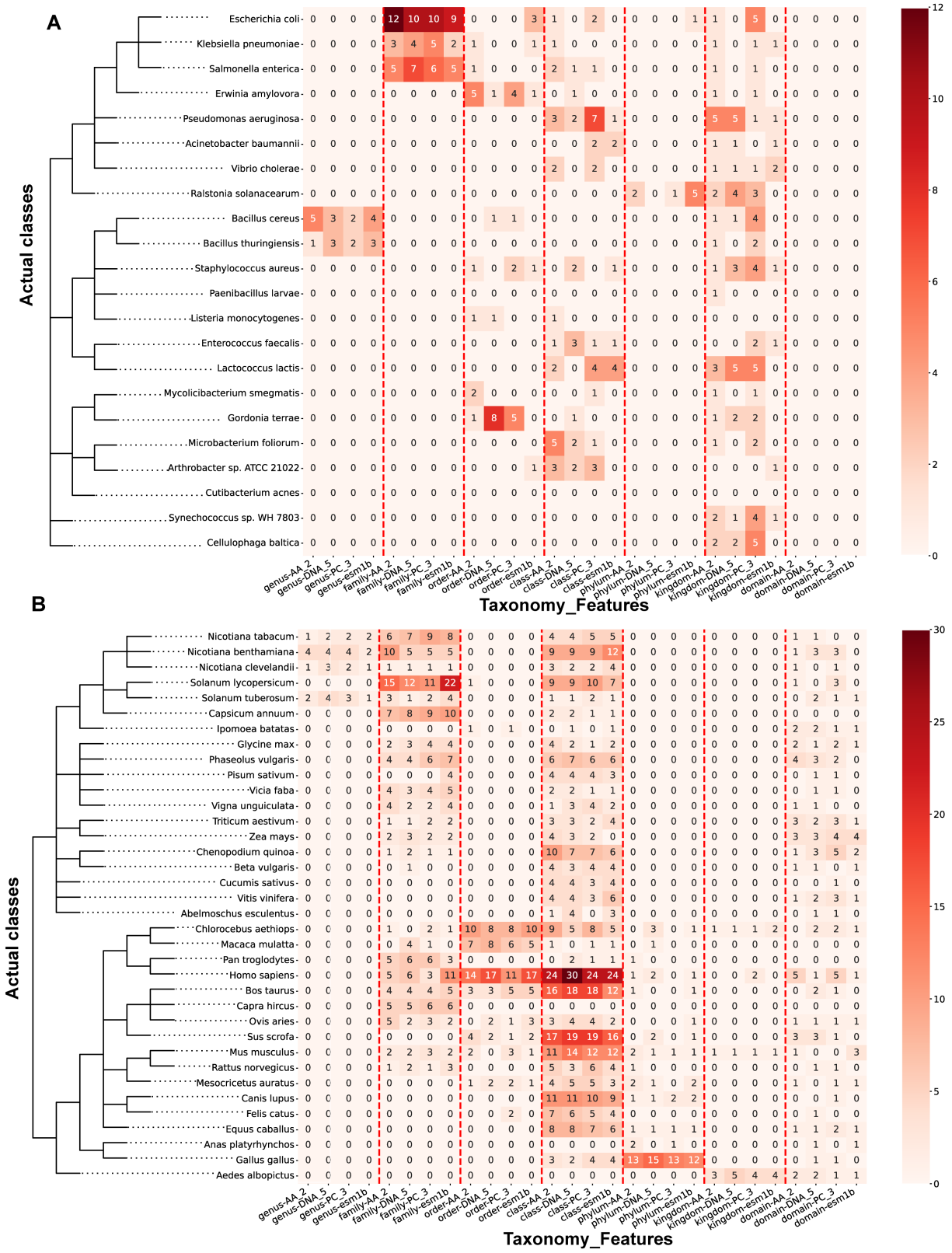

Figure 2.9. The heatmap of different features for each prokaryotic (A) and eukaryotic (B) host. The values in the heatmap are the total number of predicted hosts which belong to the same taxonomy as the true host.

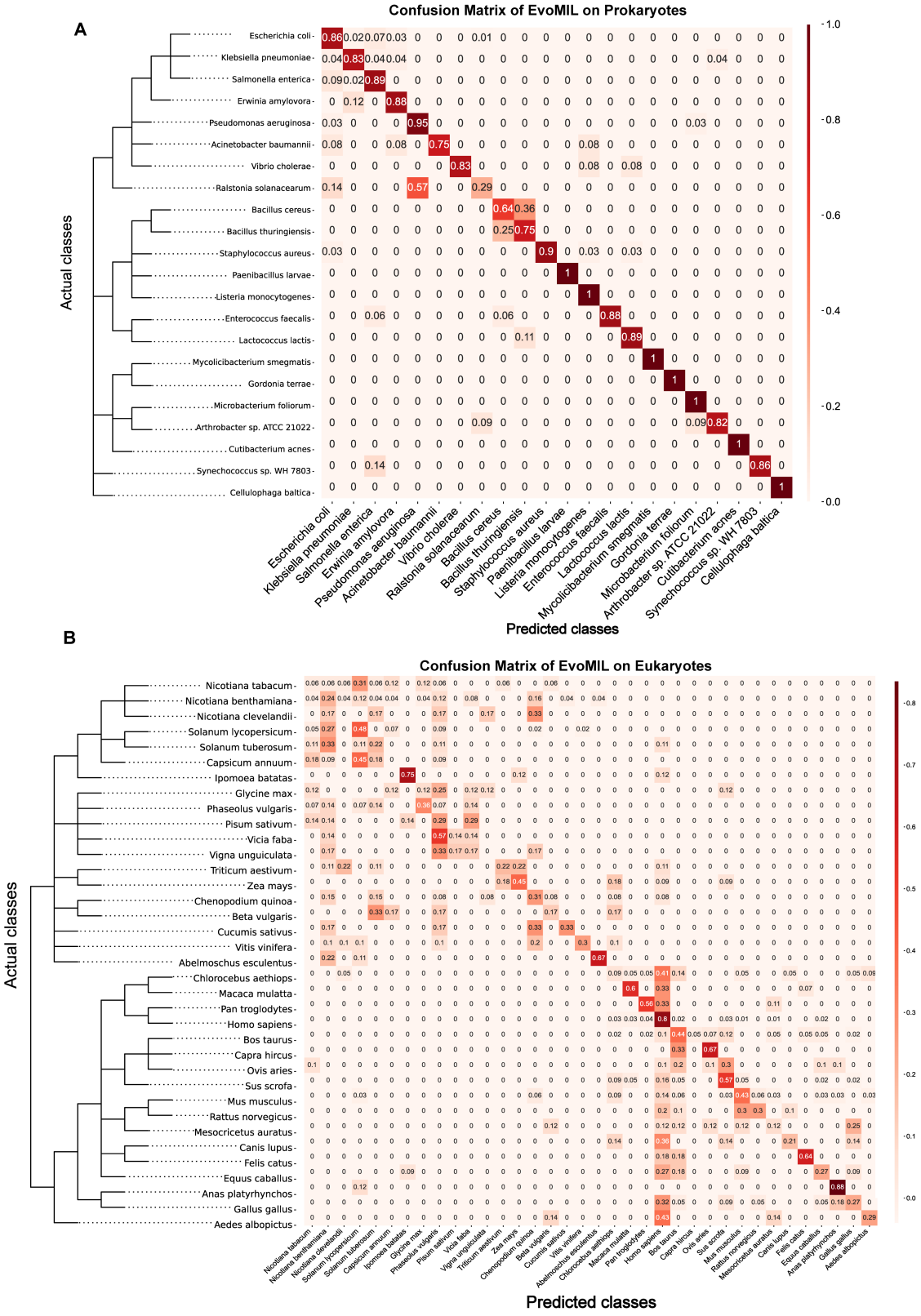

Figure 2.10. The Confusion matrix plot of prokaryotic hosts (A) and eukaryotic hosts (B) based on EvoMIL. The confusion matrix plots A and B represent the performance of the EvoMIL model on 22 prokaryotic hosts and 36 eukaryotic hosts, respectively. It is constructed by evaluating the model's predictions on a test set comprising 20% of the dataset, while the EvoMIL model was trained on the remaining 80% of the data. This plot provides insights into the model's accuracy in predicting the host species for the tested viruses.

level as *H. sapiens*; and *S. scrofa* belongs to the same class level as *H. sapiens*.

Overall, viruses are more likely to be misclassified as closely related hosts during multi-class classification, confirming that those hosts sharing common parent nodes in the taxonomic tree tend to be infected by similar viruses (Young et al. 2020).

2.4.4 Benchmarking EvoMIL with prokaryotic host predictors

Here we compare EvoMIL's prediction performance with 9 prokaryotic host predictors including the state-of-the-art iPHoP (Roux et al. 2022), BLASTn, BLASTp, CRISPR, WisH (Galiez et al. 2017), VirHostMatcher (Ahlgren et al. 2017), PHP (Lu et al. 2021), SpacePHARER (Zhang et al. 2021), VirHostMatcher-Net (Wang et al. 2020) and vHULK (Amgarten et al. 2020). We chose 364 viruses across the 22 prokaryotic species that matched the host labels in our training set, which were chosen from the benchmarking test dataset from the iPHoP paper. The full list of the testing viruses can be found in Table 2.8. For details about prediction settings for each approach, please see the Materials and Methods section.

BLASTn, BLASTp, CRISPR, PHP, WisH, VirHostMatcher and iPHoP make multiple host predictions for a given virus. We considered any predicted host species in their prediction list that matched the true host species as a correct prediction. EvoMIL, SpacePharer, VHM-Net, and vHULK are evaluated based on their top-1 prediction accuracies. As shown in Figure 2.11, EvoMIL achieves the highest accuracy of 75.27% on the test viruses, with 274 correct predictions out of 364 viruses and has a 33.65% improvement over the state-of-the-art approach, iPHoP.

2.4.5 MIL attention weights can be used to interpret which virus proteins are important

Next, we investigate if the proteins identified as important for the prediction are key contributors to virus-host specificity, i.e., whether the transformer embeddings are encoding biologically meaningful information about the viral proteins. As we described in the introduction, attention-based MIL learns the weight of each protein in a virus bag. A high weight indicates that a protein is important for host prediction, and by implication is more likely to be important to virus-host specificity. We looked at Gene Ontology (GO) annotations of these highly weighted proteins and their proximity in the embedding space with hierarchical clustering.

In Figure 2.12, we can see that the top 5 ranked proteins annotated by GO terms include im-

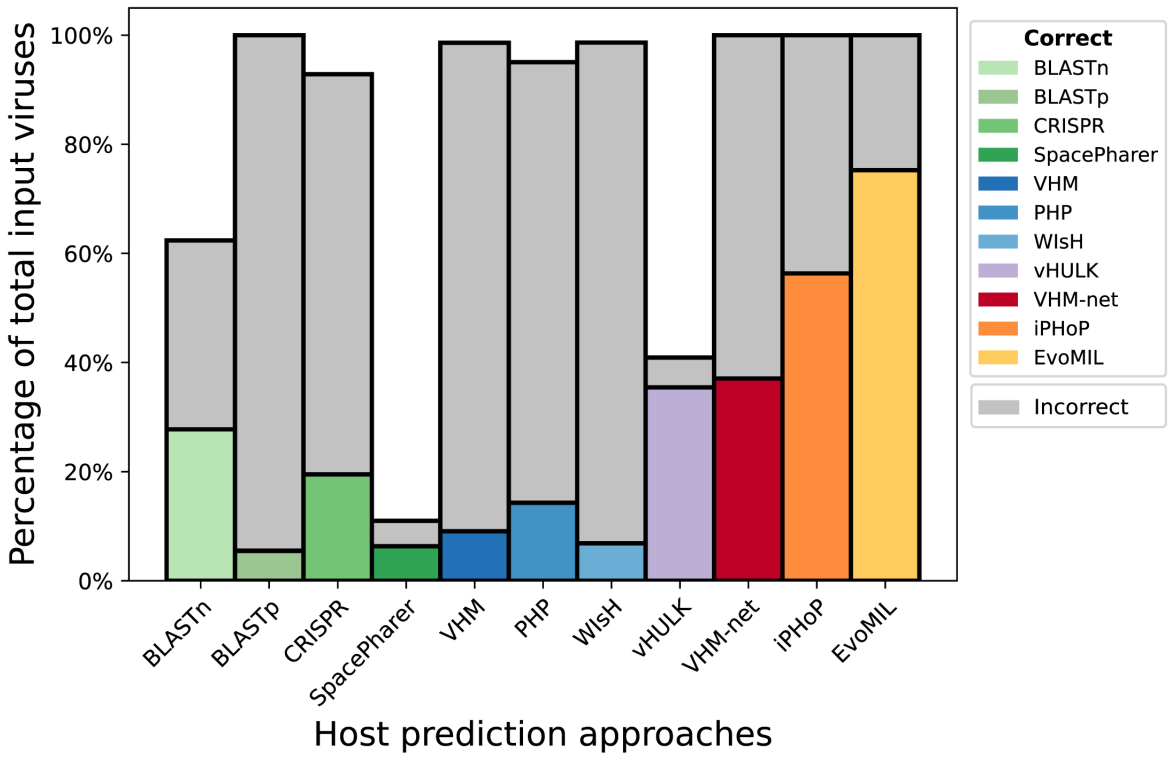

Figure 2.11. Comparison of EvoMIL and other host prediction approaches on an independent test dataset. The y-axis presents the number of correct predictions (coloured bar) and the number of incorrect predictions (grey bar) for each tool (x-axis) on the chosen benchmarking test dataset (Table 2.8). This plot shows the percentage of correct and incorrect host species predictions on the test dataset, and prediction source results are available in Table 2.9.

portant proteins with high weights, and proteins from a virus might contain weights of 0, so selecting the top 5 ranked proteins allows us to collect proteins with GO annotations and high weights, as the aim is to present key proteins which are assigned high weights based on binary and multi-class models. We mark the GO annotation for proteins and select three GO annotations as examples. For example, the top 5 ranked proteins (Figure 2.12 A) and all ranked viral proteins (Figure 2.12 B) associated with *E. coli* shared the same GO annotation (viral capsid assembly) and have the identical rank index within their respective ranked protein sets, meaning that selecting the top 5 ranked proteins of each virus still allows us to identify key GO annotations.

As shown in Figure 2.12, this selection allows us to focus on proteins that are assigned high weights, considering the possibility that some proteins from a given virus may have weights of 0. By choosing the top 5 ranked proteins, we are able to gather proteins with GO annotations that also possess high weights. This approach allows us to highlight key proteins assigned high weights based on binary and multi-class models. Figure 2.12 A and B demonstrate that some ranks of GO annotations for the top 5 ranked proteins aligned with those ranks of all proteins, indicating that GO annotations can still be obtained even when considering only the top 5 ranked

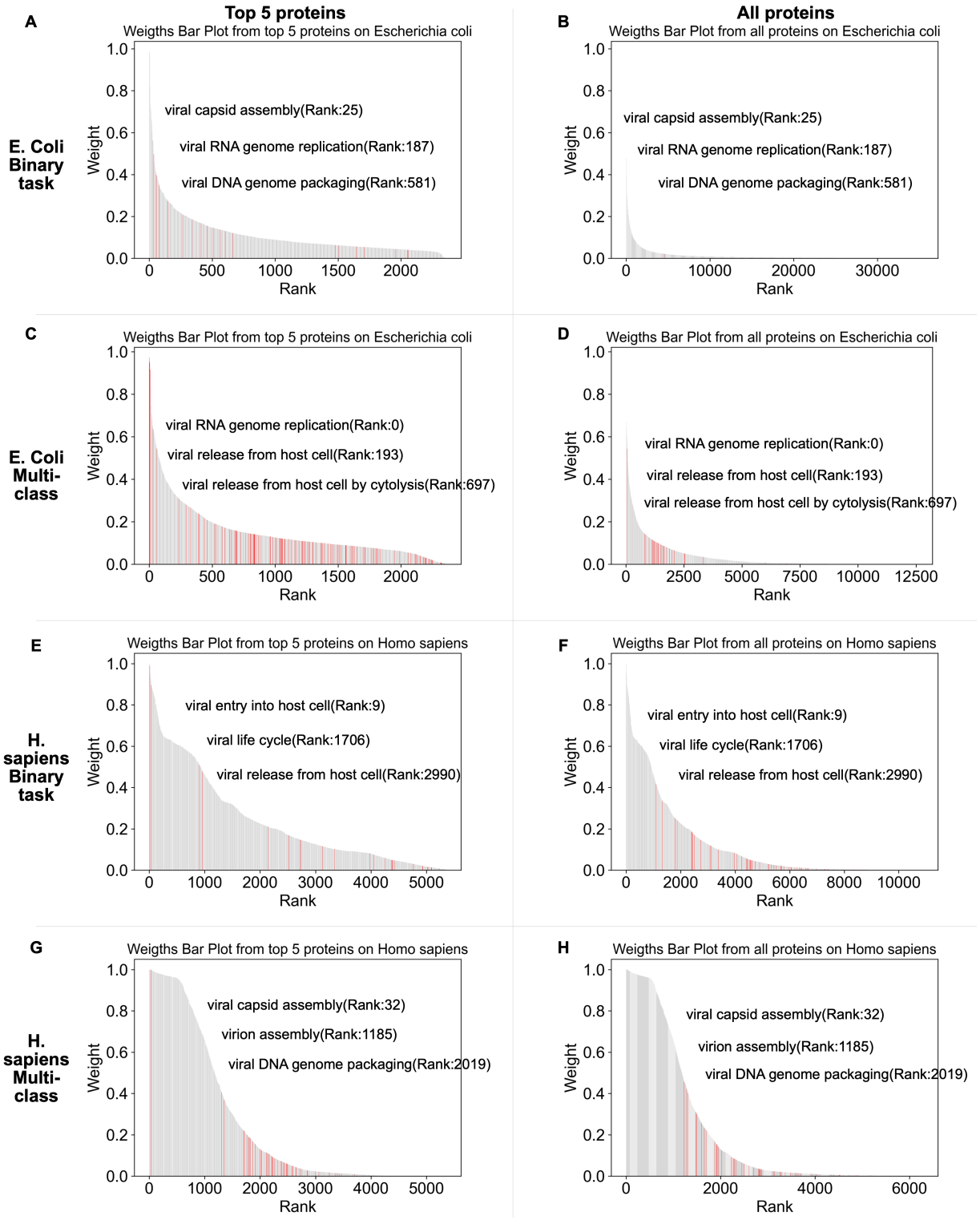

Figure 2.12. The bar plots display the ranking of weights for the top 5 proteins and all proteins of viruses associated with *E. coli* and *H. sapiens*, respectively. The top four bar plots illustrate protein weights obtained for *E. coli* based on binary classification models(A, B) and multi-class classification models (C, D), respectively. Similarly, the bottom four bar plots depict the protein weights obtained for *H. sapiens* based on binary classification models(E, F) and multi-class classification models (G, H), respectively. Each host has two sections: the left subplot shows the top 5 ranked protein weights, while the right subplot displays all protein weights sorted in descending order.

proteins.

Here we highlight results from viruses infecting *E. coli* and *H. sapiens*. These viruses contain the most GO annotations, because they are the most extensively studied hosts from prokaryotic and eukaryotic categories. Using the learned model parameters of the best-performing cross-validation model in the binary and multi-class classification tasks respectively, we ranked the attention weights of the proteins for the positive viruses and selected the top 5 ranked proteins from each virus. Furthermore, we collect two groups of top protein embeddings based on binary (Figure 2.13 A and C) and multi-class classification (Figure 2.13 B and D) models, respectively. These proteins were annotated with functions related to the viral life cycle using GO terms obtained from InterProScan (Jones et al. 2014).

For the set of top-ranked proteins of the viruses associated with *E. coli* based on the binary model, roughly 1.5% (35 out of 2359) were assigned a viral life cycle GO term. In Figure 2.13 A, there are four different GO terms related to the viral life cycle including viral capsid assembly (GO:0019069), viral DNA genome packaging (GO:0019073), viral RNA genome replication (GO:0039694), and viral entry into host cell (GO:0046718). In Figure 2.13 B, ten different GO annotations of virus proteins associated with *E. coli* based on the multi-class model, revealing a greater diversity of viral GO terms, approximately 3.7% (87 out of 2359) were assigned a viral life cycle GO term. Viral capsid assembly (GO:0019069), viral entry into host cell (GO:0046718), viral RNA genome replication (GO:0039694) are the top 3 GO annotations.

The top-ranked viral protein associated with *H. sapiens* displays significantly more GO annotations (Figure 2.13 C and D). In the binary case (Figure 2.13 C), we found 23.9% (1280 out of 5357), in the binary case, and 12.6% (676 out of 5357), in the multi-class case, of the top ranked proteins to have GO terms related to viral life cycle. It is worth noting that binary (Figure 2.13 C) and multi-class (Figure 2.13 D) shared most of the GO terms, and viral RNA genome replication (GO:0039694), attachment to host cell, virion attachment to host cell (GO:0019062) and viral life cycle (GO: 0019058) are the most prevalent.

Furthermore, we used hierarchical clustering (Nielsen 2016) to test if different proteins with the same functional annotation have similarities in the embedding space. In Figure 2.14, we clustered the ESM-1b embeddings of the abovementioned top-ranked proteins with viral life cycle GO annotations. We found apparent clusters across the viral proteins for *E. coli* (Figure 2.14 A and B) and *H. sapiens* (Figure 2.14 C and D). These results suggest ESM-1b embeddings encode functional signals, and MIL learns some consistent features associated with host specificity. While having clusters of proteins with the same GO annotation, we also observe that not all proteins with the same GO annotation are clustered into a single cluster. This suggests that richer information is contained in the embedding space beyond existing function annotations. Whether this information is biologically meaningful warrants a future study.

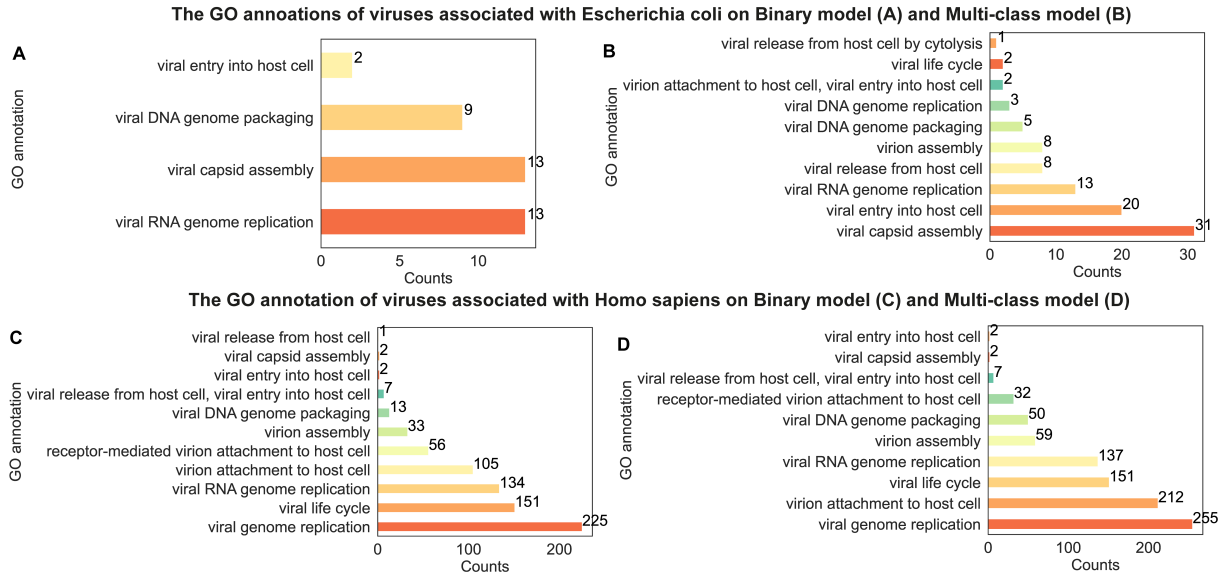

Figure 2.13. The bar plot of GO annotations of viral viruses associated with *E. coli* (top) and *H. sapiens* (bottom). Panels show the number of each GO annotation of the top 5 ranked proteins for each virus associated with *E. coli* (A, B) and *H. sapiens* (C, D). Here, the protein weights in A and C are obtained by binary models, whereas in B and D, the weights are obtained by multi-class classification models.

Table 2.7. Table of GO ID and GO terms

GO ID	GO terms
GO:0019076	viral release from host cell
GO:0046718	viral entry into host cell
GO:0044659	viral release from host cell by cytolysis
GO:0039693	viral DNA genome replication
GO:0019069	viral capsid assembly
GO:0019068	virion assembly
GO:0019073	viral DNA genome packaging
GO:0039694	viral RNA genome replication
GO:0019058	viral life cycle
GO:0019062, GO:0046718	virion attachment to host cell, viral entry into host cell
GO:0019062	virion attachment to host cell
GO:0019079	viral genome replication
GO:0046813	receptor-mediated virion attachment to host cell
GO:0019072	viral genome packaging
GO:0019076, GO:0046718	viral release from host cell, viral entry into host cell

The GO terms are coloured in the hierarchical clustering dendrograms. We retrieve protein functions for each GO ID, and GO terms are shown in the table.

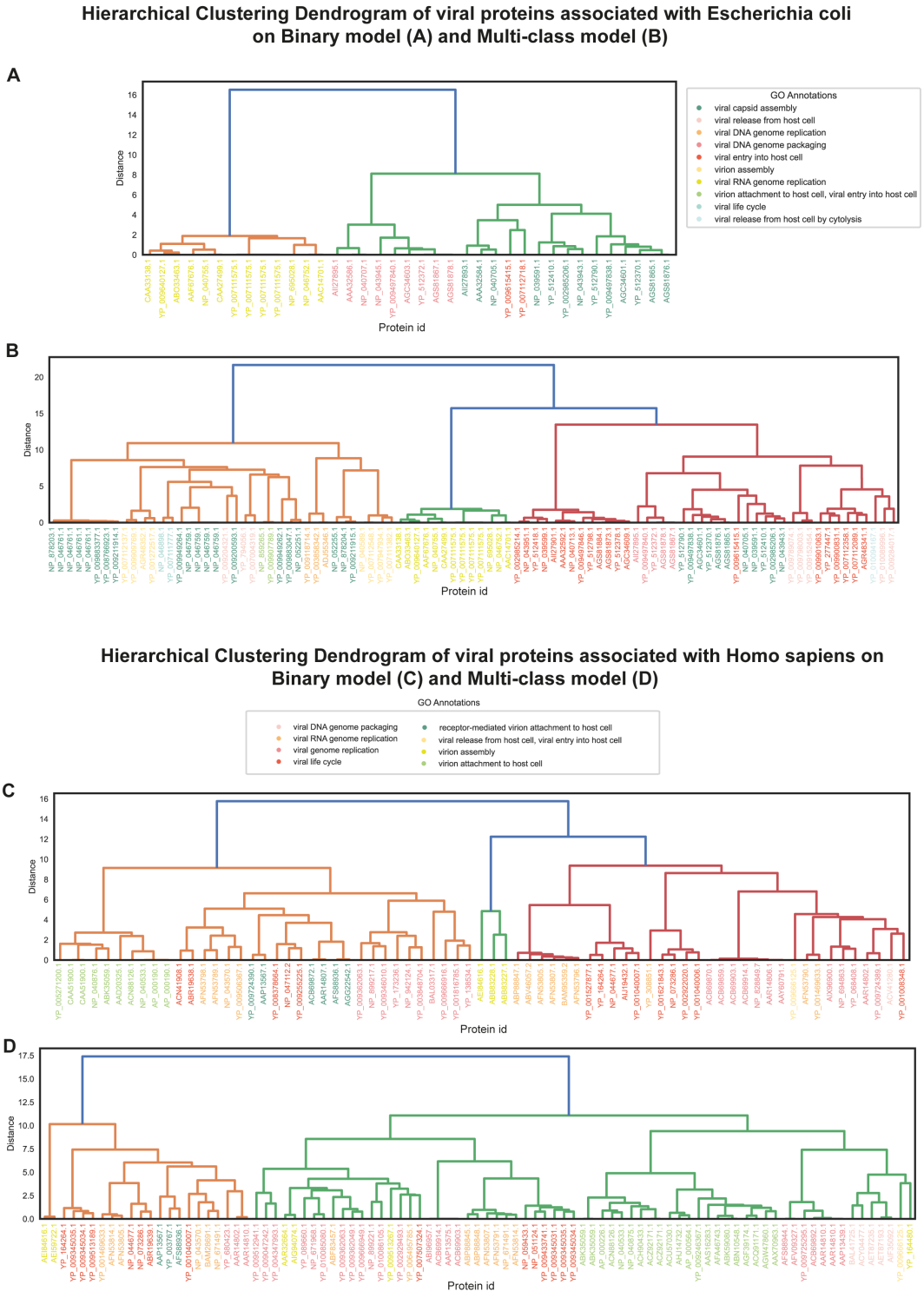

Figure 2.14. The dendrogram of protein embeddings of viruses associated with *E. coli* (top) and *H. sapiens* (bottom). Figures show the hierarchical clustering dendrograms of the top 5 ranked protein embeddings for each virus associated with *E. coli* (A, B) and *H. sapiens* (C, D). Here, the protein weights in A and C are obtained by binary models, whereas in B and D, the weights are obtained by multi-class classification models. Protein GO terms regarding the viral life cycle of viruses are highlighted with different colours, allowing us to understand predictive signals of proteins captured by the pre-trained transformer model.

2.4.6 EvoMIL identifies key proteins of SARS-CoV-2

To highlight EvoMIL performance on an unseen virus, we looked at the prediction and attention weights of SARS-CoV-2 (Wu et al. 2020) using the trained *H. sapiens* binary classifier from Section 2.4.2.1. Wuhan-Hu-1 with NCBI Reference Sequence NC_045512.2, which is not included in the training, was predicted to be a human virus with a probability of 0.97 compared to the SVM-kmer binary classifiers (Young et al. 2020) of less than 0.01. The top three ranked proteins contributing to host prediction are spike sub-sequences, Non-structural proteins sub-sequences and Nucleocapsid protein, with weights of 0.298, 0.221 and 0.189, respectively. The spike protein contains the receptor binding site which is responsible for cell entry. Nucleocapsid protein is an RNA binding protein that plays a critical role at many stages of the viral life cycle making direct interactions with many host proteins (Caruso et al. 2022). More N protein and host protein interactions can be retrieved from protein-protein interaction databases in UniPort (https://www.uniprot.org/uniprotkb/P0DTC9/entry#interaction). The viral process GO terms of these top-ranked proteins are fusion of virus membrane with the host plasma membrane (GO:0019064), fusion of virus membrane with host endosome membrane (GO:0039654), receptor-mediated virion attachment to host cell (GO:0046813) and endocytosis involved in viral entry into host cell (GO:0075509). It has been reported that these GO terms are associated with viral infections, indicating that protein attention weights obtained by EvoMIL perform well in identifying important proteins contributing to virus-host associations.

2.5 Discussion

In this chapter, we introduce EvoMIL, a novel method for virus-host prediction. Inspired by the success of NLP approaches in biology, we demonstrate the power of using a protein transformer language model (ESM-1b) to generate embeddings of viral proteins that are highly effective features for virus-host classification tasks. ESM-1b is capturing meaningful biological/host information from viral proteins, despite the ESM-1b training dataset being mainly comprised of proteins from cellular life with only 1% being viral proteins. We demonstrate that attention-based MIL can identify which of a virus's proteins are most important for host prediction, and by implication which proteins may be key to virus-host specificity.

Our classification results show that EvoMIL is able to predict the host at the species level with high AUC, accuracy and F1 scores. The prokaryotic binary classifiers achieved a mean AUC of 0.992 whereas the eukaryotic classifiers achieved a mean AUC of 0.851 during selecting negative samples by **Strategy 1**. We also evaluate the model's performance using different levels of negative sampling from **Strategy 2**, our findings indicate that training binary classifiers

becomes more challenging when those hosts associated with negative and positive samples are more closely related. Additionally, results demonstrated that eukaryotic host prediction is a more difficult task for two reasons: viruses associated with eukaryotic hosts are much more diverse across all seven Baltimore classes compared to the prokaryotic viruses which are mainly double-stranded DNA viruses; secondly, the eukaryote datasets contain many RNA viruses which have far fewer proteins, this makes it more challenging for MIL which needs many instances in each bag to perform well. Table 2.5 shows that ESM-1b features outperform more conventional sequence composition-based features on multi-class classification tasks. Furthermore, the clusters seen in the *E. coli* and *H. sapiens* hierarchical clustering dendrogram(Figure 2.14) indicates that embeddings are able to capture functionally related information of the high-ranked proteins associated with host specificity. Moreover, EvoMIL is able to find SARS-CoV-2 spike proteins from top-ranked proteins obtained by attention weights of the model.

Transformers enable a one-step pipeline to generate dense feature vectors from viral genomes and learn multiple layers of information about the structural and functional properties of proteins in an unbiased way. Multiple-instance learning is able to capitalize on any host signal contained within individual protein sequences, circumnavigating the need to represent each virus with a single feature vector. This enables it to find common patterns/signals across bags of proteins that differ greatly in both size and content. An additional advantage of attention-based MIL is the ability to identify which proteins are important in prediction. The identification of important proteins of viruses that contribute to host prediction is essential in understanding the specific viral proteins that play a critical role in host infection. Further analysis is needed to test whether this can be exploited to uncover the mechanisms behind virus-host specificity.

In this chapter, we took a virus-based approach in which only information from the viral genomes is used to generate features, this limits us to classifying viruses for those hosts that have a minimum number of known viruses. Also, by limiting the viral sequences to only Refseq sequences to reduce redundancy, we have further narrowed the range of hosts we can predict. Additionally, we acknowledge that Virus-host DB is biased towards well-studied organisms such as humans, to mitigate this, combining different virus-host association databases could enhance the diversity and comprehensiveness of the training samples. The ultimate host predictor would be able to make predictions for hosts which have no or few known viruses, to do this we will include host information. Combining virus and host-based approaches has been shown to greatly increase the host range of a prediction tool whilst maintaining a low false discovery rate, see Roux et al. (2022) and Wang et al. (2020). In the future, we will make use of the much larger numbers of host-annotated virus genomes available in public databases and include host information to construct a model that can make predictions using associations for any virus-host pair.

2.6 Supporting information

Table 2.8. Table of test virus samples used to benchmark EvoMIL on prokaryotic host species prediction. This table includes the test viruses for comparing EvoMIL and recent prokaryotic host species prediction approaches.

This table can be downloaded at: https://doi.org/10.1371/journal.pcbi.1012597.s007

Table 2.9. Table of benchmarking results of the test viruses on prokaryotic host species prediction. This table includes prediction results of prokaryotic host prediction methods on the test dataset provided in Table 2.8.

This table can be downloaded at: https://doi.org/10.1371/journal.pcbi.1012597.s008

Chapter 3

Protein-protein interaction prediction within species

Attention is all you need.

Vaswani et al. (2017)

3.1 Abstract

Computational prediction of protein structure from amino acid sequences alone has been achieved with unprecedented accuracy, yet the prediction of protein-protein interactions (PPIs) remains an outstanding challenge. Here, we assess the ability of protein language models (PLMs), routinely applied to protein folding, to be retrained for PPI prediction. Existing PPI prediction models that exploit PLMs use a pre-trained PLM feature set, ignoring that the proteins are physically interacting. Our novel method, PLM-interact, goes beyond a single protein, jointly encoding protein pairs to learn their relationships, analogous to the next-sentence prediction task from natural language processing. This approach provides a significant improvement in performance: Trained on human-human PPIs, PLM-interact predicts mouse, fly, worm, *E. coli* and yeast PPIs, with 16-28% improvements in AUPR compared with state-of-the-art PPI models. Additionally, it can detect changes that disrupt or cause PPIs. Our work demonstrates that large language models can be extended to learn the intricate relationships among biomolecules from their sequences alone.

3.2 Introduction

Proteins are the main structural components of cells and mediate biological processes by interacting with other proteins. Approximately 80% of proteins perform their biological function as part of protein complexes (Berggård et al. 2007). Human PPIs are associated with essential biological activities and have complex actions. Disruption of these protein-protein interactions (PPIs), e.g., mediated by mutations can underlie human disease (David et al. 2015). Understanding PPI mechanisms offers the potential to develop novel therapy strategies for both human disease and pathogen infections (Vassilev et al. 2004). Unfortunately, experimentally identifying PPIs is both costly and time-consuming, such that interaction datasets remain sparse with only a few species having comprehensive coverage (Kotlyar et al. 2015; Shin et al. 2020).

Computational algorithms offer an efficient alternative to the prediction of PPIs at scale. Existing prediction approaches mainly leverage protein properties such as sequence composition, evolutionary information and protein structures (Ben-Hur et al. 2005; Hashemifar et al. 2018; Huang et al. 2023). Applying these features to pairs of proteins, classifiers have been trained using classical machine learning (Shen et al. 2007) and deep learning approaches (Chen et al. 2019). Recently, protein language models (PLMs) trained on large public protein sequence databases have been used for encoding sequence composition, evolutionary and structural features, becoming the method of choice for representing proteins in state-of-the-art PPI predictors (Hallee et al. 2023; Sledzieski et al. 2023). A typical PPI prediction architecture uses a pre-trained PLM to represent each protein in a pair separately, then a classification head is trained for a binary task that discriminates interacting pairs from non-interacting pairs (Figure 3.1 A). For example, D-SCRIPT (Sledzieski et al. 2021) constructed a residue contact map using protein features extracted from Bepler and Berger's PLM (Bepler et al. 2021) to predict PPIs. TT3D (Sledzieski et al. 2023) combined protein embeddings obtained from a pre-trained PLM and one-hot encoding of 3D interaction (3Di) structural sequence predicted by FoldSeek (Van Kempen et al. 2024) to train a convolutional neural network PPI classifier. Despite this use of PLMs, identifying positive PPIs remains challenging.

The main issue is that PLMs are primarily trained using single protein sequences, i.e., while they learn to identify contact points within a single protein (Zhang et al. 2024), they are not 'aware' of interaction partners. In a conventional PLM-based PPI predictor architecture, a classification head is used to extrapolate the signals of inter-protein interactions by grouping common patterns of intra-protein contacts in interacting and non-interacting pairs, respectively (Figure 3.1 A). This strategy relies on the classification head being generalisable. Unfortunately, with a feedforward neural network being the dominant option, these classifiers often do not generalise well.

To address the lack of inter-protein context in training, we propose a novel PPI prediction model, PLM-interact, that directly models PPIs by extending and fine-tuning a pre-trained PLM, ESM-2 (Lin et al. 2023). PLM-interact (trained on human PPI data) achieves a significant improvement compared to other predictors when applied to mouse, fly, worm, yeast and *E. coli* datasets. We demonstrate that PLM-interact is capable of identifying mutations that cause and disrupt interactions, while the molecule interaction prediction tools AlphaFold3 (Abramson et al. 2024) and Chai-1 (Discovery et al. 2024) fail to correctly detect these mutation impacts (Figure 3.6). Finally, we released PLM-interact human-human PPI models, which were trained on two datasets: the benchmarking human PPI dataset from STRING V11 (Szklarczyk et al. 2019) and our created human PPI dataset source from STRING V12 (Szklarczyk et al. 2023).

3.3 PLM-interact

Inspired by the BERT architecture, which uses masked language modelling (MLM) to learn contextual representations, we developed PLM-interact, a PLM-based model trained on protein sequence pairs using MLM for data augmentation and regularisation. The model is designed to go beyond recognising exact matches, incorporating which proteins are most likely to be somewhere and the factors driving their interaction. We propose to use the transformer architecture to model inter-protein interaction, such that amino acids in a protein can be attended by amino acids from a different protein.

To directly model PPIs, two extensions to ESM-2 are introduced (Figure 3.1 B): (1) longer permissible sequence lengths in paired masked-language training to accommodate amino acid residues from both proteins; (2) implementation of the next sentence prediction task (Reimers et al. 2019) to fine-tune ESM-2 where the model is trained with a binary label indicating whether the protein pair is interacting or not (see Methods for more details). Our training task is, thus, a mixture of the next sentence prediction and the masked language modelling tasks. This architecture allows amino acids in a protein to be linked by amino acids from a different protein through the transformer's attention mechanism.

The training of PLM-interact begins with the pre-trained large language model ESM-2. We fine-tune it for PPIs, by showing it pairs of known interacting and non-interacting proteins. In contrast to similar training strategies in machine learning (Reimers et al. 2019), we find that the next sentence prediction and mask language modelling objectives need to be balanced. We therefore conducted comprehensive benchmarking for different weighting options, before selecting a 1:10 ratio between classification loss and mask loss, combined with initialisation using the ESM-2 (with 650M parameters), as this achieved the best performance (see Methods and Figure 3.2).

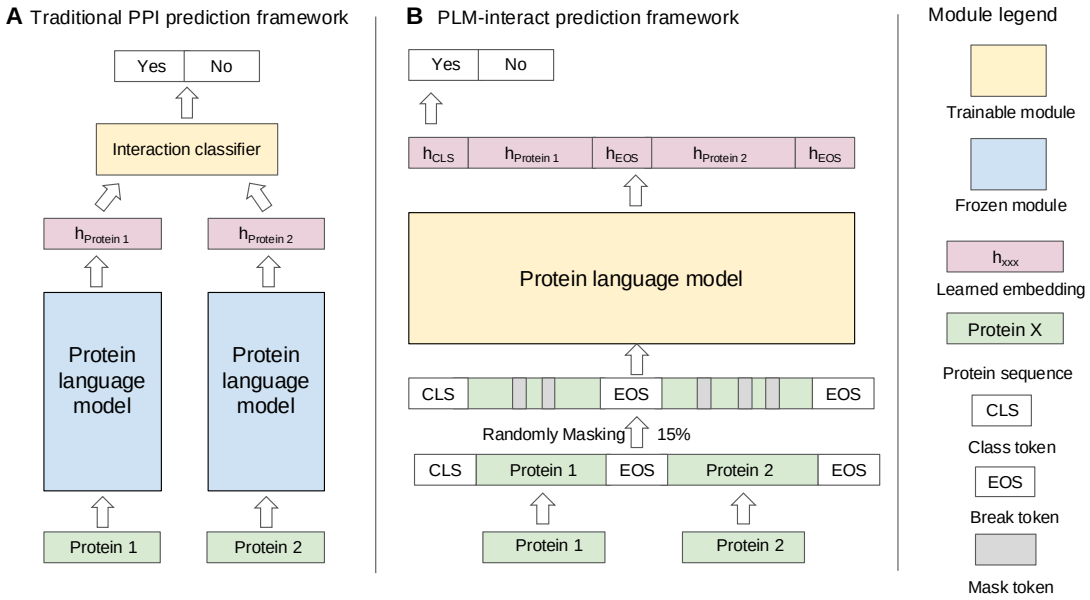

Figure 3.1. A comparison of PLM-interact to the typical existing PPI prediction framework. A. PPI prediction models that use pre-trained protein language models to extract single protein embeddings. Then, an interaction classifier is trained using these single protein embeddings. **B**. PLM-interact uses a protein language model with a longer context to directly handle a pair of protein sequences. Both the mask language modelling task and a binary classification task predicting interaction status are used to train the model (see Figure 3.2).

3.4 Methods

3.4.1 Datasets

The benchmarking human PPI dataset, from Sledzieski et al. (2021), comprises human training and validation data and test data from five other species: mouse, *Mus musculus*; fly, *Drosophila melanogaster*; worm, *Caenorhabditis elegans*; yeast, *Saccharomyces cerevisiae*; and *E. coli*, *Escherichia coli*), all retrieved from STRING V11 (Szklarczyk et al. 2019). We train and validate our model on human PPIs and then conduct inference on PPIs from five other species. All training, validation and test datasets maintain a 1:10 ratio of positive to negative pairs, reflecting the fact that positive PPIs are significantly fewer than negative pairs in the actual host PPI networks. The length of protein sequences ranges from 50 to 800, and PPIs are clustered at 40% identity using CD-HIT (Li et al. 2006) to remove the high-identity PPIs. There are a total of 43,138 positive human PPIs; roughly 10% of positive human PPIs are used to construct a validation dataset, and the remaining human PPIs are used to create a training dataset. The human training dataset includes 38,344 positive PPIs, whereas the validation set includes 4,794 positive PPIs. Each of the five species includes 5,000 positive interactions, except for *E. coli*, which only has 2,000 positive interactions due to the fewer positive PPIs in the STRING dataset used (Sledzieski et al. 2021).

In addition, we provide models trained on human PPIs from STRING V12 (Sledzieski et al. 2021). The positive PPIs are selected by collecting physical links with positive experimental scores while excluding PPIs with positive homology scores and confidence scores below 400. Previous studies have typically limited the maximum length of protein sequences to 800 or 1000 due to GPU memory limitations. We process the length of protein sequences up to 2000, with a combined length threshold for PPIs of nearly 2500. This human dataset includes 60,308 positive PPIs for training and 15,124 positive PPIs for testing. Furthermore, protein sequences are clustered at 40% identity using MMSeq2 (Steinegger et al. 2017) and only PPIs from the distinct clusters are chosen to eliminate high-identity PPIs. Again, the positive-to-negative protein pair ratio is 1:10, consistent with the aforementioned two benchmarking datasets.

3.4.2 Model architecture

We use ESM-2 as the base model in PLM-interact. ESM-2 is an encoder transformer model with a parameter size range from 8 million to 15 billion. The results presented are PLM-interact based on ESM-2 with 650M parameters. We also provided PLM-interact models with ESM-2 35M on our GitHub repository to help with quick testing. The input representation contains amino acid token representations from two proteins. This setup is similar to the original BERT model (Devlin et al. 2018), also known as the cross-encoder which simultaneously encodes a pair of query and answer sentences.

A standard input sequence of PLM-interact, x, can be shown as the following:

$$x = [CLS, P_1, EOS, P_2, EOS], \tag{3.1}$$

where CLS is the classification token, P_1 contains amino acid tokens of protein 1, P_2 contains amino acid tokens of protein 2, and EOS is the end-of-sentence token. The first EOS token marks the end of the amino acid sequence in protein 1. This setting allows us to use the original ESM-2 tokenizer to generate embedding vectors e, and pass them to the transformer encoder of the ESM-2:

$$h = f(e), \tag{3.2}$$

where f is ESM-2, e contains the token embeddings of x, and h contains the output embeddings of all input tokens. h can be presented as:

$$h = \{h_{CLS}, h_{a_1}, \ldots, h_{EOS}, \ldots, h_{a_n}, \ldots, h_{EOS}\}, \tag{3.3}$$

where h_{a_1} and h_{a_n} represent amino acid tokens in proteins 1 and 2. Then, we use the *CLS* token embedding to aggregate the representation of the entire sequence pair and as the features for a linear classification function φ, and parameterised as a single fully connected feed-forward (FF) layer with a ReLU activation function. The output of the FF layer is converted by the sigmoid function σ to obtain the predicted interaction probability g,

$$g = \sigma(\varphi(h_{CLS})) \tag{3.4}$$

3.4.3 Model Training

PLM-interact is trained with two tasks: 1) a masked language modelling (MLM) task predicting randomly masked amino acids and 2) a binary classification task predicting the interaction label of a pair of proteins. PLM-interact is trained for 10 epochs using a batch size of 128 on the human PPI benchmarking dataset. For all training runs, the input protein pairs are trained on both orders as the interaction between protein 1 and protein 2 is the same as the protein 2 and protein 1, which leads to double the size of the training set. The validation and testing sets are not subject to the same data argumentation. The learning rate is 2e-5, weight decay is 0.01, warm up is 2000 steps, and the schedular is WarmupLinear which linearly increasing the learning rate over the warmup steps. During training, we evaluate the model's performance at every 2000 steps on the validation set. For every evaluation, a set of 128 protein pairs is randomly sampled from the validation set and the results are averaged over 100 times to ensure metric reliability. Here, we use both masking and classification losses to optimize our model, the loss function for each data point l can be represented as:

$$l = \alpha l_{\text{mlm}} + \beta l_{\text{ce}}, \tag{3.5}$$

where l_{mlm} and l_{ce} separately represent the MLM loss and classification (ie cross entropy) loss. l can be written as:

$$l = -\frac{\alpha}{M} \sum_{i=1}^{M} \ln p(x_i \mid x_{(-i)}) - \beta \left(y \ln(g) + (1-y) \ln(1-g) \right), \tag{3.6}$$

where M is the number of the masked tokens, x_i is the true token at position i, $p(x_i \mid x_{(-i)})$ is

the probability of the true token x_i given the unmasked amino acid x_{-i}. y is the label of the interaction and g is the predicted probability for $y = 1$, obtained from Equation 3.4. α and β are weights for the MLM and classification losses.

All of the models are trained on the DiRAC Extreme Scaling GPU cluster Tursa. A typical 10-epoch training run of the model with ESM-2 (650M) with human PPIs takes 31.1 hours on 16 A100-80 GPUs. The model with ESM-2 (650M) trained on STRING V12 human PPIs used 16 A100-80 GPUs for 86.4 hours. For model training time with different ratios and model sizes, see the following Section 3.4.4 optimisation experiment and supplementary Table 3.1. We provide model checkpoints that include human PPI models trained on Sledzieski et al.'s benchmarking datasets retrieved from STRING V11 (Szklarczyk et al. 2019), as well as human PPIs collected by us from STRING V12 (Szklarczyk et al. 2023).

Table 3.1. The table of different models' GPU hours (GPUhs).

Models/GPUhs	650M	35M
0:1	522.4	238.2
1:1	514.3	299.1
1:5	508.4	238.1
1:10	496.9	240.4
Binary	355.8	130.9

PLM-interact has two kinds of training strategies: masking language modelling and binary classification, binary classification only. The ratios in this table present the different weights between mask loss and classification loss, binary indicates the binary classification task.

3.4.4 PLM-interact optimisation experiments

To find the optimal value of α and β in Equation 3.6, we benchmark a range of different options between mask loss and classification loss on human benchmarking data. For each ESM-2-35M and ESM-2-650M model, we train five models with different settings of ratios $\alpha : \beta$ between mask loss and classification loss. The ratios are $\alpha : \beta$ = 1:1, 1:5, 1:10, 0:1 (with mask), and 0:1 (without mask, denoted as classification) (Figure 3.2 B and C). We used the human validation set for each model to identify the optimal epoch checkpoint achieving the best AUPR. Next, the final model is selected based on testing on five other host PPIs. The results of optimisation experiments are shown in Section 3.5.1.

3.4.5 Baselines

We compute the prediction interaction probabilities based on checkpoints of TT3D (Sledzieski et al. 2023), Topsy-Turvy (Singh et al. 2022) and D-SCRIPT (Sledzieski et al. 2021) to generate precision-recall (PR) curves in Figure 3.4 B. Due to the absence of publicly available checkpoints for DeepPPI (Richoux et al. 2019) and PIPR (Chen et al. 2019), these methods are excluded from the PR curve comparison. The AUPR values for DeepPPI and Topsy-Turvy are sourced from the Topsy-Turvy paper, those for D-SCRIPT and PIPR are from the D-SCRIPT paper, and the AUPR value of TT3D is obtained through email communication. A complete list of each baseline method's main features, architectures, references, and code links can be found in Table 3.2.

Table 3.2. The table that shows main features, architectures, references, and code links of state-of-the-art models in this chapter.

Models	Protein features	Model architecture	Source	Tool
TT3D	Pre-trained Bepler & Berger PLM embeddings and one-hot encoding of the Foldseek 3Di	Convolutional neural network	Sledzieski, S. et al. 2023	GitHub
D-SCRIPT	Pre-trained Bepler & Berger PLM embeddings	Convolutional neural network	Sledzieski, S. et al. 2021	Website
Topsy-Turvy	Pre-trained Bepler & Berger PLM embeddings and network structure	Integrates D-SCRIPT and a network-based model GLIDE	Singh, R. et al. 2022	Website
PIPR	Pre-trained amino acid embeddings using Skip-Gram model	Siamese residual RCNN	Chen, M. et al. 2019	GitHub
DeepPPI	One-hot encoding amino acids	Fully connected model architecture	Richoux et al., 2019	GitLab
STEP	Pre-trained PLM ProtBERT embeddings	Siamese neural network	Madan, S. et al. 2022	GitHub

3.4.6 MMseq2

We used MMseq2 (Steinegger et al. 2017) to obtain the protein sequence-based alignment results between each pair of proteins; the parameter setting is: –threads 128 –min-seq-id 0.4 –alignment-mode 3 –cov-mode 1.

3.4.7 Chai-1

Chai-1 (Discovery et al. 2024) is a state-of-the-art model for molecular structure prediction, available at https://lab.chaidiscovery.com/. We use Chai-1 with the "specify restraints" option to predict protein-protein structure complexes and visualise predicted PPI structures using the molecular visualisation program ChimeraX (Pettersen et al. 2021).

3.4.8 AlphaFold3

AlphaFold3 is a tool to predict biomolecular interactions, including protein, DNA, small molecules, ions and modified residues (Abramson et al. 2024), available athttps://alphafoldserver.com/. We use AlphaFold3 in its PPI mode to predict protein structure complexes. The results are visualised with the molecular visualisation program ChimeraX (Pettersen et al. 2021).

3.4.9 Evaluation metrics and statistical analysis

3.4.9.1 Evaluation metrics

In this study, we use AUPR to evaluate models' performance and plot Precision-Recall (PR) curves (Davis et al. 2006) to compare our model with the state-of-the-art PPI models. AUPR is calculated based on precision and recall at different classification thresholds and represents the area under the PR curve, summarising the model's overall performance across different decision thresholds. Therefore, the PR curve is more informative than the ROC curve in evaluating the model's performance on our imbalanced training and test PPIs.

Precision is the proportion of true positives on all predicted positive samples.

$$Precision = \frac{TP}{TP+FP}, \tag{3.7}$$

Recall is the proportion of all true class samples predicted to be true classes.

$$Recall = \frac{TP}{TP+FN}, \tag{3.8}$$

True positive (TP): The model correctly predicts the positive samples. True Negative (TN): The model correctly predicts the negative samples. False Positive (FP): The model predicts negative

samples to be positive. False Negative (FN): The model predicts positive samples to be negative.

We used perplexity (Jelinek et al. 2005) to evaluate our models' performance on masking token prediction. Perplexity is the exponentiated average negative log-likelihood of the predicted tokens in a sequence and is a common evaluation metric for language models. Perplexity is used to show the uncertainty and complexity of a model; the lower perplexity indicates that the model has more confident predictions and better performance.

3.4.9.2 McNemar's test

McNemar's Chi-Square test (McNemar 1947) is a statistical test that determines if there are significant differences between paired nominal data.

$$\text{McNemar's test} = \frac{(b-c)^2}{b+c} \tag{3.9}$$

Here, b represents the count of protein pairs where model 1 obtained the correct prediction and model 2 obtained the incorrect prediction, while c represents the count of protein pairs where model 1 obtained the incorrect prediction and model 2 obtained the correct prediction.

To investigate if our models with varying ratios of mask-to-classification loss perform significantly differently, we conducted a McNemar's test for any pairs of models in optimisation experiments. This test is based on the number of correct and incorrect between two models. Predicted interaction probabilities from each model are used to get predicted labels, which are used to obtain the counts of correct and incorrect predictions. A McNemar's test p-value ≤ 0.05 indicates a significant difference between the predictive performance of the two models. The model with more correct predictions is considered superior to the other.

3.4.10 Data availability

Sledzieski et al.'s benchmarking PPI data is available at https://d-script.readthedocs.io/en/stable/data.html

STRING V12 PPIs database: https://stringdb-downloads.org/download/protein.physical.links.v12.0.txt.gz

The mutations that cause or disrupt PPIs are from the IntAct Database; the link is https://ftp.ebi.ac.uk/pub/databases/intact/current/various/mutations.tsv

UniProt: https://www.uniprot.org/

The analysis scripts are available at https://github.com/liudan111/ESM2-interact.git

The model checkpoints are available at https://huggingface.co/danliu1226

3.5 Results

3.5.1 PLM-interact optimisation results

To investigate the importance of both MLM and binary classification on PPI prediction performance, we conducted a benchmarking analysis using the different ratios between mask loss and classification loss on Sledzieski et al.'s benchmarking data. For details on the training and evaluation process, refer to Section 3.4.3. For ESM-2-650M, the 1:10 ratio is the optimal choice (McNemar test, the greatest counts of p-values less than 0.05) compared to other options (Figure 3.2 D). The model with a 1:5 ratio of mask to classification loss based on ESM-2-35M achieved the highest number of significant improvements (McNemar test, the greatest counts of p-values less than 0.05) compared to other models (Figure 3.2 E). According to these results, we select a loss ratio of 1:10 for ESM-2-650M and 1:5 for ESM-2-35M. The ratio setting is implemented in benchmarking of human PPI training, as well as human PPI training using the STRING V12 database.

To investigate how our model leverages the benefits of mask loss for the model performance, we use token prediction accuracy and perplexity to evaluate the model's performance on three tasks with varying ratios of mask to classification losses (1:1, 1:5 and 1:10). High token prediction accuracy is better, whereas low perplexity is better. Figure 3.3 shows that the mask token accuracy of mouse and fly is higher than yeast and *E. coli* in the 35M (Figure 3.3 A) and 650M models (Figure 3.3 C). Additionally, the 650M model is typically better than the 35M model. We further discovered that a loss ratio 1:1 of both 35M and 650M models achieved the best performance on the mouse with the highest token accuracy and lowest perplexity. The mouse, fly, and worm have a closer evolutionary relationship to the training species humans, resulting in higher token prediction accuracy and lower perplexity than the evolutionary divergent species yeast and *E. coli*. These results are consistent with the benchmarking results of held-out species PPI prediction (Figure 3.3 B). We demonstrated that the varying ratios of the mask-to-classification losses impact model performance, and the larger model generally performs better.

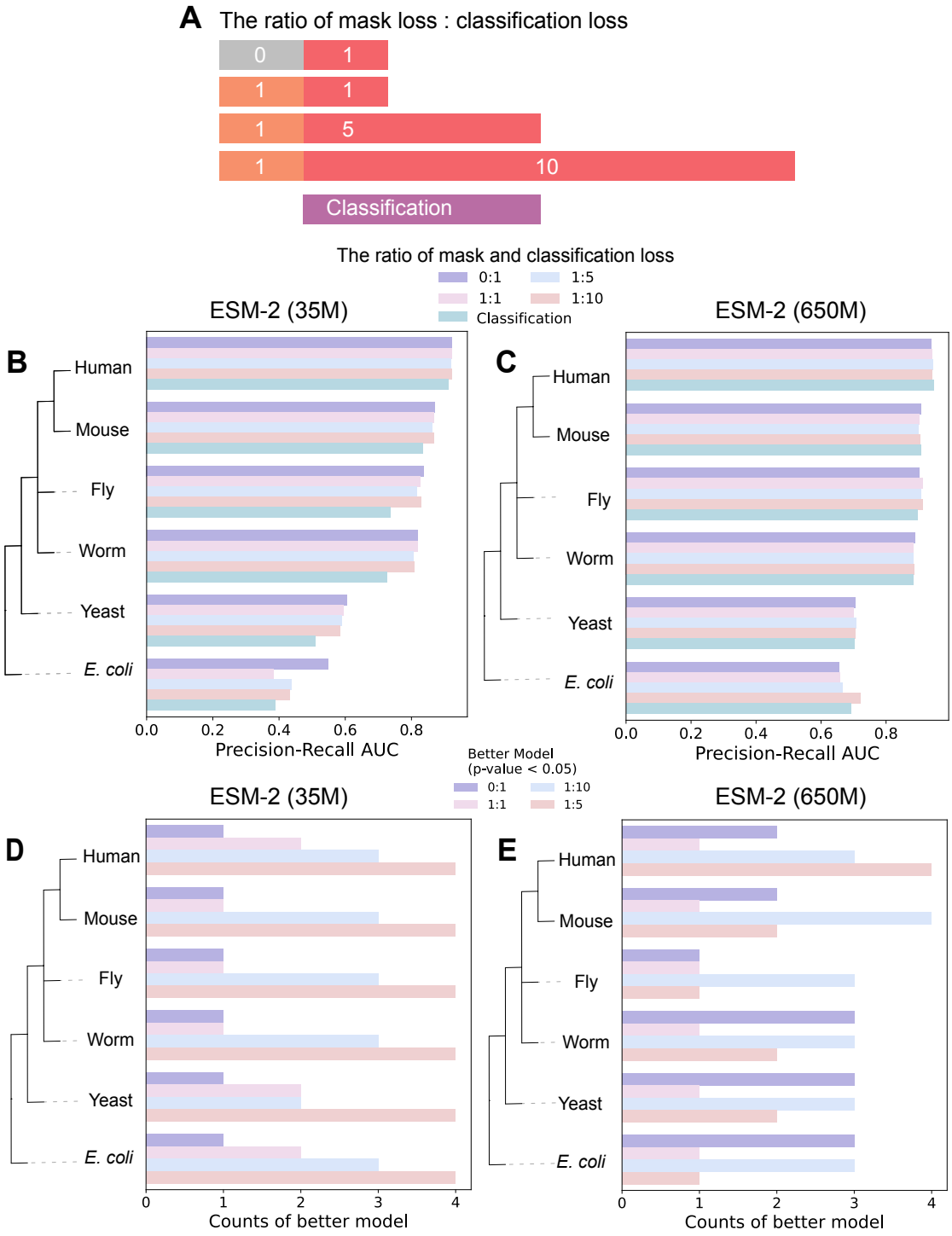

Figure 3.2. The benchmarking of different ratios of mask to classification loss on five species PPI prediction. **A**.The bar plot shows the ratio between mask loss and classification loss. **B** and **C** respectively represent the performance of our model with the different ratios between mask and classification loss on 650M and 35M of ESM-2 models. The left is aligned with the taxonomy tree of the hosts that are used for evaluating our human PPI model. **D** and **E** show the better model with significant improvements than the other, p-value < 0.05 with McNemar's Test.

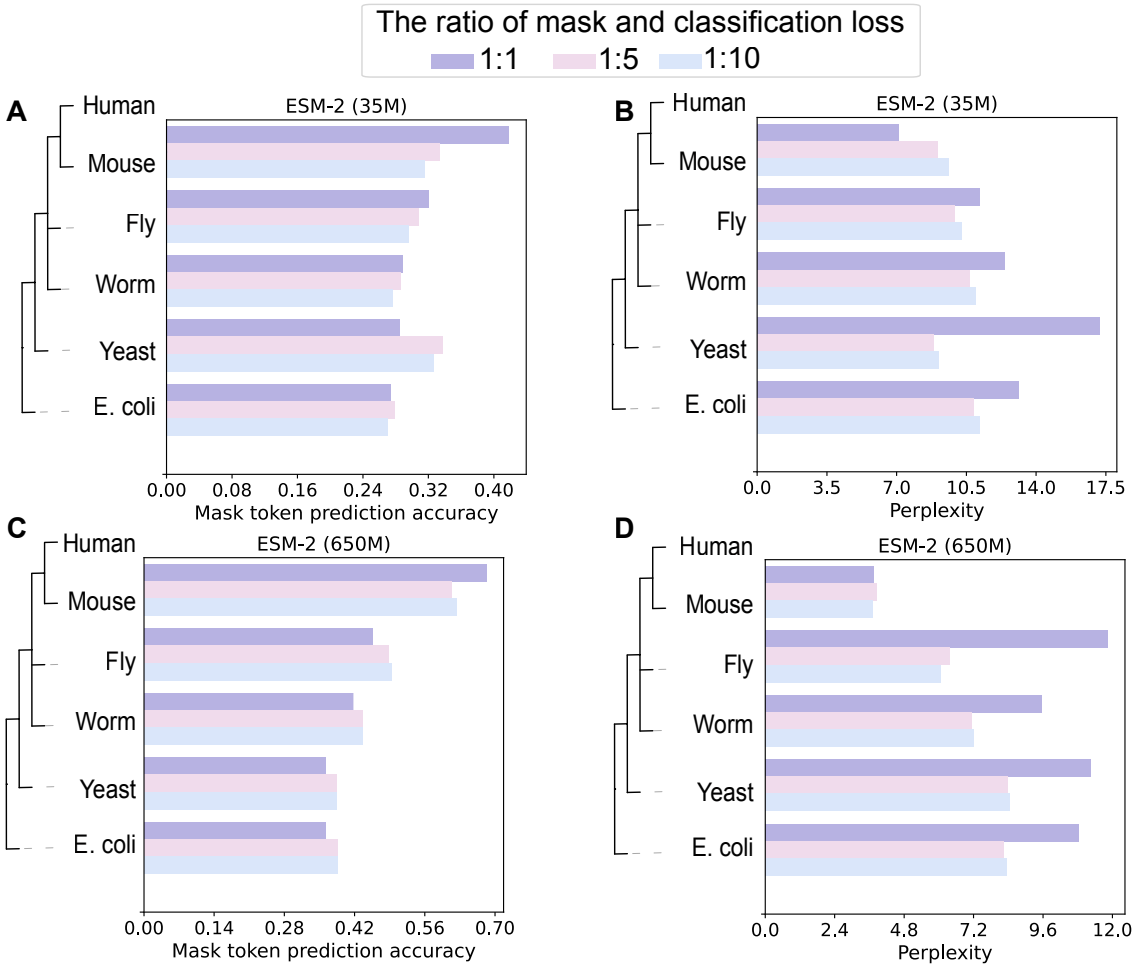

Figure 3.3. The mask token prediction accuracy and perplexity of PLM-interact with the different ratios between mask loss and classification loss.

3.5.2 PLM-interact improves prediction performance

To examine the performance of PLM-interact, we benchmark the model against five PPI prediction approaches: TT3D (Sledzieski et al. 2023), Topsy-Turvy (Singh et al. 2022), D-SCRIPT (Sledzieski et al. 2021), PIPR (Chen et al. 2019) and DeepPPI (Richoux et al. 2019). We use a multi-species dataset created by Sledzieski et al.(Sledzieski et al. 2021). Each model is trained on human protein interaction data and tested on five other species. The human training dataset in this multi-species dataset includes 421,792 protein pairs (38,344 positive interaction pairs and 383,448 negative pairs), human validation includes 52,725 protein pairs (4,794 positive interaction pairs and 47,931 negative pairs) and the mouse, worm, fly and yeast test datasets each includes 55,000 pairs (5,000 positives interaction pairs and 50,000 negative pairs), except for the *E. coli* test dataset, which includes 22,000 pairs (2,000 positive interaction pairs and 20,000 negative pairs). The positive PPIs in these datasets are experimentally-derived physical interactions, while the negative pairs are randomly paired proteins not reported to interact.

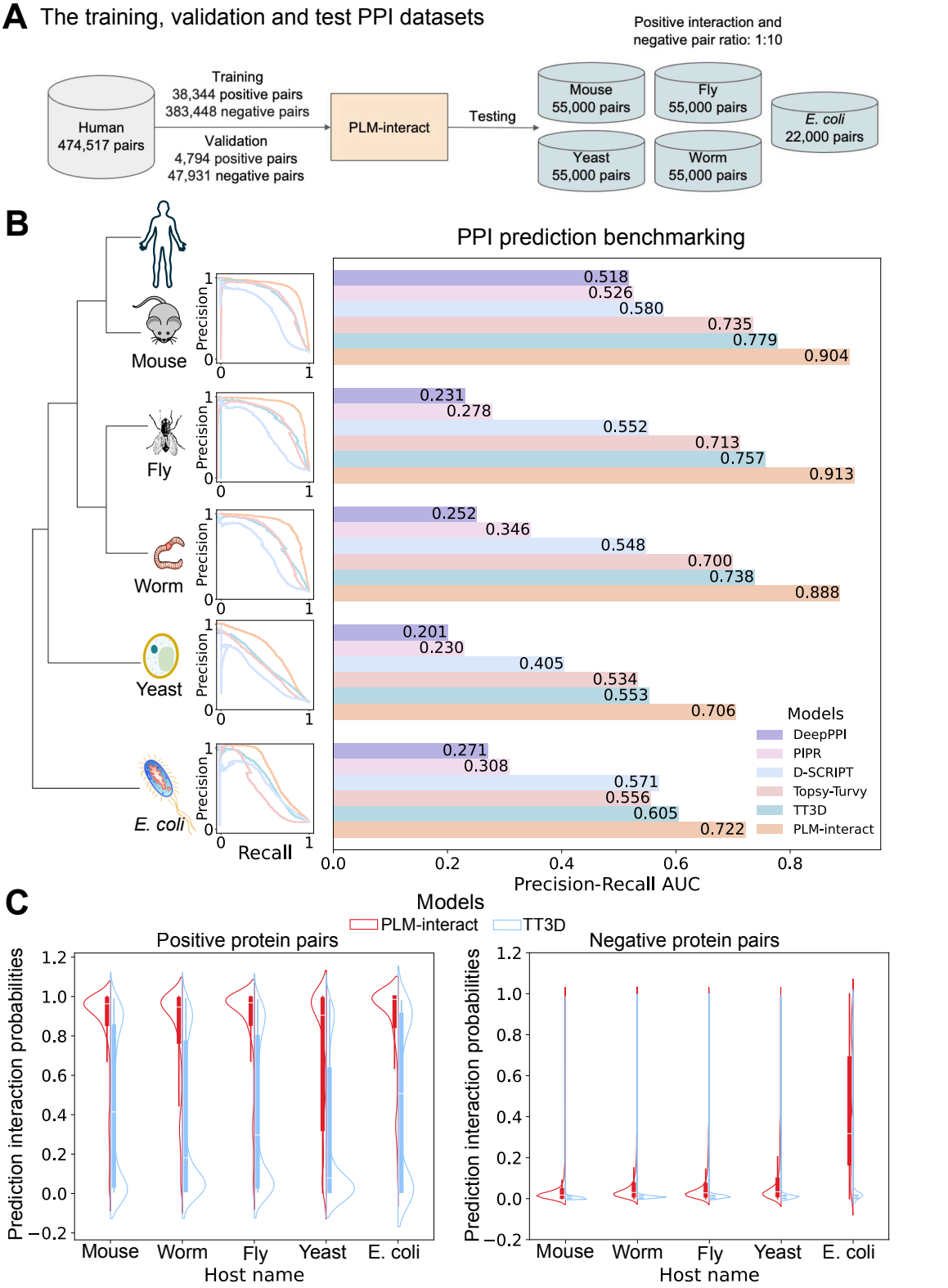

Figure 3.4. PLM-interact achieves the highest PPI prediction performance. The benchmarking results of PLM-interact with state-of-the-art PPI prediction models. **A**. The data size of training, validation and test PPIs. **B**. The taxonomic tree of the training and test species, precision-recall curves of each test species and a bar plot of AUPR values on PPI prediction benchmarking. **C**. Violin plots of predicted interaction probabilities of PLM-interact and TT3D on positive and negative pairs, respectively. (see Figure 3.5 for more information).

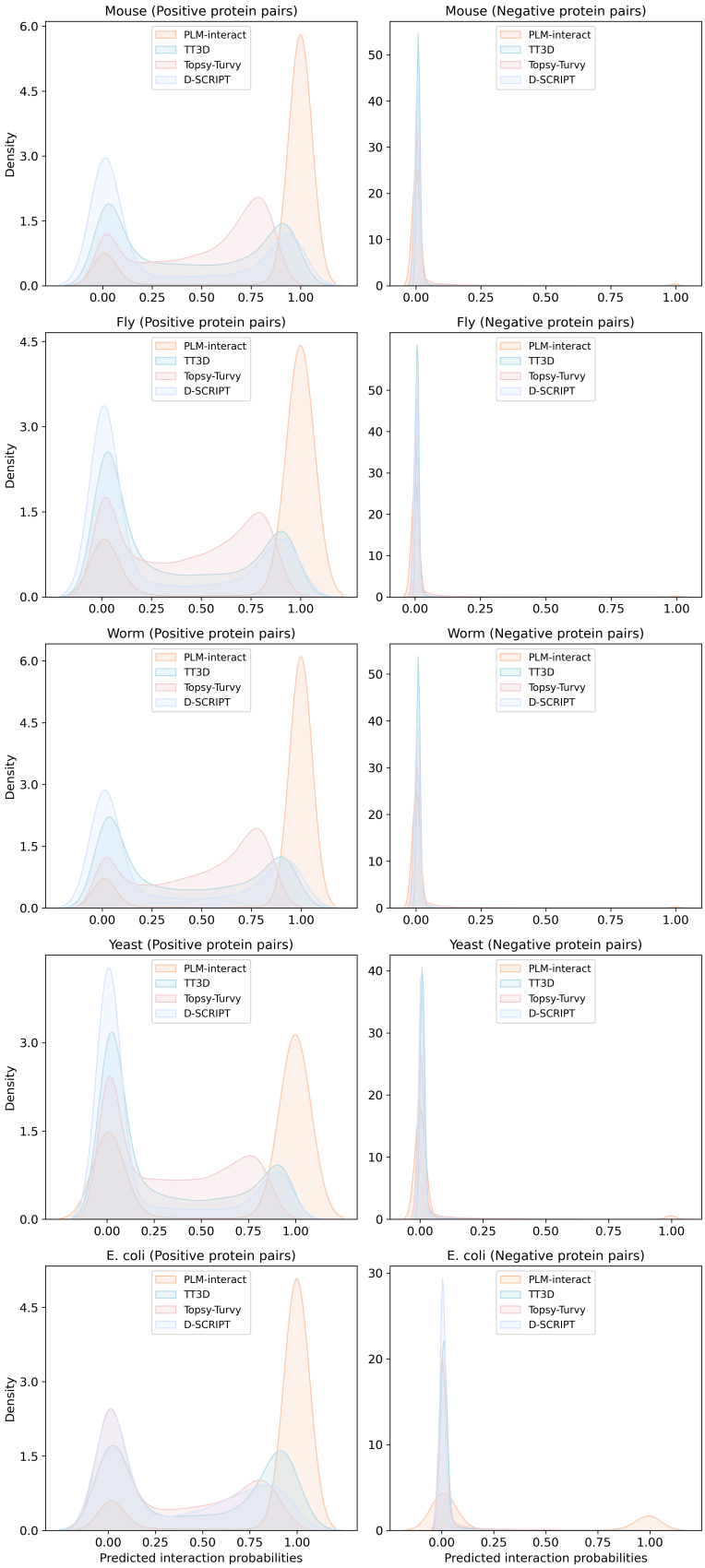

Figure 3.5. The distribution of prediction scores of positive and negative protein pairs of PLM-interact, TT3D, TT and D-SCRIPT. PLM-interact outperforms other models by identifying the greatest number of true positive pairs (predicted interaction probability > 0.5) and true negative pairs (predicted interaction probability < 0.5), demonstrating that PLM-interact achieves the best performance, except for negative pairs of *E. coli*.

PLM-interact achieves the highest AUPR (area under the precision-recall curve) (Davis et al. 2006) with the next best performer, TT3D, although similar performance to Topsy-Turvy (Figure 3.4 B). Tested on mouse, fly and worm test species datasets, PLM-interact has an improvement of 16%, 20% and 20% AUPR compared to TT3D, respectively. The predictions for yeast and *E. coli* PPIs are more challenging because they are more evolutionarily divergent from the human proteins used for training than the other species (see Figure 3.4 B): Our model achieved an AUPR of 0.706 on yeast, a 28% improvement over TT3D's AUPR of 0.553 and a 19% improvement on *E. coli* with an AUPR of 0.722.

To further present the performance of our model against recent PPI models, we show the precision-recall curve for each test species alongside the AUPR bar plot (see Figure 3.4 B). DeepPPI and PIPR were excluded from precision-recall curve plotting due to the absence of model checkpoints, their AUPR values in the bar plot were obtained from the published paper. To illustrate the prediction performance on positive and negative PPIs separately, we present violin plots of the prediction interaction probabilities of PPIs for PLM-interact and the second-best predictor TT3D (see Figure 3.4 C). The results show that PLM-interact demonstrated superior performance in predicting PPIs, except for negative pairs of *E. coli*. The distribution of PLM-interact's prediction interaction probabilities of negative protein pairs in *E. coli* is broader than TT3D and shows more false positive protein pairs with a prediction interaction probability higher than 0.5, indicating that the prediction of negative protein pairs is more challenging than positive protein pairs in *E. coli*.

Importantly, the improvement in PLM-interact is due to its ability to correctly identify positive PPIs: Comparing the prediction interaction probability of PLM-interact with the second-best predictor TT3D, PLM-interact consistently assigned higher probabilities of interaction to true positive PPIs. TT3D, in contrast, despite using a broader feature set, produces a bimodal distribution for interaction probabilities in all held-out species (more details see Figure 3.5).

Next, we showcase five positive PPI instances, one for each test species, for which our model produces a correct prediction, but TT3D produces an incorrect prediction (Figure 3.6). These PPIs are necessary for essential biology processes including ubiquinone biosynthesis, RNA polymerisation, ATP catalysis, transcriptional activation and protein transportation. We use Chai-1 (Discovery et al. 2024) and AlphaFold3 (Abramson et al. 2024) to predict and visualise these interacting protein structures (Figure 3.6 and Figure 3.7). Notably, both Chai-1 and AlphaFold3 have only 1 out of 5 structures with close to high-confidence prediction (ipTM close to 0.8). Both methods give failed prediction scores (ipTM <0.6) for 4 out of 5 structures. PLM-interact gives correct predictions with high confidence in all cases. Here, ipTM is interface predicted template modelling and pTM, a confidence score used for measuring the entire structure accuracy (Xu 2010; Zhang 2007). The ipTM scores above 0.8 represent confident pre-

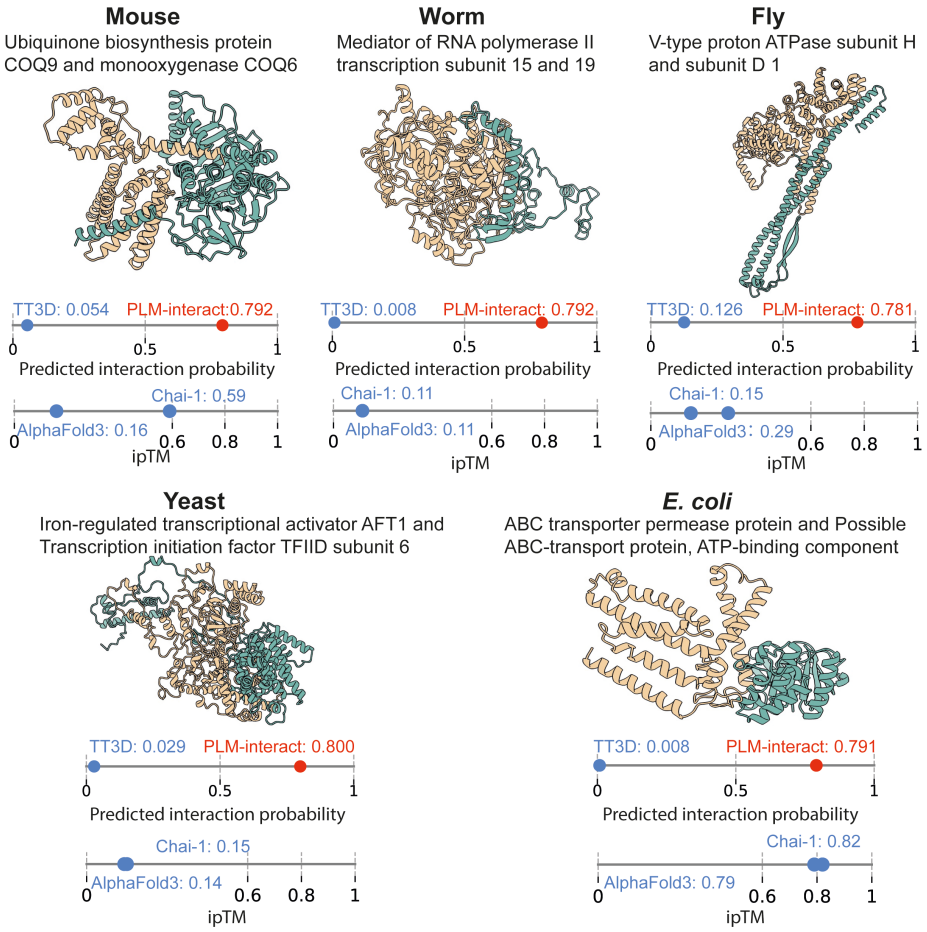

Figure 3.6. PPI example for each species that was predicted correctly by PLM-interact but not by TT3D. Protein-protein structures are predicted by Chai-1 (Discovery et al. 2024) and visualised with ChimeraX (Pettersen et al. 2021). Both models' prediction interaction probabilities range between 0 and 1. A predicted interaction probability >0.5, is predicted as a positive PPI, while <0.5 is a negative pair. Interacting proteins are shown from left (yellow) to right (green), respectively, for **Mouse**: Q8K1Z0 (Ubiquinone biosynthesis protein COQ9, mitochondrial) and Q8R1S0 (Ubiquinone biosynthesis monooxygenase COQ6, mitochondrial); **Worm**: Q21955 (Mediator of RNA polymerase II transcription subunit 15) and Q9N4F2 (Mediator of RNA polymerase II transcription subunit 19); **Fly**: Q9V3J1 (V-type proton ATPase subunit H) and Q9V7D2 (V-type proton ATPase subunit D 1); **Yeast**: P22149 (Iron-regulated transcriptional activator AFT1) and P53040 (Transcription initiation factor TFIID subunit 6); and ***E. coli***: A0A454A7G5 (ABC transporter permease protein) and A0A454A7H5 (Possible ABC-transport protein, ATP-binding component). See Figure 3.7 for AlphaFold (Abramson et al. 2024) predicted structures. The ipTMs for both Chai-1 and AlphaFold3 are shown for each structure, where ipTM <0.6 indicates failed predictions and ipTM >0.8 indicates high confidence predictions.

dictions, values below 0.6 indicate a failed prediction, and scores between 0.6 and 0.8 could be correct or wrong.

3.5.3 PLM-interact can identify mutational impact in human PPIs

Here, we demonstrate examples of PLM-interact correctly predicting the consequences of mutations on PPIs. Canonical protein sequences (ie proteins without mutations) are retrieved from UniProt (The UniProt Consortium 2009). Amino acid substitutions associated with changes in PPIs are obtained from InAct (Kerrien et al. 2012). We obtained predicted interaction probabilities for canonical and mutants from PLM-interact trained on human data (see Methods). For mutations associated with interaction gain 'mutation-causing' interactions (see Figure 3.8 A), PLM-interact predicted interaction probabilities of the negative pair are below 0.5, whereas the interaction probabilities of the mutant PPI exceed 0.5. For mutation-disrupting PPIs (Figure 3.8 B), the predicted interaction probabilities of the positive PPI exceed 0.5, whereas the interaction probabilities of the mutant PPI are below 0.5. Examples are presented in Figure 3.8, where complex structures are predicted by Chai-1 (Discovery et al. 2024).

We show two examples of mutations that cause PPIs in Figure 3.8 A, the protein Ataxin-1 and SOD1 (Superoxide dismutase [Cu-Zn]), which in humans are encoded by the ATXN1 and SOD1 genes separately. Ataxin-1 interacts with many other proteins and the expansion of a glutamine(Q)-encoding repeat can affect the function of PPIs and cause the genetic disease spinocerebellar ataxia type1 (SCA1) and other polyglutamine diseases (Lim et al. 2008). Predictions from PLM-interact show that prior to mutation, this PPI has a predicted interaction probability of 0.411, correctly indicating non-interaction with E2s (Ubiquitin-conjugating enzyme E2 E3). Following mutation, PLM-interact increases this score to 0.771, correctly predicting that the mutation induces interaction. In the second example (Figure 3.8 A), the SOD1 gene encodes superoxide dismutase enzymes that break down human superoxide radicals. SOD1 is linked to the nervous system disease amyotrophic lateral sclerosis (ALS) (Rosen et al. 1993), with the A4V mutation being the most common variant in North America (Wj et al. 2008). PLM-interact predicts 0.465 and 0.851 before and after the mutation, correctly capturing the change in interaction with CDK4 (Cyclin-dependent kinase 4).

Next, we show two examples of mutations that disrupt PPIs (Figure 3.8 B). First, GRB2 (growth factor receptor bound protein 2) is associated with signal transduction. GRB2 mutations are associated with multiple cancers, including breast cancer (Daly et al. 1994) and leukaemia (Ohanian et al. 2018). The canonical protein interacts with PLEKHS1 (Pleckstrin homology domain-containing family S member 1). PLM-interact predicts that the missense mutation R86K reduces the interaction probability from 0.503 to 0.477. Second, CLCP1 is a transmembrane protein that

regulates cell growth and this protein is identified as a cancer marker, exhibiting up-regulated expression in lung cancer (Koshikawa et al. 2002). Again, PLM-interact predicts that the missense mutation Y732F reduces the interaction probability from 0.873 to 0.425.

3.6 Discussion

In this chapter, we have developed a novel PPI prediction method, PLM-interact, that extends single protein-focused PLMs to their interacting protein pairs. We have conducted a benchmarking study, evaluating PLM-interact against five distinct PPI prediction approaches: TT3D (Sledzieski et al. 2023), Topsy-Turvy (Singh et al. 2022), D-SCRIPT (Sledzieski et al. 2021), PIPR (Chen et al. 2019) and DeepPPI (Richoux et al. 2019). We demonstrate PLM-interact's performance in a held-out species PPI prediction task, showing a significant improvement over the state-of-the-art prediction approaches (Figure 3.4 B). We further highlight five positive PPI examples where PLM-interact produces a correct prediction, while the next best performer TT3D produces an incorrect prediction (Figure 3.6 and 3.7). Additionally, we show successful examples of predicting mutational effects on human protein interactions by investigating changes in predicted interaction probabilities before and after the mutations (Figure 3.8 and 3.9).

We further observe that positive protein pairs achieve the largest number of correct predictions (interaction probabilities > 0.5) compared to other PPI approaches, as seen in the distribution of predicted interaction probabilities (Figure 3.5). Underlying the benefit of PLM-interact is the improved capability of correctly predicting positive PPIs in the held-out species. However, positive PPIs are challenging to predict due to a lack of high-quality PPI data for training. TT3D (Sledzieski et al. 2023) includes explicit structural information, the per-residual structural alphabet in FoldSeek (Van Kempen et al. 2024), to improve its prediction over Topsy-Turvy (Singh et al. 2022), which incorporated network data. Our improved performance relative to TT3D is particularly impressive given that we only use sequences.

The PLM ESM-2 is trained on large-scale proteins and has been successfully used for several downstream tasks, including protein folding (ESMFold (Lin et al. 2023)) and long-range contact prediction (Ames et al. 2016) etc. Unlike previous PPI prediction approaches that extract pre-trained PLM embeddings for each protein individually but ignore interaction-aware patterns of amino acids in each input protein pair, PLM-interact retrains PLM ESM-2 with both masking and classification losses to optimize the model. We observed that the ratio of 1:10 is the optimal choice for 650M ESM-2 (Figure 3.2). We train the PLM-Interact model on a benchmark human PPI dataset and a larger human PPI dataset we created, sourced from STRING V12. The STRING V12 includes larger training datasets, enabling better generalization in human PPI

prediction.

Furthermore, our results show the potential of predicting mutational effects on PPIs from sequence alone. This could lead to a new generation of interaction-aware in-silico variant effect predictors where methods rely on PLMs of the single proteins (Rives et al. 2021; Frazer et al. 2021; Cheng et al. 2023). However, current training data remains limited. The number of high-quality structures of mutant proteins and their interaction partners are low. Algorithmically, models with long and multimodal context (Hayes et al. 2024; Cornman et al. 2024; Shao et al. 2024) that include multiple proteins, structures and nucleotides could be specialised for interaction tasks.

Based on our results, we demonstrate that retraining ESM-2 with paired proteins allows the model to become 'aware' of interaction patterns. PLM-interact expands the scope of PPI prediction within species and can analyse the mutational effects of human PPIs. Despite the current limitations of PLM-Interact in predicting PPIs across all species, its strong performance on five held-out species PPI prediction highlights its important potential to extend to more species. We expect that PLM-interact can be further extended to handle downstream tasks and fine-tune for specific tasks. For instance, PLM-interact can extract protein representations, encoding protein properties and interactions, which can be used for specific tasks such as host prediction. We anticipate that our model can be harnessed for cross-species PPI prediction and the construction of PPI networks in future research. The PLM-interact model, trained on human PPIs, can be adapted to predict PPIs of other species by fine-tuning additional datasets.

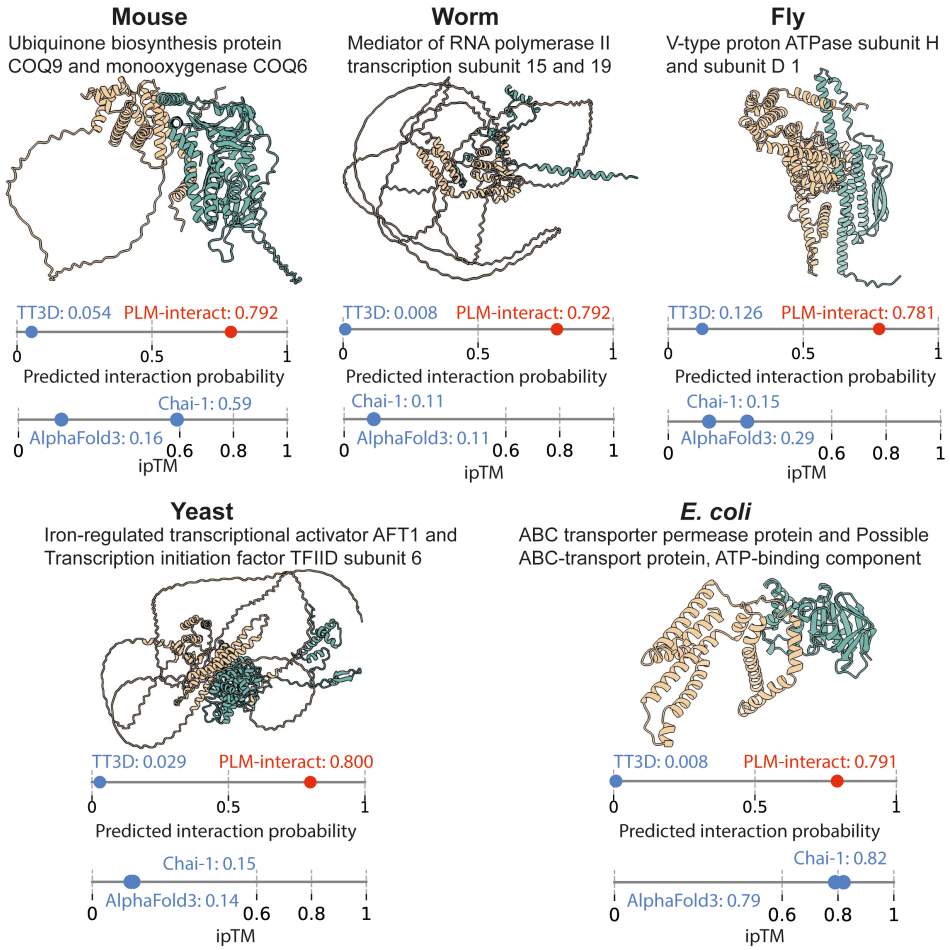

Figure 3.7. PPI example for each species that was predicted correctly by PLM-interact but not by TT3D. Protein-protein structures are predicted by AlphaFold3 (Abramson et al. 2024) and visualised with ChimeraX (Pettersen et al. 2021). Both models' prediction interaction probabilities range between 0 and 1. A predicted interaction probability >0.5, is predicted as a positive PPI, while <0.5 is a negative pair. Interacting proteins are shown from left (yellow) to right (green). For information about these PPIs, see Figure 3.6

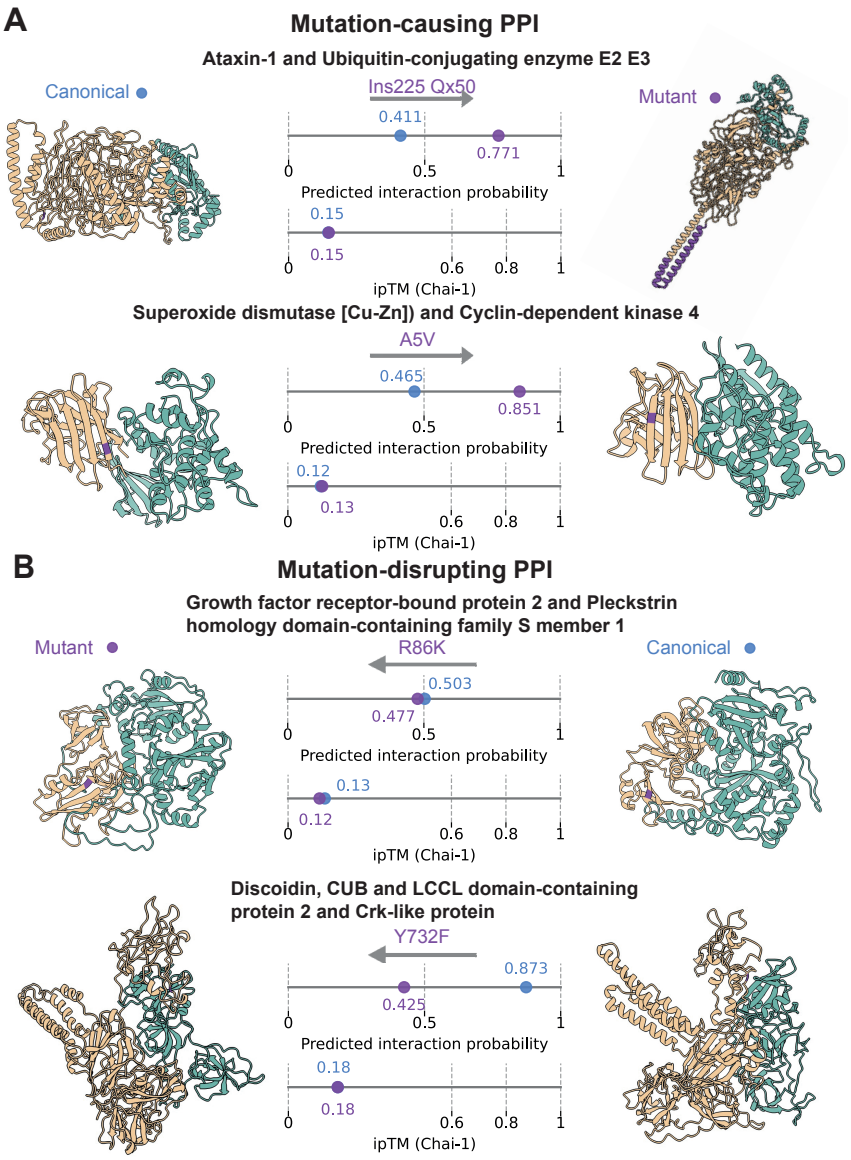

Figure 3.8. Demonstration of PLM-interact detecting changes in human PPIs associated with mutations. **A** shows two mutation-causing interaction examples, while **B** shows two mutation-disrupting PPI examples. These PPI structures are predicted using Chai-1 (Discovery et al. 2024) and visualised with ChimeraX (Pettersen et al. 2021); here, the mutated amino acids are highlighted in purple. Prediction interaction probabilities exceeding 0.5 indicate the proteins interact, while below 0.5 indicate non-interact. Chai-1's ipTM scores give the structure prediction confidence where <0.6 indicates failed predictions. Interacting protein structures are shown from left (yellow) to right (green): **A**. Residue 225 Glutamine (Q) of P54253 (Ataxin-1) is mutated to 50 Q, causing interaction with Q969T4 (Ubiquitin-conjugating enzyme E2 E3) (C et al. 2020); Residue 5 Alanine (A) of P00441 (Superoxide dismutase [Cu-Zn]) is mutated to Valine (V), causing interaction with P11802 (Cyclin-dependent kinase 4) (Kabuta et al. 2013). **B**. Residue 86 Arginine(R) of P62993 (Growth factor receptor-bound protein 2) is mutated to Lysine (K), disrupting its interaction with Q5SXH7-1 (Pleckstrin homology domain-containing family S member 1) (Grossmann et al. 2015); Residue 732 Tyrosine (Y) of Q96PD2 (Discoidin, CUB and LCCL domain-containing protein 2) is mutated to Phenylalanine (F), disrupting its interaction with P46109 (Crk-like protein) (Aten et al. 2013). See Figure 3.9 for AlphaFold3 (Abramson et al. 2024) predicted structures.

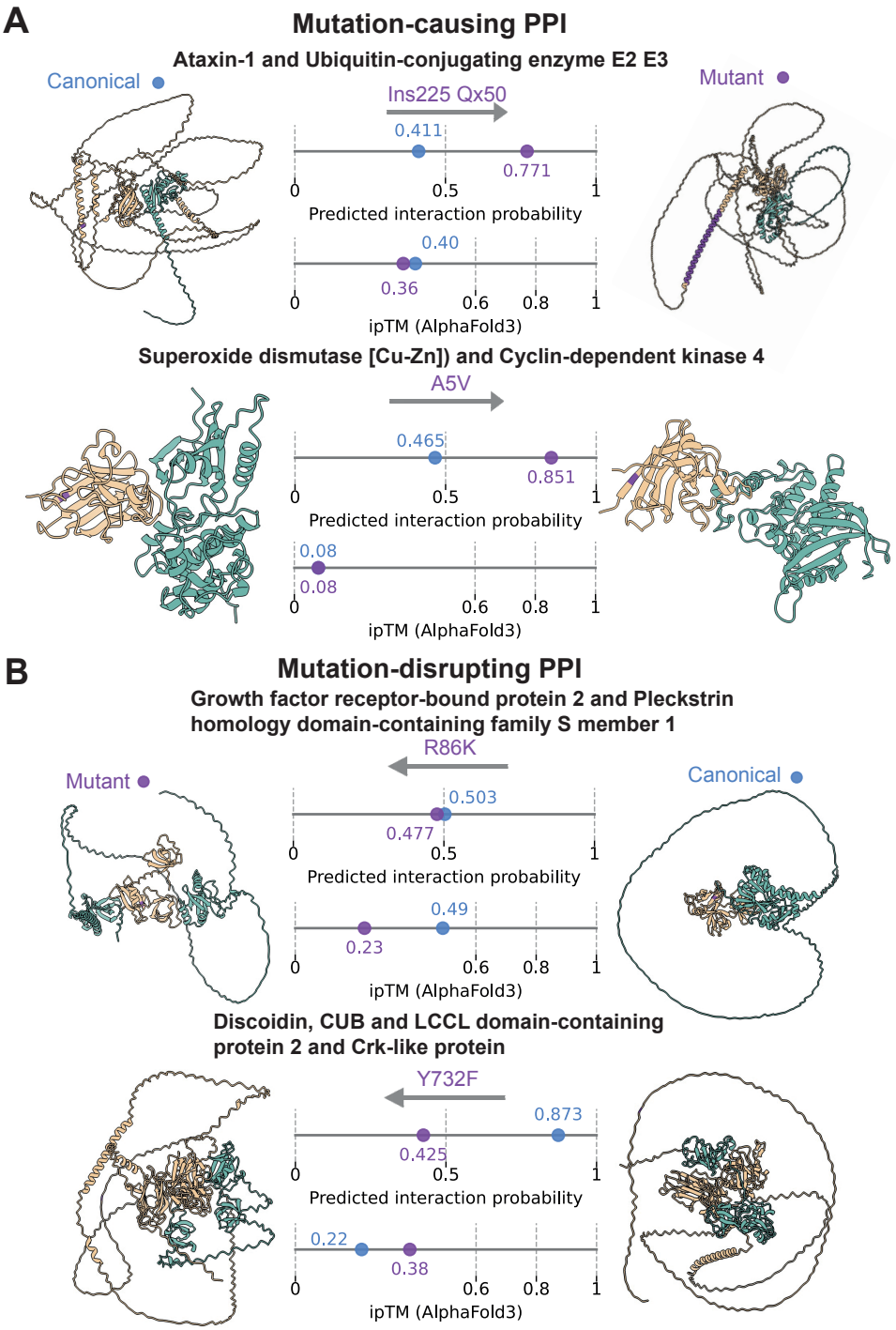

Figure 3.9. Demonstration of PLM-interact detecting changes in human PPIs associated with mutations. **A** shows two mutation-causing interaction examples, while **B** shows two mutation-disrupting PPI examples. These PPI structures are predicted using AlphaFold3 (Abramson et al. 2024) and visualised with ChimeraX (Pettersen et al. 2021); the mutated amino acids are highlighted in purple. Prediction interaction probabilities exceeding 0.5 indicate the proteins interact, while below 0.5 indicate non-interact. AlphaFold3's ipTM scores give the structure prediction confidence, where <0.6 indicates failed predictions. Interacting protein structures are shown from left (yellow) to right (green). See Figure 3.8 for information about these protein pairs.

Chapter 4

Virus-human protein-protein interaction prediction

If I have seen further, it is by standing on the shoulders of Giants.

Isaac Newton

4.1 Abstract

Viral infections frequently cause human disease. Determining virus-human interaction mechanisms and discovering the interface of interacting proteins helps develop antiviral drugs and vaccines. However, most virus-human protein-protein interactions (PPIs) are unknown. Existing state-of-the-art virus-human PPI models applied protein features extracted from pre-trained protein language models (PLMs), ignoring the contextualised inter-molecular amino acid interactions within the entire protein pair. We trained a language model PLM-interact-VH, jointly encoding paired virus and human proteins, to learn their protein interactions (see Chapter 3). Our results demonstrated that PLM-interact-VH effectively captures predictive signals of virus-human PPIs, achieving 5-23% improvements in AUPR compared to state-of-the-art models. To evaluate the model's performance across varied levels of prediction difficulty, we constructed a virus-human PPI dataset and split it into training, validation and testing sets using three strategies: random splitting, Park and Marcotte's C1/C2/C3 and hold-out virus families. We demonstrated PLM-interact-VH's robust performance on these data-splitting datasets, except for hold-out virus families. We further show that fine-tuning the human PPI model on virus-human PPI

datasets improves the virus-human PPI prediction. Finally, we show that training with both intra-species and inter-species PPIs has the potential to develop a generalizable PPI model.

4.2 Introduction

In virology, PPIs are particularly important as viruses depend entirely on the host cell for replication, achieved mainly through specific interactions with host molecules (Jangra et al. 2022). Hundreds of protein interactions are involved in this viral life cycle. Viruses bind with host receptors to enter host cells, and the mutation of virus binding sites might lead to disruption with host receptors. Identifying virus-human PPIs has the potential to develop novel therapeutic interventions.

PPIs in human, yeast and fly are well-studied. However, PPIs from most other species remain sparse (Kotlyar et al. 2022), particularly virus-host PPIs. Virus-human PPIs in public databases are limited (Brito et al. 2017). With the development of sequencing technologies, many novel viruses have been discovered, but the hosts of these viruses and virus-host PPIs are typically unknown. Experimental approaches such as the yeast two-hybrid system(Y2H) (Suter et al. 2008) and co-immunoprecipitation (Co-IP) (Iqbal et al. 2018) are often used to identify PPIs. However, these approaches are time-consuming and labour-intensive, making identifying PPIs on a large scale challenging.

Computational algorithms offer an alternative way to predict virus-human PPIs at scale. Conventional prediction approaches mainly leverage protein features including domains (Zhang et al. 2017), structures (Lasso et al. 2019) and evolutionary alignments (Hamp et al. 2015a). Additionally, NLP algorithms word2vec (Church 2024) and doc2vec (Le et al. 2014) have been used for encoding the composition information of protein sequences. These encoded features are then used to train virus-human PPI classifiers based on classical machine learning such as random forest (Yang et al. 2020) and deep learning approaches such as LSTM (Tsukiyama et al. 2021). InterSPPI (Yang et al. 2020) used a distributed-memory (DM) model architecture to train a doc2vec model with k-mer residue segments and protein sequences for protein representations, which are then used to train a random forest for virus-human PPI prediction. LSTM-PHV (Tsukiyama et al. 2021) trained a word2vec to learn k-mer patterns of protein sequences, which are then used to train an LSTM (Hochreiter 1997) to predict virus-human PPIs. Both sequence embedding methods, word2vec and doc2vec, aim to encode the textual information of sequences into dense representations of protein sequences.

Recently, protein language models (PLMs) trained on a large scale of protein sequence databases have been deployed to encode structural features, protein properties, sequence composition and

evolutionary information, which has emerged as the approach for representing proteins in virus-human PPI predictors. The state-of-the-art PPI prediction model STEP (Madan et al. 2022) used a pre-trained PLM ProtBERT (Elnaggar et al. 2021) to extract protein embeddings and then trained on a Siamese neural network for the PPI binary classification. The architecture of the STEP exploits the PLM to independently represent each protein in a protein pair, followed by a classification layer to discriminate between interacting and non-interacting pairs. This model used the pre-trained PLM to extract a protein embedding for each protein, ignoring inter-protein contacts in the entire protein pairs. PLMs are primarily trained on single proteins and designed to identify contact points within a single protein, failing to learn interaction patterns in a protein pair. Another issue is the high PPI protein sequence identity between training and test datasets. The latest benchmarking virus-human PPI dataset removes high-identity PPIs at an identity of 95%, much higher than the 40% used in existing intra-species PPI prediction tasks (Sledzieski et al. 2021; Liu et al. 2024b).

To improve performance, we trained the deep learning model PLM-interact-VH (introduced in Chapter 3) to predict virus-human PPIs. PLM-interact-VH fine-tuned ESM-2 by encoding paired proteins and learned interaction patterns within each paired protein sequence. PLM-interact-VH significantly improves performance compared to other predictors in the virus-human PPI benchmarking task. To ensure no high-identity protein pairs between training and test sets, we constructed a virus-human PPI dataset by removing protein pairs with an identity of 40%. Virus and human proteins are clustered with an identity of 40% and then filter protein pairs using the representative protein in each cluster, which aims to keep all virus-human protein pairs sharing an identity of no more than 40%. We further evaluated the model's performance and robustness on different groups of PPI datasets split with three strategies. We discovered that the prediction task becomes harder when we test on hold-out virus families, as virus-human PPI mechanisms are divergent across different virus families. We demonstrated that fine-tuning human-human PPI models on virus-human PPI datasets improves the performance of models that are only trained with virus-human PPI data.

4.3 Methods

4.3.1 Datasets

4.3.1.1 The virus-human benchmarking PPI dataset.

The benchmarking dataset of 22,383 virus-human PPIs includes 5,882 human and 996 virus proteins. This dataset was obtained from Tsukiyama et al. (2021), sourced from the HPIDB 3.0

database (Ammari et al. 2016); the ratio of positive to negative pairs is 1:10 and negative pairs are chosen based on sequence dissimilarities. The length of protein sequences ranges from 30 to 1,000 and the high-identity PPIs are filtered based on a threshold of 95% identity using CD-HIT (Szklarczyk et al. 2023).

4.3.1.2 Integration of virus-human PPIs from multiple datasets

The benchmarking virus-human PPI dataset was sourced from HPIDB 3.0 (Ammari et al. 2016), which integrated multiple virus-human PPI databases, including IntAct (Kerrien et al. 2012) and VirHostNet 3.0 (Guirimand et al. 2015). Recently, new versions of virus-human PPI databases have been released and more virus-host PPIs have been integrated. Especially during the pandemic, a large amount of SARS-CoV-2-related data has been collected in these PPI databases. In this chapter, we integrate the seven latest versions of PPI databases to create a bigger virus-human PPI dataset. This expanded dataset includes positive interaction proteins from IntAct (Kerrien et al. 2012), VirHostNet3.0 (Guirimand et al. 2015), BioGRID (Oughtred et al. 2019), HPIDB 3.0 (Ammari et al. 2016), MINT (Zanzoni et al. 2002), EBI-GOA-nonIntAct (Orchard et al. 2014) and HVIDB (Yang et al. 2021). To ensure high-quality interactions, we focus on three types of interactions: physical association, association, and direct interaction.

1. **IntAct** (http://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact-micluster.zip, 245, October 2023 release), protein-protein interaction database. There are 805,932 interactions between 143,510 interactors. We extract 28,455 virus-host PPIs between 1,165 virus proteins and 9,035 human proteins, with 27,647 virus-human PPIs between 8,382 virus proteins and 1,095 human proteins.

2. **VirHostNet 3.0** (https://virhostnet.prabi.fr/, March 2022 release), virus-host protein-protein interaction database. This release includes 55,115 manually curated PPIs for both virus-host and virus-virus. We extracted 37,948 virus-host PPIs between 973 virus proteins and 8,581 host proteins, with 37,566 virus-human PPIs between 968 virus proteins and 8,199 human proteins.

3. **BioGRID** (https://downloads.thebiogrid.org/File/BioGRID/Release-Archive/BIOGRID-4.4.235/BIOGRID-MV-Physical-4.4.235.mitab.zip, 25th June 2024 release), protein-protein interaction database. This database includes 1,513,281 PPIs. We extracted 6,662 virus-host PPIs between 111 virus proteins and 3,008 host proteins, with 6,655 virus-human PPIs between 111 virus proteins and 3,001 human proteins.

4. **Host-Pathogen Interaction Database 3.0** (HPIDB 3.0) (https://cales.arizona.edu/hpidb/, accessed on 26 September 2024), virus-host protein-protein interaction database. HPIDB

includes 55,505 PPIs between 55 hosts and 523 pathogen species in release 3.0. The top 3 ranked virus species are influenza, Herpes viruses and Saccharomyces cerevisiae. We extracted 32,030 virus-host PPIs between 1,240 virus proteins and 7,864 human proteins, with 30,959 virus-human PPIs between 1,181 virus proteins and 6,968 human proteins.

5. **MINT** (https://mint.bio.uniroma2.it/index.php/sample-page/, accessed on 26 September 2024), protein-protein interaction database. MINT collected experimentally verified PPIs from literatures. We extracted 8,675 virus-host PPIs between 411 virus proteins and 3,153 host proteins, with 8,505 virus-human PPIs between 391 virus proteins and 3,003 human proteins.

6. **EBI-GOA-nonIntAct** (https://www.ucl.ac.uk/cardiovascular/research/pre-clinical-and-fundamental-science/functional-gene-annotation/psicquic?conversationContext=2, July 2019 release), protein-protein interaction database. This database includes 84,087 PPIs created by the GO consortium and removes the interactions that overlap with the IntAct. We extracted 527 virus-host PPIs between 154 virus proteins and 451 host proteins, with 396 virus-human PPIs between 126 virus proteins and 336 human proteins.

7. **HVIDB** (http://zzdlab.com/hvidb/, accessed on 26 September 2024), virus-human protein-protein interaction database. This database includes 48,643 virus-human protein-protein interactions. Here, we extract 4,8026 virus-human PPIs between 1,928 virus proteins and 7,759 human proteins.

Here, we combined virus-human PPIs from the seven public PPI databases outlined above. To visualize the distribution of virus families in this integrated virus-human PPI dataset, we plot a pie chart for virus families in the dataset. The proportion of the top 8 ranked virus families within this dataset is shown in Figure 4.1 A. Coronaviridae, Retroviridae and Orthomyxoviridae are the top 3 virus families. Coronaviridae has the largest percentage at 19.4%, while Retroviridae and Orthomyxoviridae have roughly 16.5%. We plot the pie chart of virus species for these three virus families receptively (Figure 4.1 B). SARS-related coronavirus, HIV and Influenza A virus are the most significant species (over 80%) in Coronaviridae, Retroviridae and Orthomyxoviridae, respectively. This indicates that each of these three virus families is dominated by one species. All species of these virus families are shown in Table 4.1.

Table 4.1. The table that shows species names of three hold-out virus families in Section 4.3.2.3. The columns are virus family, virus species and the number of PPIs of each virus species.

Family	Species	PPI Count
Coronaviridae	Severe acute respiratory syndrome-related coronavirus	11727
Coronaviridae	Middle East respiratory syndrome-related coronavirus	670
Coronaviridae	Human coronavirus HKU1	578
Coronaviridae	Human coronavirus NL63	438
Coronaviridae	Human coronavirus 229E	420
Coronaviridae	Betacoronavirus 1	310
Coronaviridae	Pipistrellus bat coronavirus HKU5	44
Coronaviridae	Murine coronavirus	39
Coronaviridae	Alphacoronavirus 1	27
Coronaviridae	Rousettus bat coronavirus HKU9	17
Coronaviridae	Avian coronavirus	5
Coronaviridae	Pangolin coronavirus	4
Coronaviridae	Bat coronavirus BM48-31/BGR/2008	2
Coronaviridae	Rhinolophus bat coronavirus HKU2	1
Coronaviridae	Tylonycteris bat coronavirus HKU4	1
Retroviridae	Human immunodeficiency virus 1	11489
Retroviridae	Primate T-lymphotropic virus 1	213
Retroviridae	Primate T-lymphotropic virus 2	77
Retroviridae	Human immunodeficiency virus 2	45
Retroviridae	Simian immunodeficiency virus	37
Retroviridae	Rous sarcoma virus	23
Retroviridae	Bovine leukemia virus	20
Retroviridae	Murine leukemia virus	12
Retroviridae	Simian foamy virus	7
Retroviridae	Walleye dermal sarcoma virus	6
Retroviridae	Mason-Pfizer monkey virus	5
Retroviridae	Equine infectious anemia virus	5
Retroviridae	Primate T-lymphotropic virus 3	3
Retroviridae	Feline immunodeficiency virus	2
Retroviridae	Visna-maedi virus	2
Retroviridae	Caprine arthritis encephalitis virus	2
Retroviridae	Avian myeloblastosis virus	2
Retroviridae	Jaagsiekte sheep retrovirus	2

Continued on next page

Table 4.1 – continued from previous page

Family	Species	PPI Count
Retroviridae	UR2 sarcoma virus	1
Retroviridae	Abelson murine leukemia virus	1
Retroviridae	Avian leukosis virus	1
Retroviridae	Bovine immunodeficiency virus	1
Retroviridae	Y73 sarcoma virus	1
Orthomyxoviridae	Alphainfluenzavirus influenzae	12046
Orthomyxoviridae	Betainfluenzavirus influenzae	14
Orthomyxoviridae	Thogotovirus thogotoense	5
Orthomyxoviridae	Thogotovirus dhoriense	3
Orthomyxoviridae	Gammainfluenzavirus influenzae	1

These PPI databases contain overlaps. Therefore, we combined the datasets and removed redundant PPIs, obtaining 74,178 PPIs. The proteins with a length longer than 2000 were excluded, resulting in a total of 62,250 PPIs. To identify and remove high-identity PPIs, we cluster proteins with a threshold of 40% identity using MMseq2 (Steinegger et al. 2017). Protein pairs (A, B) and (C, D) are identified as high-identity PPIs when proteins A and C belong to the same cluster, as well as B and D in the same cluster. This clustering method removes high-identity PPIs and ensures less than 40% identity between each pair, resulting in 31,466 virus-human PPIs (Figure 4.4). Figure 4.2 shows the Venn diagram of our constructed PPI dataset and the benchmarking PPI dataset in Section 4.3.1.1. The result shows that our dataset contains 16,763 new PPIs and has 14,703 PPI overlaps with the benchmarking dataset. To visualize the database source of PPIs in our dataset, we show the UpSet plot for these seven public datasets above (Figure 4.3). Each database has specific PPIs, and there are no PPIs that exist in all seven databases. IntAct and VirHostNet have over 3000 PPIs that do not exist in other databases; combining these databases can enhance our PPI training data.

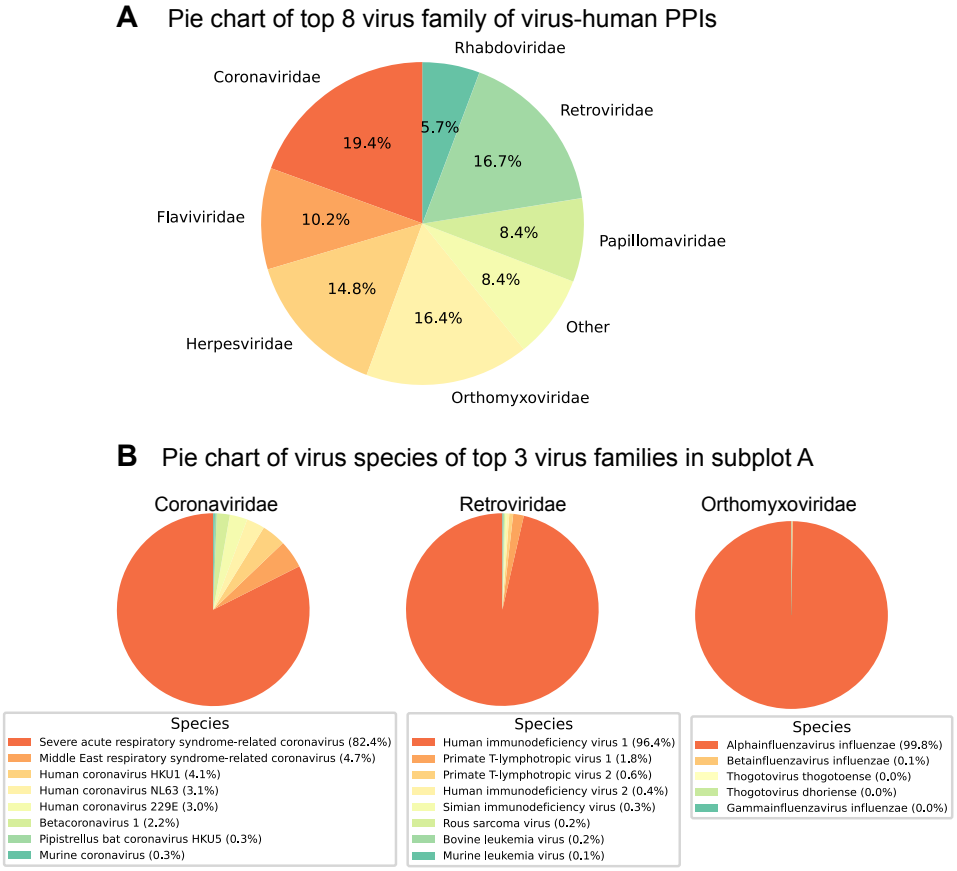

Figure 4.1. The distribution of virus families and species in the virus-human PPIs. Panel A shows the percentage of the top 8 ranked virus families in our collected virus-human PPI dataset. The top 3 ranked virus families are Coronaviridae, Retroviridae and Orthomyxoviridae. Coronaviridae has the largest percentage at 19.4%, while Retroviridae and Orthomyxoviridae have roughly 16.5%. Panel B shows the top 8 virus species of these three virus families. SARS-related coronavirus, HIV, and Influenza A virus dominated the three virus families, respectively. It indicates that the PPIs of each virus family mainly belong to one virus species. Table 4.1 shows all virus species from these three virus families.

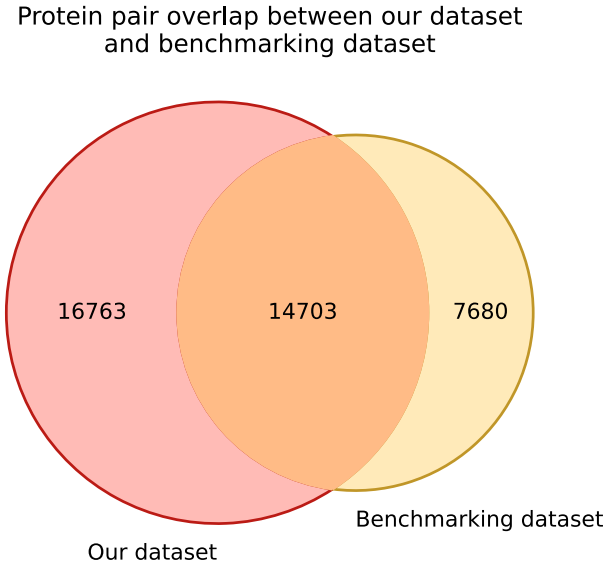

Figure 4.2. The Venn Diagram of virus-human PPIs dataset from our and Tsukiyama et al. (2021). It shows the overlaps and differences between our constructed PPI dataset and Tsukiyama et al.'s benchmarking PPI dataset.

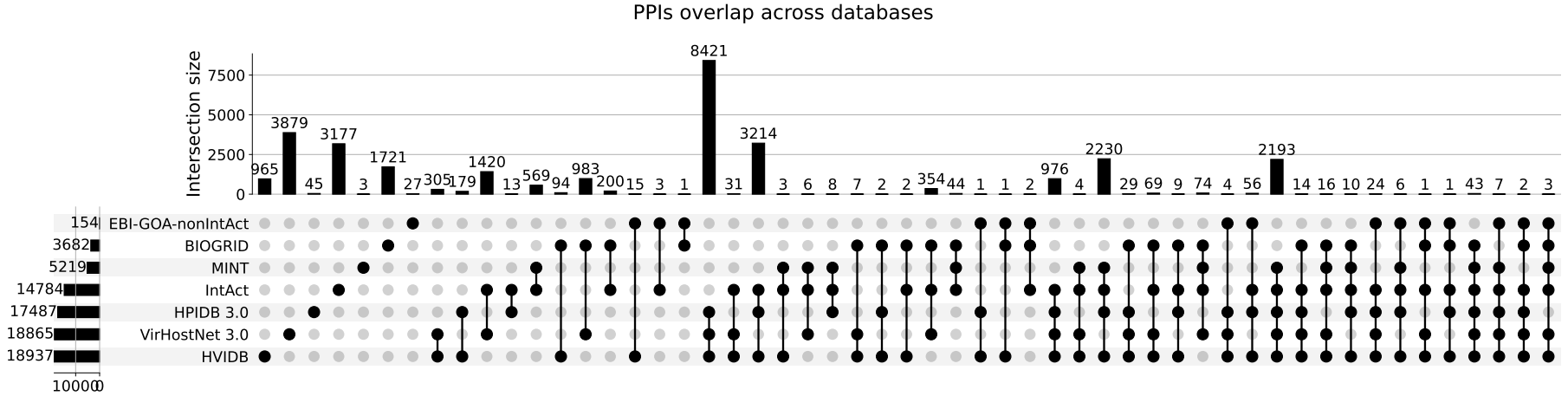

Figure 4.3. The UpSet plot of seven public virus-human PPI databases that are used to construct the virus-human PPI dataset. This Upset diagram shows the overlaps and differences of seven virus-human PPI databases in our constructed 31,466 virus-human PPIs.

The negative protein pairs are randomly chosen from any paired proteins that are not reported as positive protein pairs. The virus-human PPI network is sparse, so randomly sampling negative protein pairs minimizes the likelihood of false negative (Park et al. 2011). The ratio between positive and negative is 1:10, following the previous PPI prediction models (Sledzieski et al. 2021). To optimize GPU storage and training efficiency, we set the maximum protein length to 2000 and the maximum protein pair length to 3000. The threshold for the protein pair length was determined based on the 98th and 95th percentiles of the length distribution of all protein pairs. If the 98th percentile exceeded 3000, the 95th percentile was the threshold. If the 95th percentile also exceeded 3000, the threshold was set to 3000. This constructed dataset is the standard virus-human PPI dataset for analysing three data-splitting strategies: 1) Randomly split; 2) Park and Marcotte's, C1, C2 and C3; 3) Hold out the virus family. For details about each split strategy, see Section 4.3.2.

4.3.2 Splitting strategies for virus-human PPI datasets

We employ three strategies to split the constructed virus-human PPI dataset into training, validation and test sets. Each strategy results in different data sizes and prediction outcomes. In this section, the training, validation and test sets are split into 80%:10%:10%, a ratio commonly used in machine learning tasks. Note: As mentioned above, we randomly select negative protein pairs with a positive-to-negative ratio of 1:10. Next, we filter protein pairs based on their combined length using the threshold setting method described in Section 4.3.1.2. This removes varying numbers of protein pairs from each set. Therefore, after filtering based on the combined length, the positive-to-negative ratio of each dataset is not exactly 1:10 but remains close to it.

4.3.2.1 Random splitting

Firstly, we randomly split the constructed standard dataset into 80% training, 10% validation and 10% test protein pairs. In this case, we obtained 23,707 positive protein pairs in the training dataset, as well as 29,70 and 2,968 positive protein pairs in the validation and testing datasets, respectively (see Figure 4.5).

4.3.2.2 Park and Marcotte's C1, C2 and C3

The model's performance might be overestimated by random splitting, because the training dataset probably includes similar patterns to the test dataset (Park et al. 2012; Hamp et al. 2015b). Park et al. (2012) introduced three PPI classes with varying prediction difficulty: 1)

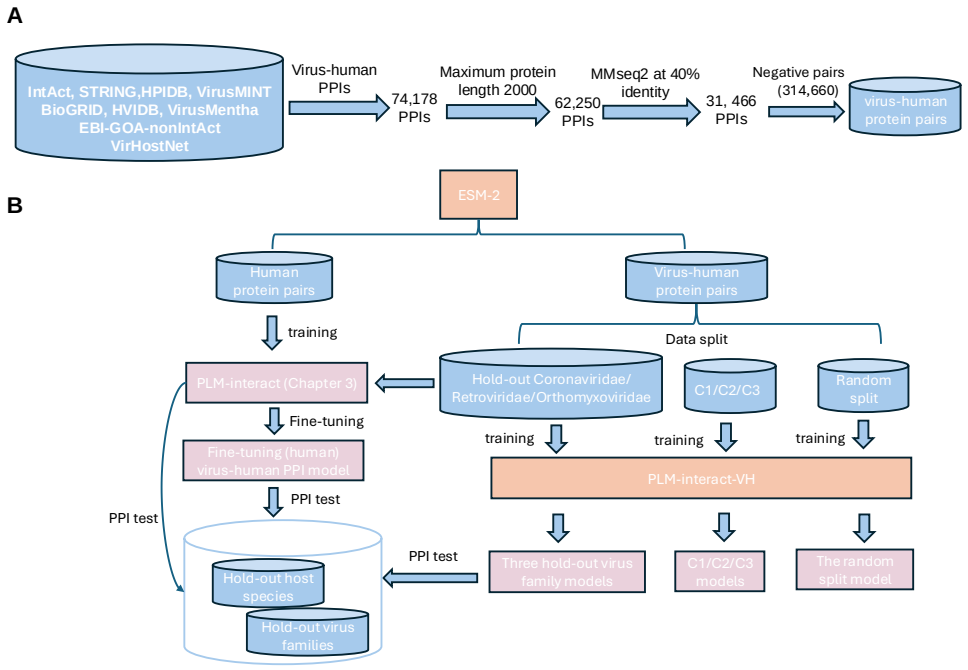

Figure 4.4. The preprocessing steps for the integrated virus-human protein interaction dataset (A); data-splitting and training steps of our dataset (see Section 4.3.2, 4.4.2 and 4.4.4) (B). Virus-human PPIs are extracted from IntAct (Kerrien et al. 2012), VirHostNet3.0 (Guirimand et al. 2015), BioGRID (Oughtred et al. 2019), HPIDB 3.0 (Ammari et al. 2016), MINT(Zanzoni et al. 2002), EBI-GOA-nonIntAct (Orchard et al. 2014) and HVIDB (Yang et al. 2021). The maximum length of protein sequences is 2000. MMSeq2 (Steinegger et al. 2017) is used to cluster proteins, and the high-identity protein pairs with 40% protein similarities are removed. After filtering, we obtained 31,466 virus-human PPIs for downstream training tasks. Next, we combined these positive protein pairs with randomly selected negative PPIs to create a virus-human PPI dataset. Finally, we use three data-splitting strategies to construct training, validation and test datasets for the evaluation of PLM-interact-VH. Details about PLM-interact-VH training, fine-tuning and testing are described in Section 4.4.2 and 4.4.4.

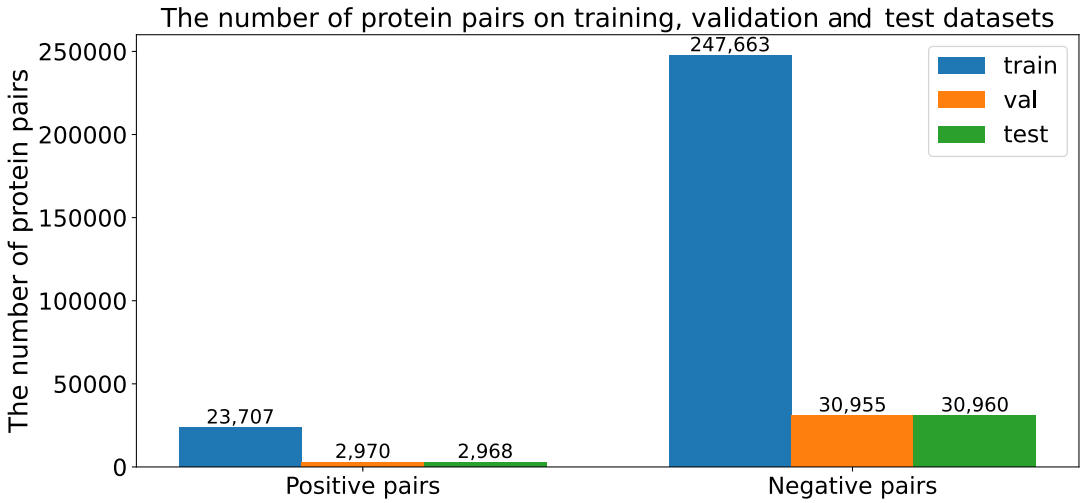

Figure 4.5. The bar plot shows the number of positive and negative protein pairs on training, validation and test datasets obtained by the randomly split strategy.

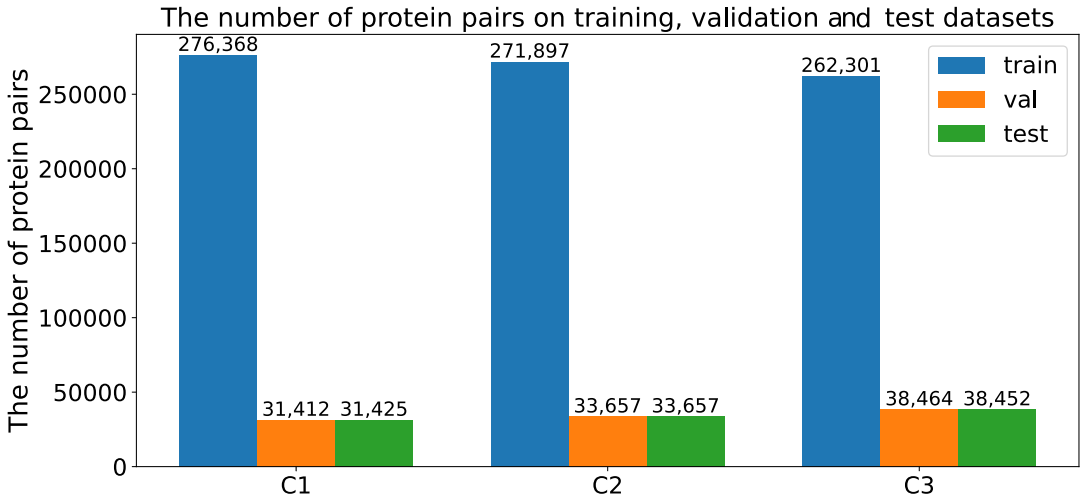

Figure 4.6. The bar plot shows the number of protein pairs on training, validation and testing split by Park and Marcotte's C1, C2 and C3 (Park et al. 2012), separately.

C1: both proteins in the test pair exist in the training protein pairs; 2) C2: only one protein in a test pair exists in the training protein pairs; 3) C3: no protein overlap exists between training and test protein pairs. We use this split strategy to evaluate the model's performance with different PPI conditions. Here, we constructed the C1, C2 and C3 datasets, respectively. The training protein pairs of C1 and C2 are over 270K, while C3 includes approximately 260K protein pairs. Validation and test sets of both C1 and C2 consist of no more than 34K protein pairs, while C3 contains over 38K protein pairs (see Figure 4.6).

4.3.2.3 Hold-out virus family

As shown in Figure 4.1, the top 3 ranked virus families are Coronaviridae, Retroviridae and Orthomyxoviridae in the constructed dataset. We hold out each of these three virus families as our validation (50%) and test sets (50%), and the rest of the PPIs are the training sets. The number of training, validation and test sets on each hold-out virus family is shown in Figure 4.7. The Retroviridae family has the most significant number of training protein pairs at over 317K, while Coronaviridae and Orthomyxoviridae have roughly 272K and 266K, respectively. The test and validation set of hold-out Coronaviridae and Orthomyxoviridae have more test protein pairs with roughly 33K and 36K, respectively, while Retroviridae has around 10K.

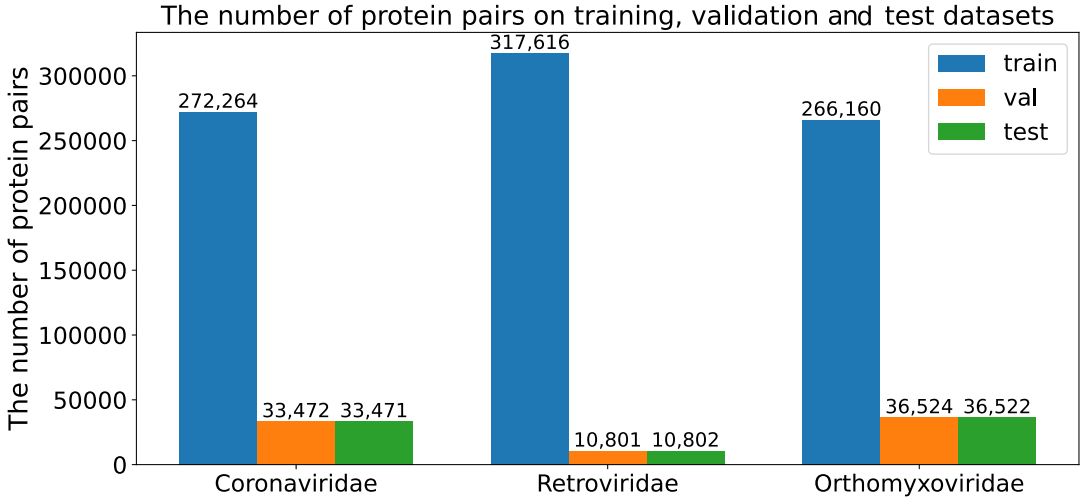

Figure 4.7. The bar plot shows the number of training, validation and test sets obtained by holding out three virus families separately. The training data size for Coronaviridae and Retroviridae is around 272K and 266K, respectively, while Retroviridae is larger, at approximately 317K.

4.3.3 Model architecture

The model architecture of PLM-interact-VH is the same as PLM-interact and details are described in Section 3.4.2. For details about this model's training and evaluation, see the method Section 3.4. Figure 3.1 shows the PLM-interact model architecture, which trains ESM-2 as a cross-encoder based on masking and binary classification loss. Cross-encoder is a type of neural network architecture (Reimers et al. 2019), which is successfully implemented in various NLP tasks such as semantic textual similarity (Rep et al. 2023) and information retrieval (Reimers et al. 2019). The cross-encoder architecture simultaneously processes both input texts, allowing for an understanding of semantic content across both virus and human proteins. The self-attention mechanism of ESM-2 enables PLM-interact-VH to maintain long-range information throughout the entire virus-human protein sequence pair. For the input protein $Virus_{P_1}$ and $Human_{P_2}$, the cross-encoder processes them as:

$$P_{pair} = [CLS, Virus_{P_1}, EOS, Human_{P_2}, EOS] \tag{4.1}$$

ESM-2 has been pre-trained on billions of protein sequences from UniRef (Suzek et al. 2015). We retrain ESM-2 based on the cross-encoder architecture to model virus-human PPIs as a pairwise comparison, treating protein pairs as analogous to sentence-sentence and question-answer formats in NLP tasks.

4.3.4 Model Training

We train PLM-interact-VH using an efficient batch size of 128 on the virus-human PPI benchmarking dataset and the constructed virus-human PPI datasets. The model is trained for ten epochs on the virus-human benchmarking data. To evaluate the model's performance on the validation data during training, evaluations are conducted every 2000 steps. In each evaluation, 5000 protein pairs are randomly subsampled separately from the validation and test datasets. Evaluating a subset of validation datasets enables us to monitor models' performance effectively while optimising both time and GPU resources. For details on the training parameter settings, see Section 3.4.3.

All the models are trained on the DiRAC Extreme Scaling GPU cluster Tursa. A typical 10-epoch training run of the model with ESM-2 (650M) trained on the virus-human benchmarking dataset takes 30.5 hours on 8 A100-80 GPUs. The virus-human PPI model trained on the benchmarking virus-human PPI dataset created by Tsukiyama et al. (2021), sourced from HPIDB 3.0 (Ammari et al. 2016).

4.3.5 Evaluation metrics

In this chapter, AUPR was used to evaluate the model's performance (see Section 3.4.9.1). Additionally, F1 and Matthews Correlation Coefficient (MCC) were employed to evaluate the model's performance on the virus-human PPI prediction benchmarking task. The formulas are shown below:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{4.2}$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4.3}$$

4.3.6 Data availability

The virus-human benchmarking PPI dataset from Tsukiyama et al. (2021) is available at: http://kurata35.bio.kyutech.ac.jp/LSTM-PHV/download_page

4.4 Results

4.4.1 Improved virus-human PPI prediction

To study virus-host PPI prediction, we retrain PLM-interact on a virus-human PPI dataset from Tsukiyama et al. (2021). The dataset is derived from the Host-Pathogen Interaction Database (HPIDB) 3.0 (Ammari et al. 2016), comprising a total of 22,383 PPIs, which include 5,882 human and 996 virus proteins. We compare our model with three recent virus-human PPI models: PLM-based approach STEP (Madan et al. 2022) and the protein embeddings-based approaches LSTM-PHV (Tsukiyama et al. 2021) and InterSPPI (Yang et al. 2020). STEP is similar to existing PPI models benchmarked previously in our study; it leverages protein sequence embedding extracted by the pre-trained PLM ProtBERT (Elnaggar et al. 2021). The results show that PLM-interact outperforms the other models. For the STEP comparison, this corresponds to improvements in AUPR, F1 and MCC scores of 5.7%, 10.9% and 11.9%, respectively (Figure 4.8 A). The length of virus proteins, human proteins and the combined length of virus-human PPIs are shown in Figure 4.8 B. To further analyse our model's performance, we select three pairs of virus-human PPIs from our test data, all with corresponding experimental virus-human complex structures available in the HVIDB (Yang et al. 2021). We then use ChimeraX (Pettersen et al. 2021) to visualise these structures and present PLM-interact's predicted interaction probability for each example (see Figure 4.8 C).

On this benchmarking dataset, all metric values are higher than 0.9 or nearly 0.9, except for Yang et al. (2020). The models' strong performance might benefit from the correct prediction of negative pairs. Because negative protein pairs are selected based on sequence dissimilarities, sequence dissimilarities can also be a predictive signal, causing the model to be overestimated. Additionally, protein pairs are removed at an identity of 95% and split randomly, resulting in highly similar protein pairs between the training and testing sets. To ensure that our model's performance does not benefit from these two factors, we constructed a virus-human PPI dataset (Section 4.3.1.2). Negative samples were randomly selected from any paired proteins that were not reported to interact. Protein pairs with more than 40% identity were removed, following the previous filter threshold for high-identity protein pairs (Sledzieski et al. 2021). Section 4.3.1.2 describes the PPI preprocessing to remove high-identity and low-quality PPIs, which is illustrated in Figure 4.4 A. In the following sections, we apply the three strategies outlined in Section 4.3.2 to split our virus-human PPI dataset and analyse PLM-interact-VH's performance under varying data-splitting conditions.

4.4.2 Model evaluation using three data-splitting strategies

To evaluate the model's performance under varying levels of prediction difficulty, we train and test models on datasets created through randomly split, Park and Marcotte's C1/C2/C3 and hold-out virus families, respectively (see Section 4.3.2). For a quick evaluation, we applied the binary classification model with the 35M ESM-2 for training. To track models' training performance on the training, validation and test sets, we record losses and AUPR values for each dataset at every 2000 training steps. We then use a line graph to visualize the model's loss and AUPR across each dataset (Figure 4.9, Figure 4.11, Figure 4.13). Here, the AUPR values are obtained by testing the model on 5000 protein pairs subsampled from the test dataset.

To further evaluate the models' ability to classify positive and negative protein pairs, we plot the distribution of predicted interaction probabilities for positive and negative protein pairs separately (Figure 4.10, Figure 4.12, Figure 4.14). Additionally, we plot the precision-recall curve of the model to evaluate the model's performance across the different thresholds for positive and negative samples. The results in this chapter demonstrated that predicting PPIs for randomly split and Park and Marcotte's C1/C2/C3 test sets is much easier than predicting PPIs for hold-out virus families. This challenging prediction task on held-out virus families might be caused by diverse virus-human PPI mechanisms in different virus families.

4.4.2.1 Random split

The loss and AUPR values of the training, validation and test protein pairs are separately shown in Figure 4.9. During training, the losses decreased from 1.1 to 0.9, while the AUPR values increased from 0 to above 0.9. The losses decreased from roughly 3 to 2.2 for the validation and test datasets, whereas the AUPR values improved from 0 to 0.99. Overall, PLM-interact-VH achieved a strong performance on each dataset, with AUPR values exceeding 0.9.

We show the distribution of predicted interaction probabilities for positive and negative protein pairs (Figure 4.10 A) and the precision-recall curve on test protein pairs (Figure 4.10 B). The results show that PLM-interact-VH successfully classifies positive and negative protein pairs with a threshold of 0.5, achieving an AUPR of 0.91 on the test set.

4.4.2.2 Park and Marcotte's C1, C2 and C3

As mentioned in Section 4.3.2.2, C1, C2 and C3 represent three classes of dataset split, where the protein overlaps between each test protein pair and training proteins are different. Figure 4.11 shows that the range of loss (red line) and AUPR (blue line) differs across datasets. For

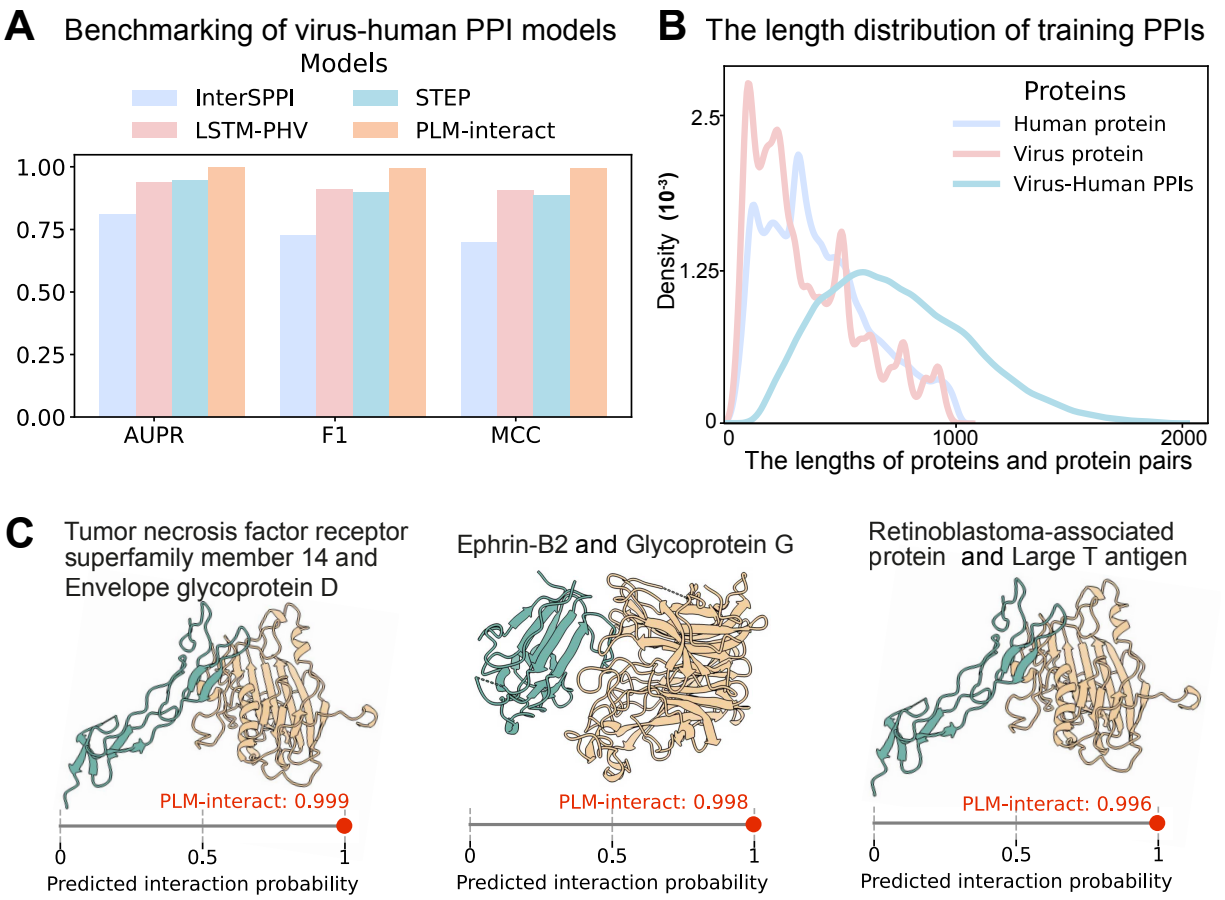

Figure 4.8. Benchmarking results of virus-human PPI models. **A**. Comparison of AUPR, F1 and MCC metrics of PLM-interact against recent virus-human PPI models. **B**. The distribution of the length of virus proteins, human and virus-human protein pairs. **C**. The virus-human PPIs are correctly predicted by our model and the 3D complex structures of virus-human PPIs are experimentally verified structures collected from the human-virus PPI database (HVIDB (Yang et al. 2021)). From left (green) to right (yellow), these interacting protein structures are: Tumour necrosis factor receptor superfamily member 14 (Human protein: Q92956) with Envelope glycoprotein D (human herpes simplex virus 1: P57083), Ephrin-B2 (human protein: P52799) with Glycoprotein G (Nipah virus protein: Q9IH62) and Retinoblastoma-associated protein (human protein: P06400) with Large T antigen (Simian virus 40: P03070). Note: The metrics results of the other three models in panel **A** are taken from STEP (Madan et al. 2022) paper.

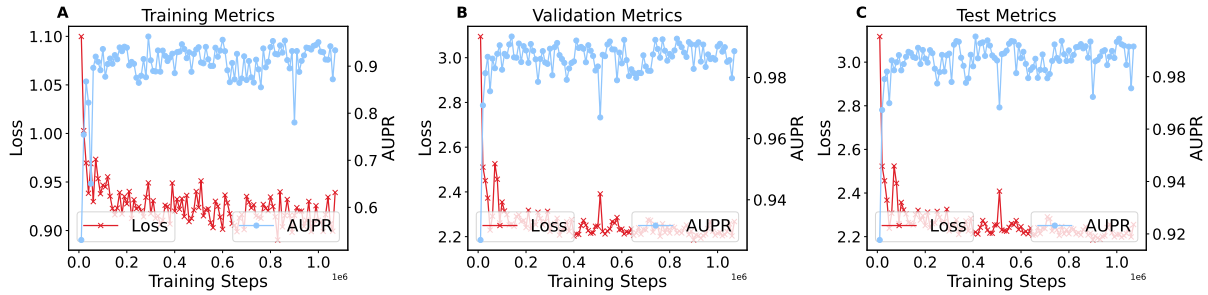

Figure 4.9. The line plots show changes in loss and AUPR values over the training steps for the training, validation and test datasets obtained by random split, respectively. The left y-axis represents the loss values, while the right y-axis shows the AUPR values.

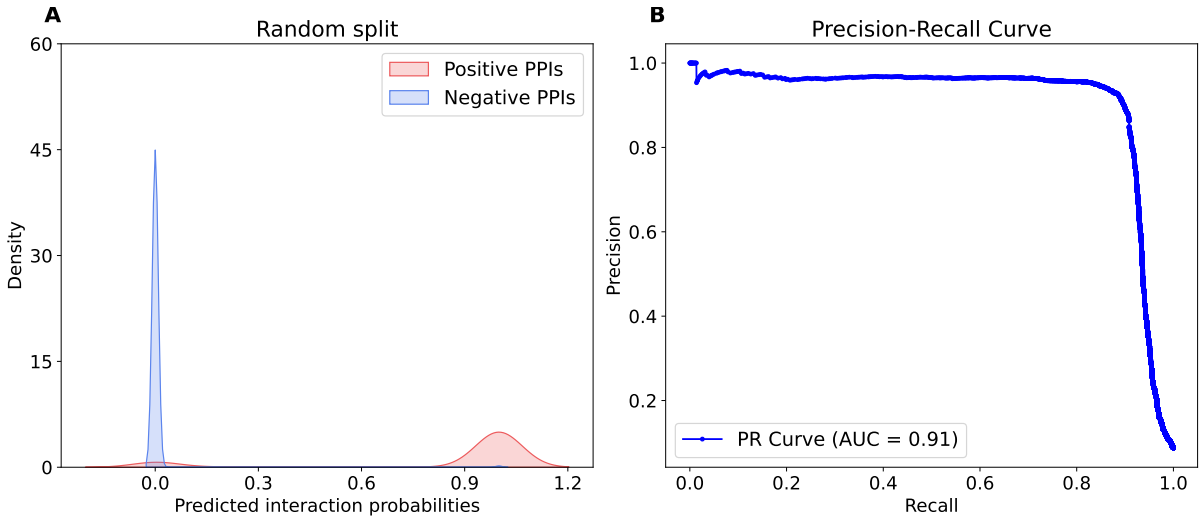

Figure 4.10. The distribution of predicted interaction probabilities for positive and negative protein pairs and the precision-recall curve on the random split test dataset. The left panel shows the predicted interaction probabilities distribution for positive and negative protein pairs. In the binary classification task of PPIs, positive and negative protein pairs are classified based on an interaction probability threshold of 0.5. The right panel displays the precision-recall curve, illustrating the model's performance on the randomly split test set.

C1, C2, and C3, the loss decreases from 1 to 0.9 in the training dataset. In the validation and test datasets, the loss for C1 decreases from 2.5 to 2.1, whereas for C2 and C3, it decreases from approximately 2.8 to 2.3. The result shows that the training dataset exhibits the smallest range, while the loss range is larger in the validation and test datasets.

For C3, after 1,000,000 training steps, the loss increases and the AUPR decreases in both the validation and test datasets. In contrast, the training dataset shows a decrease in loss and an increase in AUPR. This suggests that the model begins to overfit after 1,000,000 training steps. To prevent overfitting, the maximum number of training steps should be adjusted accordingly for this dataset. For the training sets, C1, C2 and C3 achieve AUPR values ranging from 0.7 to 0.95, while the AUPR values for validation and test sets are above 0.95.

In Figure 4.12, the left panel shows the distribution of predicted interaction probabilities of positive and negative protein pairs, while the right panel shows the precision-recall curves. In this binary classification task for PPIs, positive and negative protein pairs can be classified based on an interaction probability threshold of 0.5 (Figure 4.12 A, C, E). The AUPR values for the test protein pairs of C1, C2 and C3 are 0.88, 0.92 and 0.92, respectively (Figure 4.12 B, D, F). Interestingly, although C3 doesn't have overlap proteins within each training and test pair, it performs as well as C1 where each test pair shares two proteins with the training proteins. This suggests that training and testing protein pairs in C3 may share high protein identities, contributing to strong prediction performance. Further analysis of protein identities between the training and test datasets can be found in Section 4.4.6.

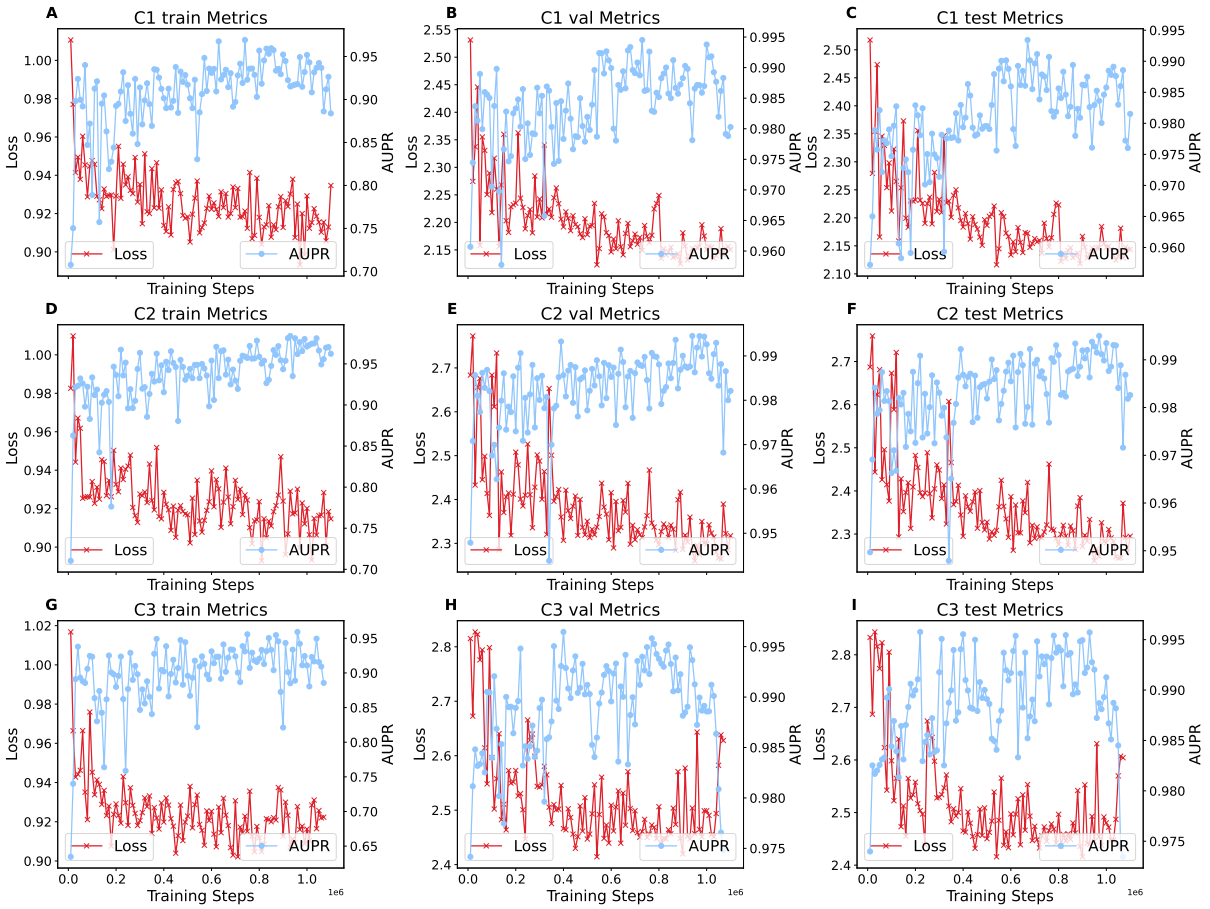

Figure 4.11. The line plots show the changes in loss and AUPR values over the training steps across the training, validation and test datasets created by Park and Marcotte's C1/C2/C3, separately. The left y-axis represents the loss values, while the right y-axis corresponds to the AUPR values.

4.4.2.3 Hold-out virus family

To evaluate the performance of PLM-interact-VH in holdout virus families, we selected each of the three most dominant virus families from our constructed dataset as a test set (Figure 4.1). Figure 4.13 respectively shows the loss and AUPR values of the model in each of the three holdout virus families: Coronaviridae, Retroviridae and Orthomyxoviridae. It shows that the model performs significantly better in the training dataset than the validation and test datasets, indicating that distinct patterns exist between the training and validation/test datasets. In the training data, the loss decreases and AUPR increases, while in the validation and test PPIs, the loss increases and AUPR fluctuates. The reason is probably that the viruses in the validation and test PPIs belong to virus families excluded from the training data; it is challenging to learn virus-human PPI mechanisms from different virus families. In each hold-out virus family classification task, the validation and test losses initially increase, indicating that the model struggles to generalize to the unseen virus families.

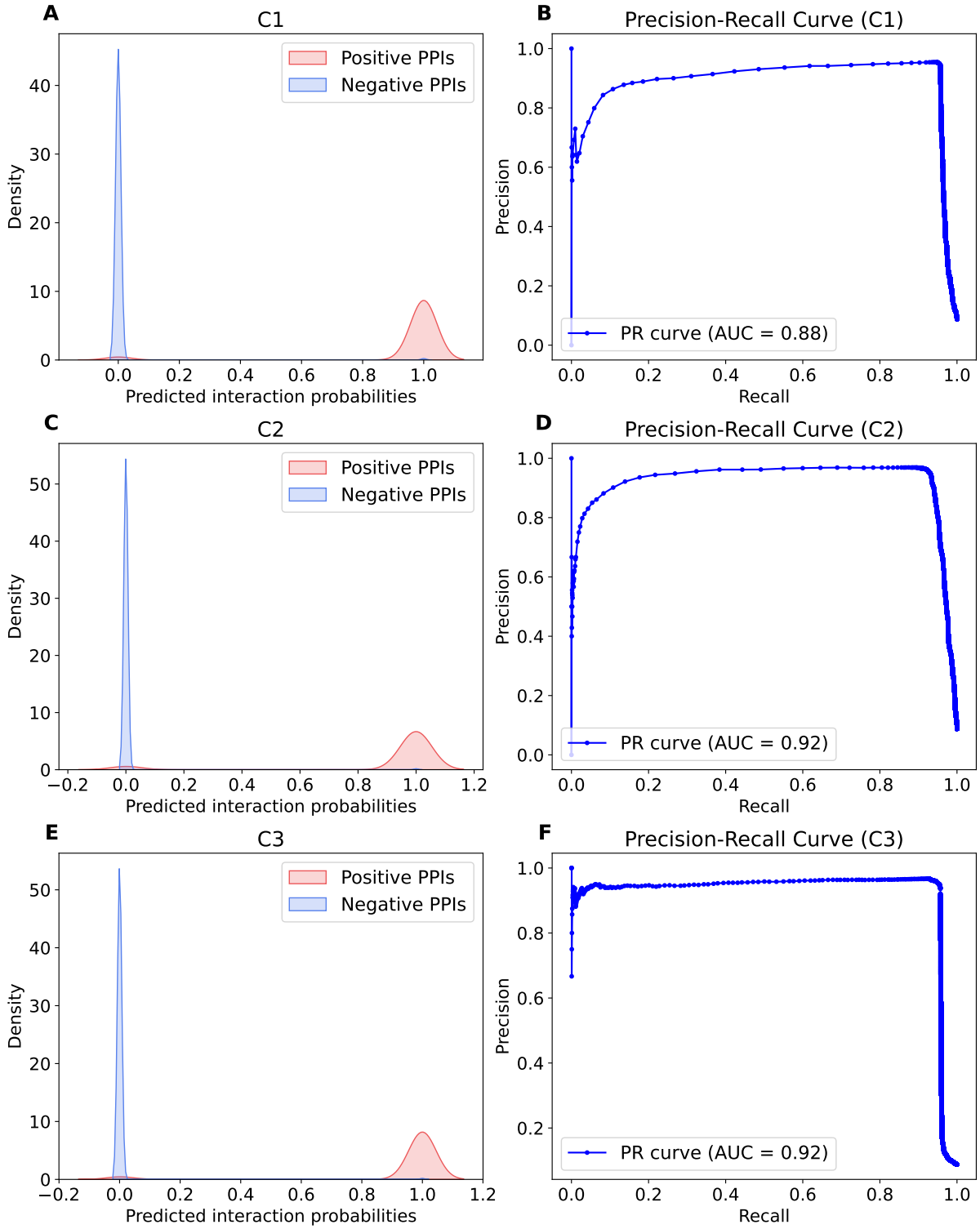

Figure 4.12. The distribution of predicted interaction probabilities for positive and negative protein pairs and the precision-recall curves on the C1, C2 and C3 test datasets. The left panel shows the predicted interaction probabilities distribution for positive and negative protein pairs. The right panel presents precision-recall curves demonstrating the model's performance separately on C1, C2 and C3 test datasets.

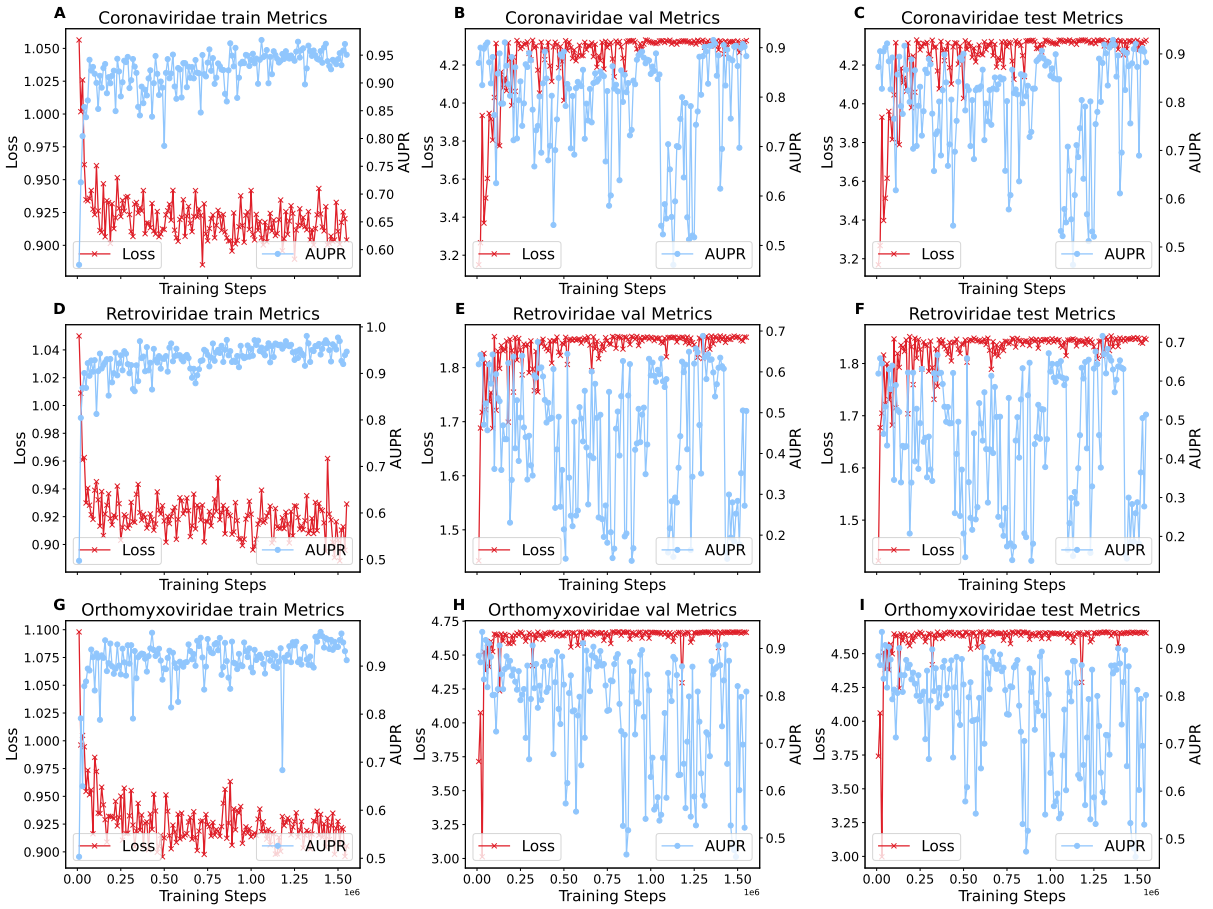

Figure 4.13. The line plots show the changes in loss and AUPR values over the training steps across the training, validation and test datasets created by holding out virus families, separately. The left y-axis represents model losses, while the right y-axis shows the AUPR values.

In the left panel of Figure 4.14, the positive and negative protein pairs cannot be classified effectively using an interaction probability threshold of 0.5. Most positive protein pairs have predicted interaction probabilities below 0.5 (Figure 4.14 A, C, E). The AUPR values for the test PPIs of the hold-out virus families are 0.44, 0.4 and 0.29, respectively (4.14 B, D, F).

In summary, random and Park and Marcotte's C1/C2/C3 splitting strategies for creating training, validation and test datasets demonstrated strong predictive performance, with AUPR values on test PPIs around 0.9. In contrast, the three hold-out virus families pose significant challenges, with AUPR values ranging from 0.29 to 0.44. The results in this section demonstrated that predicting PPIs for hold-out virus families is challenging due to the diversity among different virus families. This complexity may arise from the distinct features and patterns in training and test PPIs, as different virus families have different mechanisms to enter a human host cell (Goodacre et al. 2020) and are typically less likely to share similar patterns (Iyer et al. 2001). To further investigate the impact of similarities between training and test datasets on the model's performance, a study of protein identities between training and test PPIs is provided in Section 4.4.6.

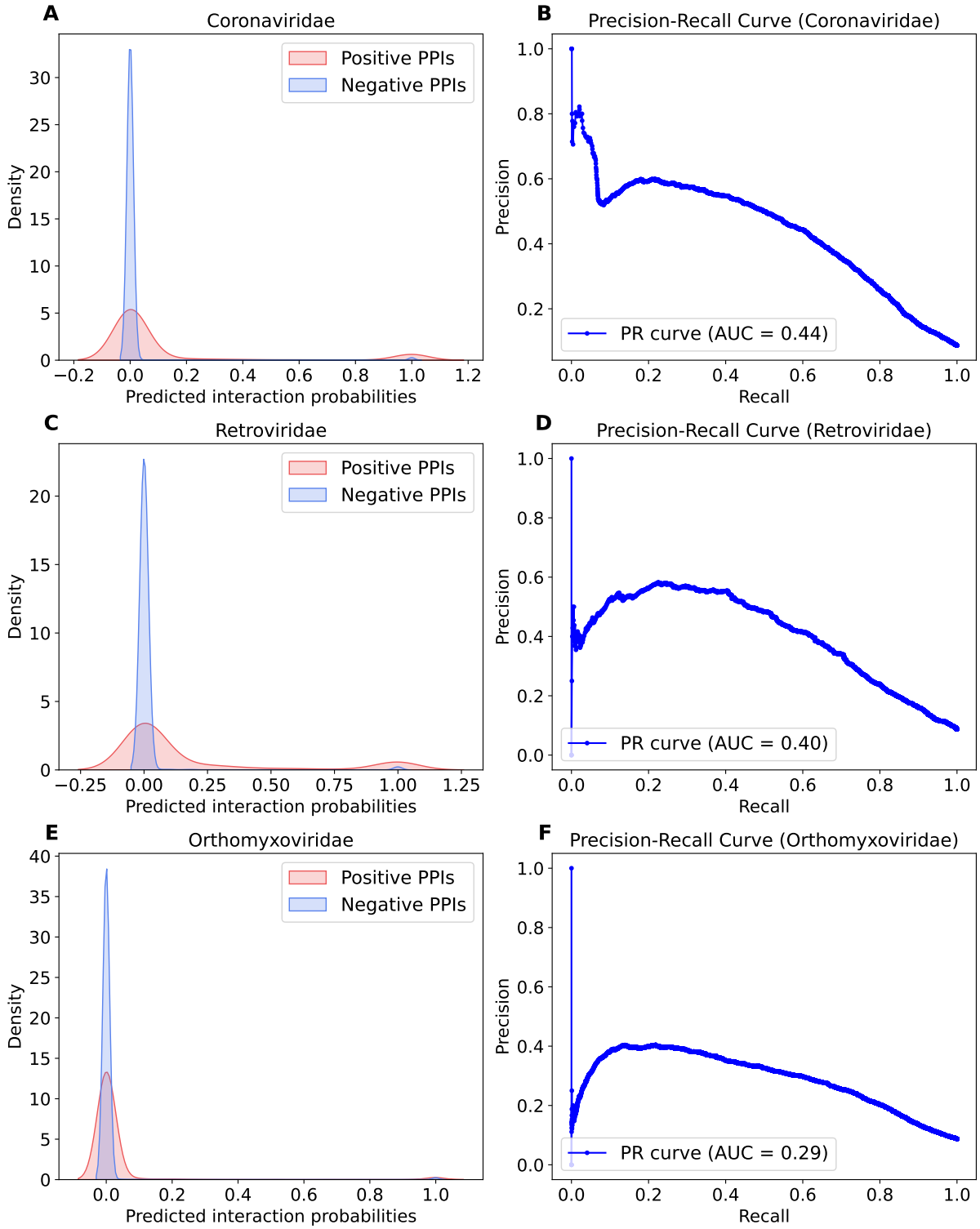

Figure 4.14. The distribution of predicted interaction probabilities for positive and negative protein pairs and the precision-recall curves in the three hold-out virus families. The left panel shows the distribution of predicted interaction probabilities for positive and negative protein pairs. The right panel is the precision-recall curves of test protein pairs.

4.4.3 Virus-human complex structures

To further analyse our model's performance, we selected virus-human PPIs from each test dataset described in Section 4.3.2. All selected virus-human PPIs have corresponding experimental virus-human complex structures available in HVIDB (Yang et al. 2021), except for PPIs in Figure 4.15 G and I, which are predicted by Chai-1 (Discovery et al. 2024), as these PPIs do not exist in HVIDB. We then use ChimeraX (Pettersen et al. 2021) to visualise these virus-human PPI structures and present the predicted interaction probability of PLM-interact-VH for each example (Figure 4.15).

PPI complex structures are shown for the three test PPIs from the random split dataset (Figure 4.15 A-C), one test PPI from each of Park and Marcotte's C1, C2 and C3 datasets (Figure 4.15 D-F), and one test PPI from each of the three hold-out virus families' datasets (Figure 4.15 G-I). Below each PPI structure, the corresponding line bar plot shows the predicted interaction probability of PLM-interact-VH. The ipTM scores for Figure 4.15 G and I are 0.21 and 0.15, respectively, indicating that Chai-1 provides failed prediction scores (ipTM < 0.6). In contrast, PLM-interact-VH gives correct predictions (> 0.5) in all cases.

4.4.4 Fine-tuning human PPI model for virus-human PPI prediction

Based on the results presented in Section 4.4.2, we demonstrated that holding out a virus family as test PPIs is more challenging than the other two data split strategies. In this section, we apply two approaches to enhance the model's performance on hold-out virus families. Firstly, PLM-interact with a loss ratio of 1:10 and 650M is the optimal training setup for model training, as outlined in PLM-interact optimisation experiments in Section 3.5.1. All models in Section 4.4.2 are trained with PLM-interact-VH's 35M binary model for quick testing. Therefore, the first approach to improve virus-human PPI prediction is to retrain the model using the optimal training setup in Section 3.5.1.

Secondly, we assume that the human PPI model in Chapter 3 has the potential to be extended for virus-human PPI prediction because viruses might take the human PPI mechanism to interact with human proteins due to viral mimicry (Section 1.1.4.2). Therefore, to further improve the model's performance, we fine-tune the human PPI model trained with human STRING V12 PPIs in Chapter 3, on each of the training datasets of hold-out virus families separately. This fine-tuned model keeps the same optimal training setup and an equal number of training epochs with the first approach for equality comparison.

To investigate if we can improve the virus-human model's performance by these two approaches,

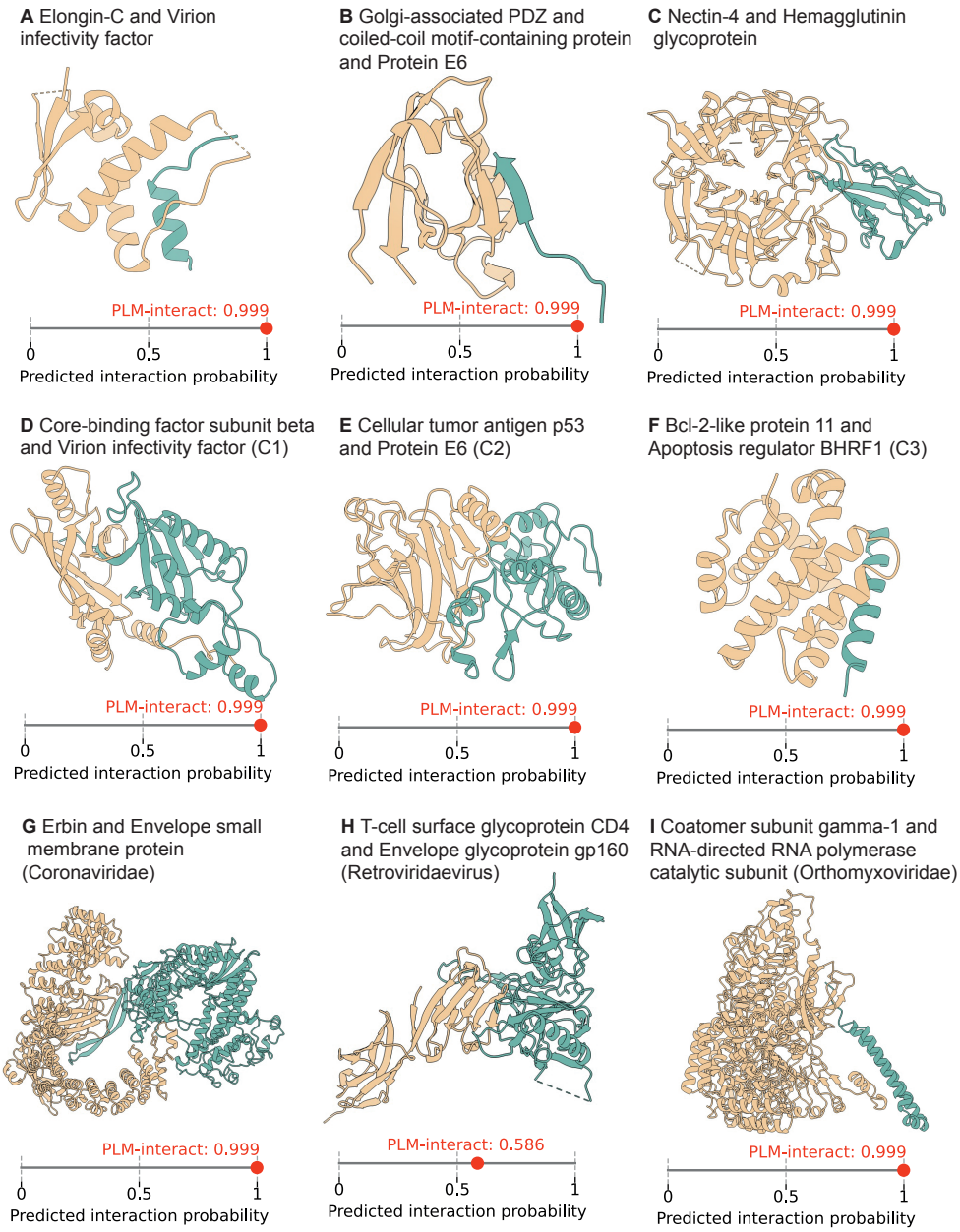

Figure 4.15. Virus-human complex structures. **A-C** is from the random split test dataset. **A**: Elongin-C (Q15369) and Virion infectivity factor (P12504) from HIV1; **B**: Golgi-associated PDZ and coiled-coil motif-containing protein (Q9HD26) and Protein E6 (P06463) from Human papillomavirus type 18; **C**: Nectin-4 (Q96NY8) and Hemagglutinin glycoprotein (Q786F2) from Measles virus (strain Ichinose-B95a) (MeV). **D-F** is from C1, C2 and C3, respectively. C1: Core-binding factor subunit beta (Q13951) and Virion infectivity factor (P12504) from HIV-1; C2: Cellular tumor antigen p53 (P04637) and Protein E6 (P03126) from Human papillomavirus type 16; C3: Bcl-2-like protein 11 (O43521) and Apoptosis regulator BHRF1 (P0C6Z1) from Human herpesvirus 4. **G-I** is from three holdout virus families, respectively. Coronaviridae: Erbin (Q96RT1) and Envelope small membrane protein (A3EXD5) from Bat coronavirus HKU5; Retroviridae: T-cell surface glycoprotein CD4 (P01730) and Envelope glycoprotein gp160 (P04578) from HIV-1; Orthomyxoviridae: Coatomer subunit gamma-1(Q9Y678) and RNA-directed RNA polymerase catalytic subunit (Q1K9H5) from Influenza A virus. The predicted interaction probabilities of PLM-interact-VH are shown below each virus-human PPI structure.

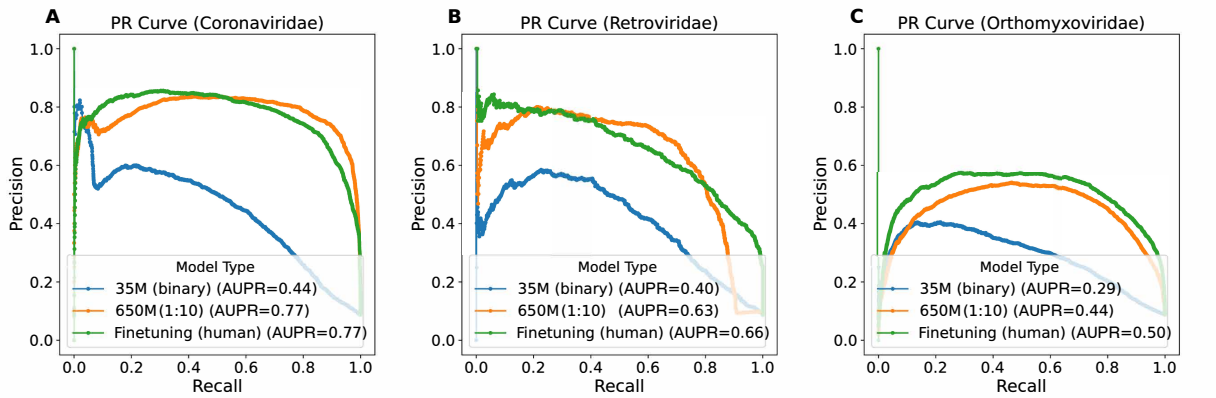

Figure 4.16. The prediction results of virus-human PPIs from hold-out virus families using three models: virus-human PPI with binary and 35M ESM-2, virus-human PPI with 650M ESM-2 (1:10), and fine-tuning the human STRING V12 PPI model using virus-human PPIs. The precision-recall curves of the three models show the models' performance on three hold-out virus families, respectively.

we plot precision-recall curves for each hold-out virus family based on three models: 35M (binary), 650M (1:10) and finetuning (human) (Figure 4.16). Firstly, the optimal training setup 650M (1:10) model has an improvement of 75%, 57.5% and 51.7% AUPR compared to the 35M binary model, respectively. The prediction for the Coronaviridae family is better than the other two virus families, with an AUPR of 0.77, while the Retroviridae and Orthomyxoviridae are 0.63 and 0.44, respectively. For the finetuning (human) model, AUPR values are 0.77, 0.66 and 0.5 on three hold-out virus families, separately. The AUPR remains consistent with the 650M (1:10) model in Coronaviridae, while AUPR values have an improvement of 4.8% and 13.6% on Retroviridae and Orthomyxoviridae families, respectively. Overall, the 650M (1:10) model shows improvement compared with the 35M binary model on all three holdout virus families. The results demonstrated that fine-tuning human PPI models improves models' performance on the Retroviridae and Orthomyxoviridae families. This indicates that combining human-human and virus-human PPIs for training can strengthen the model's performance.

4.4.5 Model generalizable experiments

In addition to improving the model's performance for hold-out virus families, we investigate which model is the best and can predict PPIs for intra-species (human, mouse, worm, fly, yeast and *E. coli*) and inter-species (virus-human). Firstly, the human STRING V12 PPI model in Chapter 3 is directly used to predict PPIs of three held-out virus families described in Section 4.3.2.3. However, this human model reported that all AUPR values are no more than 0.1. Secondly, the 650M (1:10) virus-human PPI model and the finetuning (human) model, trained on datasets of held-out virus families (see Section 4.4.4), are used respectively to predict five held-out host species PPIs in Chapter 3. Figure 4.17 shows that all AUPR values of 650M (1:10)

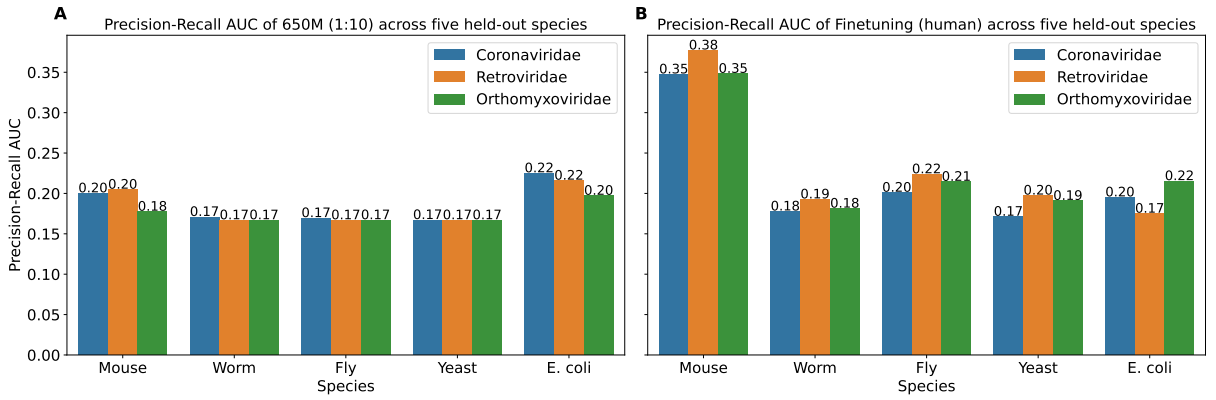

Figure 4.17. The results of the virus-human model and fine-tuning human model on five held-out host species PPIs. The bar plots of two models showed these models' performance on five held-out host species PPIs (mouse, worm, fly, yeast and *E. coli*), respectively.

models are roughly 0.2. In contrast, AUPR values of finetuning (human) models show improvements in mouse PPI prediction at roughly 0.35, while other host PPI prediction performance remains consistent with the 650M (1:10) model. All these AUPR values obtained by both models are much smaller than the human PPI model reported in Chapter 3. Based on these results, we discovered that human PPI models only work for human PPIs and held-out intra-species (mouse, worm, fly, yeast and *E. coli*) PPI prediction; virus-human PPI models only work for virus-human PPI prediction. We anticipate that training on both virus-human and human PPI datasets will provide the potential to develop a generalizable model for PPI prediction of intra-species and inter-species.

4.4.6 Protein similarities between train and test datasets

To investigate protein similarities between training and test datasets, we use the maximum percentage identity to estimate similarities between training and test protein pairs. The maximum percentage identity represents the highest identity in a protein pair with any proteins in training PPIs. The percentage identities between training and test proteins are obtained by MMseq2 (Steinegger et al. 2017). Figure 4.18 shows the max percentage identities of positive and negative protein pairs for datasets created by three strategies described in Chapter 4.4.2. For random split datasets and Park and Marcotte's C1, C2 and C3, the max percentage identity of positive protein pairs between training and test ranges from 80 to 100. The percentage identity of most protein pairs is between 90 and 100, while the counterparts of hold-out virus families are between 0 and 100, indicating that this test data includes both low identities (< 40%) and high identities (>80%) between training and test positive proteins. For all negative protein pairs, the max percentage identity ranges from 0 to 100, and some have an identity of 0.

Figure 4.19 shows the accuracy of positive test PPIs across the different thresholds of protein per-

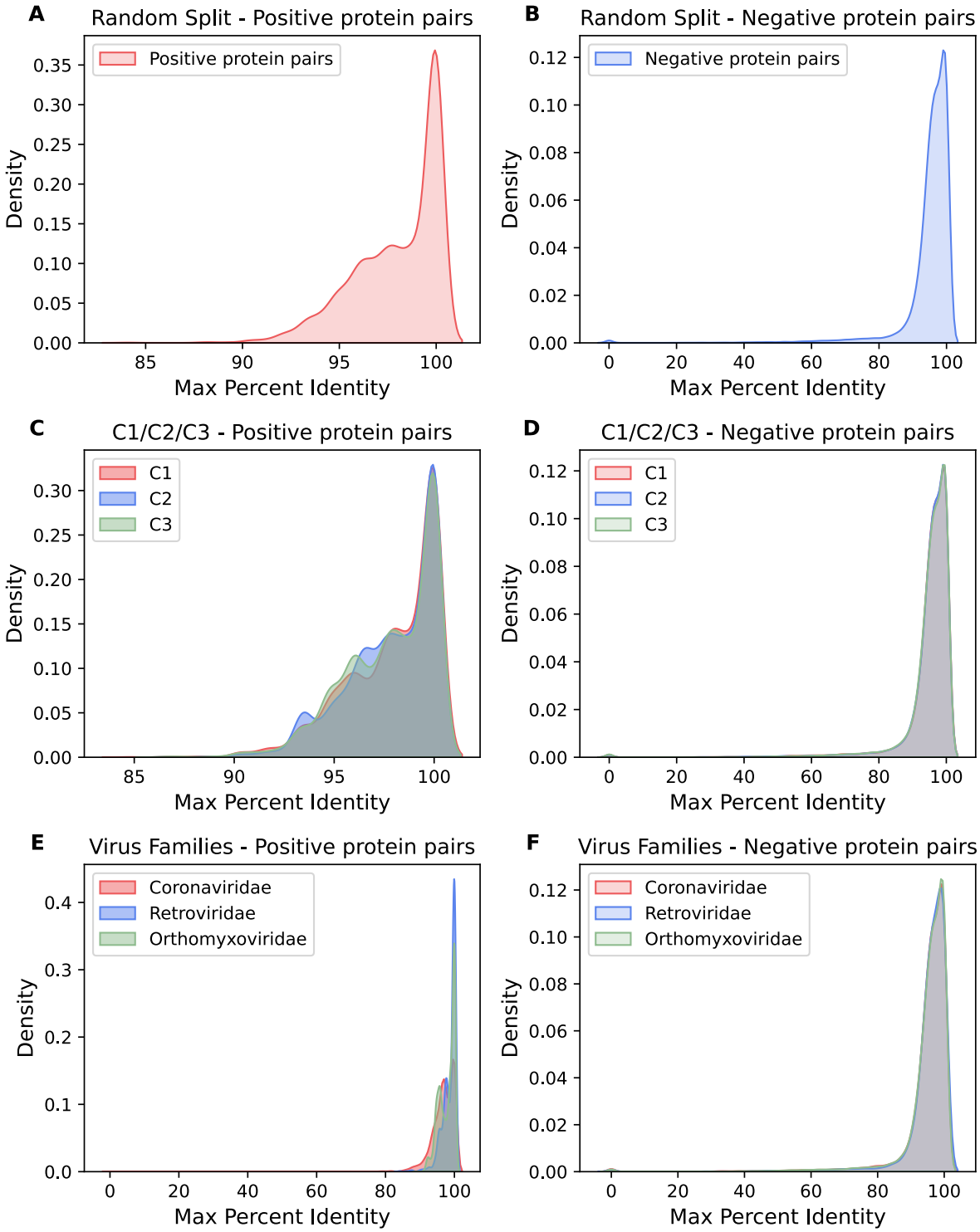

Figure 4.18. The distribution of the maximum percentage identity between training and test PPIs obtained using three data-spilling strategies. The distribution plot of the left panel is for positive protein pairs and the right panel is for negative protein pairs.

centage identity between training and test datasets. The max percentage identity of all datasets between training and test proteins is between 0 and 100, except for Park and Marcotte's C1, which is between 50 and 100 due to the two proteins in a test pair existing in the training proteins. In the random split and Park and Marcotte's C1/C2/C3, the accuracy increases from 0.2 to roughly 0.9 when the percentage identity is above 80. In contrast, accuracy for three hold-out virus families fluctuates and all accuracy values are below 0.2. We demonstrated that predicting hold-out virus families is still challenging, although the maximum protein identities increased between proteins in each test protein pair and the training proteins. It indicates that virus-human PPIs are complex across different virus families, making it difficult for the model to generalize protein interaction patterns.

4.5 Discussion

In this chapter, we utilized the human PPI model PLM-interact in Chapter 3 to train with virus-human PPIs to develop the virus-human PPI prediction model PLM-interact-VH. PLM-interact-VH extends a single protein-focused PLM ESM-2 to paired proteins and has been successfully implemented for PPI prediction in held-out intraspecies, including mouse, worm, fly, yeast and *E. coli*. We demonstrate PLM-interact-VH's performance in a virus-human PPI benchmarking dataset, showing a significant improvement over the state-of-the-art approaches. We further investigate PLM-interact-VH's performance on datasets created by split training, validation and test PPIs using three strategies: random split, Park and Marcotte's C1/C2/C3, and hold-out virus families. We discover that prediction is challenging on hold-out virus families because virus-human protein interaction patterns are complex and difficult to generalize across different virus families. We further demonstrate that fine-tuning the human PPI model on virus-human PPIs can improve the performance of the model trained on virus-human PPIs only. This result indicates that the combined training of PPIs from human-human and virus-human protein interactions provides insight into developing generalized PPI models.

Virus-host PPI prediction is challenging because virus proteins can be highly divergent. As such, making assumptions about the conservation of PPIs is problematic. Viruses can acquire genes from their hosts, undergo convergent evolution, and exploit intrinsically disordered regions in proteins. This indicates that the mechanisms enabling them to interact with or mimic host proteins can be complex. In previous PPI prediction models, NLP algorithms process protein sequences analogous to estimating context semantics and document similarities in NLP tasks. Pre-trained protein language foundation models have been developed and implemented for secondary prediction, protein 3D structure prediction and mutational effects prediction (Rives et al. 2021). The PLM-based PPI models have become state-of-the-art approaches due to the PLM's

ability to encode different levels of protein patterns, including sequence composition and contact, structures and protein properties.

As described in Chapter 3, PLM-interact is analogous to the 'next sentence' prediction task (Reimers et al. 2019), with the goal of training on binary labels that represent positive and negative protein pairs. PLM-interact trained on human PPIs improved significantly in mouse, fly, worm, yeast and *E. coli* test PPIs compared with previous PPI models (Figure 3.4). In this chapter, we implemented PLM-interact on virus-human PPI datasets to train a virus-human PPI model PLM-interact-VH. Unlike previous virus-human PPI models that only trained a classification head on protein features, PLM-interact-VH retrains ESM-2 by jointly encoding virus-human protein pairs, allowing amino acids in a virus protein to be linked by amino acids from a human protein through the attention mechanism of ESM-2. Our improved performance relative to the state-of-the-art model STEP (Madan et al. 2022) is impressive, showing a 5.7% improvement in AUPR (Figure 4.8 A). The results demonstrated that PLM-interact-VH can be implemented on virus-human PPI prediction.

The AUPR, F1-score and MCC metrics of models in the benchmarking of virus-human PPI prediction task are close to 1, except for InterSPPI (Yang et al. 2020). The main reason is the high protein similarities (>40%) of protein pairs between training and test datasets in this benchmarking dataset. Another possible reason would be that negative protein pairs are chosen based on sequence dissimilarities; this makes negative protein pair prediction easier. To train generalisable virus-human models, firstly, we collected virus-human PPIs sourced from seven public PPI databases. Next, we remove high-identity PPIs with an identity of 40% and randomly choose negative protein pairs. Then, we split this constructed dataset into training, validation and test sets using three strategies, making varied levels of difficulty in prediction. We observed that predicting hold-out virus families is the most challenging task. All AUPR values are below 0.5 on three hold-out virus families (Figure 4.14), while AUPR values are roughly 0.9 on the randomly split and Park and Marcotte's C1/C2/C3 datasets (Figure 4.10 and Figure 4.12). The reason is that virus-human PPI mechanisms are diverse due to the virus diversity between the training and test sets. Different virus families are evolutionarily divergent, making predicting virus-human PPIs of hold-out virus families difficult. The range of maximum protein identities between test and training positive protein pairs of hold-out virus families is 0 to 100 (Figure 4.18). This indicates that positive protein pairs of hold-out virus families include protein pairs with low similarities (< 20%) between training and test PPIs. This also explains the challenging prediction task of hold-out virus families. We discovered that in each family a single species dominated the PPI data for each of the three hold-out virus families (Figure 4.1 B). This indicates that holding out a virus family of top-ranked three virus families causes the training dataset to lose many PPIs from well-studied virus species. In future work, we will hold out virus species that are not involved in many PPIs as a test. The remaining training will include PPIs of well-

studied species and enhance the model to learn virus-human PPIs, which provides the potential to train a generalisable model.

Furthermore, we fine-tune the STRING V12 human PPI model on each training dataset of three hold-out virus families, separately. We discovered that fine-tuning the human PPI model can improve the performance of virus-human PPI prediction of hold-out Retroviridae and Orthomyxoviridae families (Figure 4.16). Based on this finding, we suggest that human PPI models can be extended to train virus-human PPI models. Viruses mimic human proteins to escape the host defence system, indicating that the virus-human and human-human PPIs probably share the same or similar PPI mechanisms. This provides clues for improving the virus-human model's performance by fine-tuning human PPI models.

While PLM-interact-VH has demonstrated significant improvements, we still need high-quality virus-host experimental PPI data to optimize our models and obtain reliable predictions. The attention mechanism of our model applied to long-range sequence interactions enhances our understanding of protein interactions, the fundamental 'language' of molecules. Effective virus-host PPI predictors could provide molecular details to virus-host species prediction approaches, which mainly rely on genome and protein composition signals, ignoring host molecules interacting physically with viral molecules (Babayan et al. 2018; Young et al. 2020; Liu et al. 2024a; Roux et al. 2022). Recent progress within SARS-CoV-2 PPI studies mapped out a complex interaction landscape between the virus and human proteome (Stukalov et al. 2021; J et al. 2021). Similarly, other viruses might have complex interactions with human and animal hosts. Leveraging PPIs could lead to better tools for predicting zoonotic events and offer the potential to identify hosts for novel viruses.

In conclusion, we showed that PLM-interact-VH performed best on the virus-human benchmarking task. To train generalizable virus-human PPI models, we created a virus-human PPI dataset by integrating the latest version of seven PPI datasets. We split the training, validation and test datasets using three strategies to evaluate the models' performance. We discovered that predicting hold-out virus families is challenging due to the diverse PPI mechanisms across different virus families. Furthermore, we provide fine-tuned human PPI models that improve PPI prediction of hold-out Retroviridae and Orthomyxoviridae families. Based on these findings, we anticipate that future studies will be on a large scale of high-quality PPIs to further improve virus-human PPI prediction. Moreover, virus-human PPI models learn molecular interactions between viruses and humans, which provides insights into designing models for predicting virus-host associations.

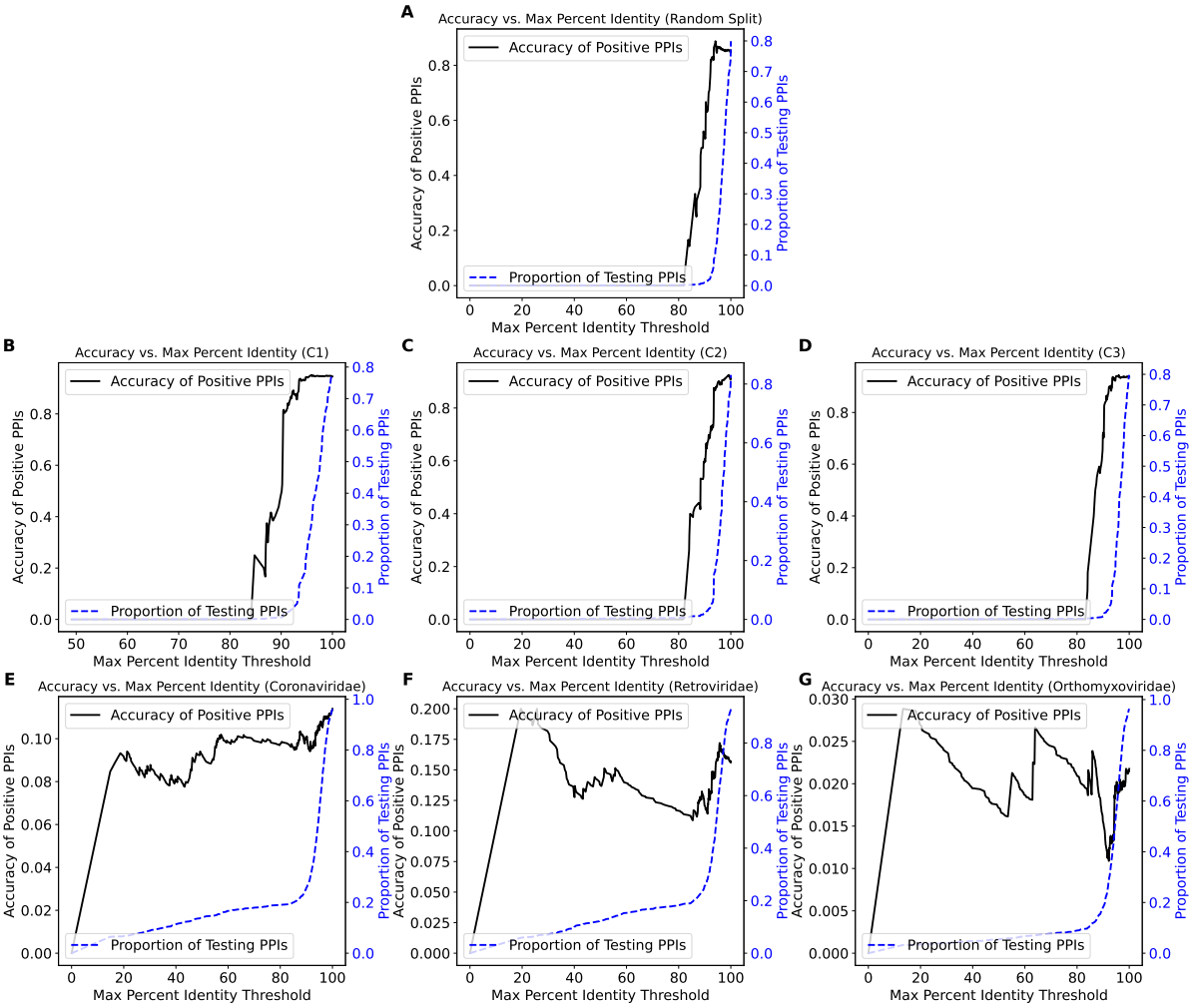

Figure 4.19. The accuracy of positive test PPIs across the different thresholds of protein percentage identity between protein and test proteins. The black line shows the accuracy of positive protein pairs with changes in the maximum percentage identity between training and test proteins. The blue line is the proportion of test PPIs. The highest identity between any training protein and a protein in a positive test pair is considered the maximum identity for the positive test pair.

Chapter 5

General Discussion

This thesis aimed to use AI to exploit the patterns captured by pre-trained PLMs to explore virus-host interactions, both at the species and the protein-protein interaction (PPI) levels. I developed deep-learning approaches to predict host specificity for prokaryotic and eukaryotic viruses and PPIs for human-human and virus-human. I have demonstrated that pre-trained PLMs can capture biological patterns from viral proteins to identify host specificity for viruses, as well as be fine-tuned to understand complex PPI mechanisms and detect their mutational impacts.

In Chapter 2, I introduced a novel virus-host association prediction model EvoMIL that combines the pre-trained PLM ESM-1b and attention-based multiple instance learning (MIL). The results demonstrated that ESM-1b captures stronger predictive signals than sequence composition features (k-mers): amino acids, physicochemical properties and DNA. EvoMIL binary and multi-host classifiers achieve impressive AUC for all prokaryotic and eukaryotic hosts (Figure 2.3 and Figure 2.6). We demonstrated that EvoMIL achieved the highest accuracy of 75.25% on the independent test dataset, outperforming the state-of-the-art methods, including iPHoP (Figure 2.11). EvoMIL identifies important viral proteins that drive host specificity in binary and multi-class host prediction tasks. We show key viral protein GO annotations associated with *E. coli* and humans, respectively (Figure 2.13). Furthermore, we show that ESM-1b embeddings can capture functionally related information of identified key proteins associated with host specificity (Figure 2.14). This chapter aims to show the pre-trained PLM's abilities to represent protein features and improve virus-host association prediction. EvoMIL only encodes viral proteins for host prediction, so the range of host prediction is limited to trained hosts: 22 prokaryotes and 35 eukaryotes. Further analysis would include extracting host features to enable the model to predict any pair of virus-host associations.

In Chapter 3, I developed PLM-interact, a novel deep-learning PPI prediction model that trains ESM-2 on paired proteins. Protein pairs are analogous to sentence pairs; the MLM training

technique of the 'next sentence' prediction is implemented for the 'next protein' prediction. The previous PPI prediction model extracts protein features using a pre-trained PLM and then only trains a classifier for PPI classification (Figure 3.1 A). PLM-interact trains all layers' parameters of ESM-2 and classification head (Figure 3.1 B), so it is more generalized than training classification head only. We demonstrated that PLM-interact trained on human PPIs shows significant improvements in AUPR (16% to 28%) on PPI prediction of held-out host species compared with the state-of-the-art PPI model TT3D (Sledzieski et al. 2023) (Figure 3.4). We discovered that PLM-interact outperforms other models by identifying the most significant number of true positive protein pairs (Figure 3.5). Furthermore, our results show that PLM-interact can identify mutational impacts of human PPIs by investigating changes in predicted interaction probabilities between canonical and mutants (Figure 3.8), which offers insights into interaction-aware in-silico variant effect predictors. This chapter aims to train ESM-2 with MLM and classification to improve performance in PPI prediction of held-out species and detect mutational impacts associated with PPIs. We are training larger models and updating PLM-interact to make it generalizable for PPI predictions in a wider range of species. We anticipate that in the future PLM-interact will be a robust tool for identifying PPIs across species by integrating more high-quality training PPIs from different species.

To identify PPIs between viruses and humans for understanding virus-human interactions, Chapter 4 creates a virus-human PPI dataset by integrating the latest version of seven PPI datasets and investigates whether PLM-interact can be implemented to predict virus-human PPIs. I retrained PLM-interact on the virus-human PPI benchmarking dataset to develop a PLM-interact-VH model, and the result showed a significant improvement (5-23%) in AUPR over state-of-the-art prediction approaches (Figure 4.8 A). To train a robust virus-human PPI model, we evaluated the model's performance on different datasets created using three data-splitting strategies. We observed that predicting PPIs for a hold-out virus family is more challenging than random split and Park and Marcotte's C1/C2/C3 (Figure 4.10, Figure 4.12, Figure 4.14). We further demonstrated that AUPR values improved in hold-out Orthomyxoviridae and Retroviridae families by fine-tuning the human STRING V12 PPI model on each training dataset of hold-out virus families (Figure 4.16). This chapter shows that PLM-interact-VH performs best among state-of-the-art virus-human PPI models. Splitting data using different strategies impacts the model's performance; higher protein similarities between training and test sets typically achieve better model performance (Section 4.4.6). I think that in the future, our constructed virus-human PPI dataset could be a new virus-human PPI benchmarking dataset. Besides, fine-tuning the human PPI model on virus-human PPIs will save training time and offer the potential to develop generalizable PPI prediction models.

Negative samples are necessary for machine learning classification models, which allows the model to learn features of positive and negative samples for classifying classes effectively. The

actual negative samples are usually lacking in public databases, as typically only positive samples are collected. Additionally, positive samples are limited, and only a few species are well-studied; virus-host association and PPI networks remain sparse. In Chapter 2, we collected virus-host associations from the public database VHDB (Mihara et al. 2016) with positive samples only. Without a curated database for negative samples, any pairs that are not reported as positive samples can be considered as negative samples. Therefore, these negative candidates might include false negative samples, as some virus-host pairs might be undiscovered positive samples. However, using all available negative samples could result in a highly imbalanced dataset. This imbalance can lead to biased model training and cause the model to learn mainly from negative samples. Therefore, subsampling from any virus-host pairs with unknown associations becomes the only scalable option. Because balanced positive and negative samples save training time and provide efficient predictors, we constructed a balanced training dataset in Chapter 2. In Chapter 3 and 4, PPI models (Liu et al. 2024b; Sledzieski et al. 2023; Sledzieski et al. 2021) used the ratio of 1:10 for positive-to-negative protein pairs because positive protein pairs are fewer in sparse PPI networks. Future work might include subsampling different ratios of positive and negative samples for EvoMIL and PLM-interact, which provides a comprehensive understanding of the impacts of negative samples in binary model training. In addition to the ratios of positive and negative samples, there are different strategies to select negative samples in the large set of negative candidates. In Chapter 2, two strategies are used for negative sampling: (1) sample negative viruses from all viruses that are in different genera than viruses in the positive dataset; (2) select negative viruses from those that infect hosts in the same taxonomic rank as the positive host. We discovered that sampling negative samples similar to positive ones makes the binary classification task more challenging. (Figure 2.3 E, F). In Chapter 3 and 4, we randomly choose any paired proteins that are not reported as positive pairs as negative protein pairs. Random sampling makes no biases for similar or dissimilar samples with positive samples. Based on this finding, we anticipate that future studies will test EvoMIL's performance using random-sampling negative samples.

The different data splitting strategies will generate different training, validation and test datasets, which leads to varied similarities between training and test datasets and impacts the model's performance. In Chapter 2, we evaluate the model's performance on binary and multi-class classification tasks by 5-fold cross-validation. In Chapter 3, Protein pairs with an identity of 40% are identified as high-identity pairs and removed in all training tasks. In Chapter 4, given that all predictors achieved an AUPR above 0.9 on the benchmarking virus-human dataset, we hypothesize that filtering high-identity PPIs with an identity of 95% remains highly similar protein pairs, making inference easier on this dataset. Therefore, we constructed a dataset removing highly similar PPIs with an identity of 40% to ensure that training datasets do not contain PPIs that share high protein similarities with test datasets. We further investigated the models' robustness by training and testing on different datasets with varying levels of protein similarities. We dis-

covered that predicting hold-out virus families is the most challenging task due to the divergent PPI patterns between training and test datasets (Figure 4.16). Overall, it is important to remove the high-identity pairs to ensure that training and test datasets do not share high similarities. Different data-splitting strategies provide a comprehensive evaluation of classification models.

In Chapter 4, predicting virus-human PPIs for hold-out family datasets remains challenging. Viruses are evolutionary divergent across families (Holmes 2011), and different viruses have different and divergent PPI mechanisms with host proteins (Goodacre et al. 2020). We discovered that fine-tuning the human PPI model on the virus-human PPI dataset fails to capture significant predictive signals and performs much worse than the human model PLM-interact in held-out host species PPI prediction (Figure 4.17 B). The main reason is that this fine-tuning model is trained for virus-human PPIs, and parameters optimized on the virus-human data lost previously obtained intra-species PPI patterns. In Figure 4.16, fine-tuning the human PPI model on virus-human PPI datasets improves the prediction of virus-human PPIs. The reason might be that high-quality and massive human PPI data provide good training data and can be extended for virus-human PPI prediction tasks. Training a wide range of PPIs allows the model to learn complex patterns from different species, although these are highly divergent. To train a generalizable model for two kinds of hold-out PPI prediction, we anticipate that in the future, combining PPIs from intra-species such as humans, yeast, and *E. coli* and inter-species such as virus-human PPIs will enhance model training and enable generalizable PPI inference.

As described in Chapter 1, viruses gain entry to their host via the specific host receptor for the viral life cycle, which involves hundreds of PPIs. Therefore, knowledge of virus-host PPIs is vital to understanding viral infection. Theoretically, it should be possible to exploit the prediction of virus-host PPIs to predict which hosts a virus can potentially infect. Because both virus and host have many proteins, it is unclear which virus-host PPIs allow virus-host associations (Goodacre et al. 2020). Recent graph-based algorithms offer the potential to apply protein-protein similarities to develop host prediction tools. For example, PGCN (Ma et al. 2025) is trained on a multi-view graph by combining two phage similarity networks obtained from protein clusters and protein sequence similarities, then trained on a graph convolutional network (GCN) to capture phage features for host prediction. Here, protein similarities are between viruses; it provides insights into using the PPI model to estimate protein-protein similarities for developing graph-based host prediction tools. Cherry (Shang et al. 2022) was developed for prokaryotic host prediction and trained a GCN on a multimodal knowledge graph that encodes virus-virus and virus-host connections. These graph-based approaches can combine multi-view features and encode connections between the virus and host in the graph. Virus-host and virus-virus connections in the graph can be determined by the predicted interaction probabilities of PPI models. We expect that in the future a deep-learning framework can be developed by using a trained PPI model to construct virus-host and virus-virus connections in the multimodal graph network and

then training a graph convolution network for host prediction.

Attention mechanisms and MLM have been implemented on translation tasks (Vaswani et al. 2017). These NLP techniques have been widely implemented in biological sequences such as protein (Rives et al. 2021; Brandes et al. 2022), DNA (Zhou et al. 2023; Nguyen et al. 2024), and RNA (Yu et al. 2024; Saberi et al. 2024). The PLM ESM-1b (Rives et al. 2021) encodes protein sequences and learns protein representations. See Section 1.2.8 for details about ESM-1b. ESM-1b has since improved in predicting protein 3D structures and generating new proteins. The number of parameters also improved from 650 million to 98 billion. ESMFold (Lin et al. 2023) leveraged the ESM-2 to predict 3D structures from protein sequences. The ESM Metagenomic Atlas (https://esmatlas.com) is constructed using ESMFold to predict over 617 million metagenomic protein sequences, which provides structure information for analysis of metagenomic data. ESM3 (Hayes et al. 2024) is a multi-track transformer that combines protein sequence structures and functions. Due to the MLM on sequences, structures and functions, ESM3 can generate three modalities for protein design and therapeutic applications. Recently, ESM-C (Scale 2024) was released and mainly designed to create protein representations, showing a significant improvement over ESM-2 on the benchmarking task. ESM-C is trained on larger datasets and better training resources; the 600M parameter ESM-C performs similarly to 15B ESM-2. Although 15B ESM-2 achieved better performance than 650M ESM-2, we trained 650M ESM-2 in Chapter 3 and 4, as 15B ESM-2 required extensive computational storage and was time-consuming. In the future, 600M ESM-C can be used to optimize our PLM-interact.

With the development of PLMs, generative models are proposed for protein and enzyme design (Strokach et al. 2022; Zeng et al. 2022), offering the potential to develop therapeutic strategies. RFdiffusion (Watson et al. 2023) and EvoDiff (Alamdari et al. 2023) are based on a diffusion framework to create new proteins, paving the way for therapeutic application and synthetic biology. RFdiffusion fine-tunes the RoseTTAFold (Baek et al. 2021) structure prediction network on protein structure denoising tasks to obtain a protein generative model. EvoDiff focused on sequence diversity and integrated evolutionary-scale data to generate proteins capturing sequence and functional space. Furthermore, genomic foundation models Evo (Nguyen et al. 2024) and NT (Dalla-Torre et al. 2024) are released for DNA and CRISPR-Cas generation. Evo is a multimodal deep-learning model designed to represent and generate genomic sequences. Evo is trained with StripedHyena architecture on bacterial, archaeal and phage genomes and plasmid sequences to learn DNA, RNA and protein modalities. This model can be applied to predict mutational effects and protein expression, as well as generate CRISPR-Cas complexes and transposable systems. NT is a nucleotide transformer model trained from DNA sequences, including humans and other species, which can be used for nucleotide representation in various genomic tasks such as molecular phenotype prediction. The protein and genomic foundation models are experiencing rapid growth, implementing these model training techniques in virology can help

design specific proteins targeting viruses, which speeds up the development of antiviral drugs and vaccines. Fine-tuning pre-trained foundation models on viral genomics or protein sequences provides insights into viral protein design. In the future, AI models will offer potential solutions for challenging problems, including viral protein structure prediction and sequence generation.

In conclusion, I introduced deep learning approaches that directly apply or fine-tune PLMs to learn complex protein patterns to understand virus-host interactions. These methods can identify the host specificity for novel viruses and identify PPIs within and between species, providing insights into understanding complex interaction mechanisms between viruses and hosts. We anticipate that in the future collecting more high-quality samples will enhance training datasets, and utilizing multi-layer classification heads and larger model sizes will make our models more generalizable and applicable to a broad range of species.

Bibliography

[1] Sephra Rampersad and Paula Tennant. “Replication and expression strategies of viruses”. In: *Viruses* (2018), p. 55.

[2] Sherlyn Jemimah and M. Michael Gromiha. “Insights into changes in binding affinity caused by disease mutations in protein-protein complexes”. In: *Computers in Biology and Medicine* 123 (Aug. 2020), p. 103829. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2020.103829. URL: https://www.sciencedirect.com/science/article/pii/S0010482520301943.

[3] Isabel Bandín and Carlos P Dopazo. “Host range, host specificity and hypothesized host shift events among viruses of lower vertebrates”. In: *Veterinary research* 42 (2011), pp. 1–15.

[4] Peter Daszak and Kate Jones. “Global trends in emerging infectious diseases”. In: *Nature* 451.7181 (2008), pp. 990–993.

[5] COVID - Coronavirus Statistics. “COVID - Coronavirus Statistics - Worldometer”. In: (2025). URL: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1?%22%20%5Cl%20%22countries%3Ca%20href=..

[6] Cassandra L Atzrodt et al. “A Guide to COVID-19: a global pandemic caused by the novel coronavirus SARS-CoV-2”. In: *The FEBS journal* 287.17 (2020), pp. 3633–3650.

[7] Sibnarayan Datta et al. “Next-generation sequencing in clinical virology: Discovery of new viruses”. In: *World journal of virology* 4.3 (2015), p. 265.

[8] Max Kotlyar et al. “Integrated interactions database: tissue-specific view of the human and model organism interactomes”. en. In: (2015). DOI: 10.1093/nar/gkv1115. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4702811%20/.

[9] Woong-Hee Shin et al. “Current Challenges and Opportunities in Designing ProteinProtein Interaction Targeted Drugs”. en. In: (2020). DOI: 10.2147/AABC.S235542. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7669531%20/.

[10] Hitoshi Iuchi et al. “Bioinformatics approaches for unveiling virus-host interactions”. In: *Computational and Structural Biotechnology Journal* 21 (Jan. 2023), pp. 1774–1784. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2023.02.044. URL: https://www.sciencedirect.com/science/article/pii/S2001037023000892.

[11] Li Deng et al. “Viral tagging reveals discrete populations in Synechococcus viral genome sequence space”. en. In: *Nature* 513.7517 (Sept. 2014). Publisher: Nature Publishing Group, pp. 242–245. ISSN: 1476-4687. DOI: 10.1038/nature13459. URL: https://www.nature.com/articles/nature13459.

[12] Dan Liu et al. “Prediction of virus-host associations using protein language models and multiple instance learning”. In: *PLOS Computational Biology* 20.11 (2024), e1012597.

[13] Qiangzhen Yang et al. “Highly accurate protein structure prediction and drug screen of monkeypox virus proteome”. In: *The Journal of Infection* 86.1 (2022), p. 66.

[14] Rachel Seongeun Kim et al. “BFVDa large repository of predicted viral protein structures”. In: *Nucleic Acids Research* 53.D1 (2025), pp. D340–D347.

[15] Dan Liu et al. “PLM-interact: extending protein language models to predict protein-protein interactions”. In: *bioRxiv* (2024), pp. 2024–11.

[16] Samuel Sledzieski et al. “TT3D: Leveraging precomputed protein 3D sequence models to predict proteinprotein interactions”. In: *Bioinformatics* 39.11 (Oct. 2023), btad663. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btad663. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10640393%20/.

[17] The UniProt Consortium. “The Universal Protein Resource (UniProt) 2009”. In: *Nucleic Acids Research* 37.suppl_1 (Jan. 2009), pp. D169–D174. ISSN: 0305-1048. DOI: 10.1093/nar/gkn664. URL: https://doi.org/10.1093/nar/gkn664.

[18] AR Mushegian. “Are there 1031 virus particles on earth, or more, or fewer?” In: *Journal of bacteriology* 202.9 (2020), pp. 10–1128.

[19] Sobia Idrees et al. “Exploring Viral–Host Protein Interactions as Antiviral Therapies: A Computational Perspective”. In: *Microorganisms* 12.3 (2024), p. 630.

[20] Aleksandra Petrovic Fabijan et al. “Phage therapy for severe bacterial infections: a narrative review”. In: *The Medical Journal of Australia* 212.6 (2019), p. 279.

[21] Felicity Aiano et al. “A cross-sectional national investigation of COVID-19 outbreaks in nurseries during rapid spread of the Alpha (B. 1.1. 7) variant of SARS-CoV-2 in England”. In: *BMC public health* 22.1 (2022), p. 1845.

[22] Haibo Yang et al. “The effect of strict lockdown on Omicron SARS-CoV-2 variant transmission in Shanghai”. In: *Vaccines* 10.9 (2022), p. 1392.

[23] Harald zur Hausen. "Papillomavirus infections a major cause of human cancers". In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1288.2 (Oct. 1996), F55–F78. ISSN: 0304-419X. DOI: 10.1016/0304-419X(96)00020-0. URL: https://www.sciencedirect.com/science/article/pii/0304419X96000200.

[24] Mu Gao, Hongyi Zhou, and Jeffrey Skolnick. "Insights into disease-associated mutations in the human proteome through protein structural analysis". In: *Structure* 23.7 (2015), pp. 1362–1369.

[25] D Baltimore. "Expression of animal virus genomes." In: *Bacteriological Reviews* 35.3 (Sept. 1971), pp. 235–241. ISSN: 0005-3678. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC378387%20/.

[26] Moriah L. Szpara and Koenraad Van Doorslaer. "Mechanisms of DNA Virus Evolution". In: *Encyclopedia of Virology* (2021), pp. 71–78. DOI: 10.1016/B978-0-12-809633-8.20993-X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7173537%20/.

[27] Fran Robson et al. "Coronavirus RNA proofreading: molecular basis and therapeutic targeting". In: *Molecular cell* 79.5 (2020), pp. 710–727.

[28] Peter Simmonds et al. "Four principles to establish a universal virus taxonomy". en. In: (Feb. 2023). DOI: 10.1371/journal.pbio.3001922. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001922.

[29] Susan Payne. "Virus evolution and genetics". In: *Viruses* (2017), p. 81.

[30] Gal Shuler and Tzachi Hagai. "Rapidly evolving viral motifs mostly target biophysically constrained binding pockets of host proteins". In: *Cell Reports* 40.7 (2022).

[31] Simon A. Babayan, Richard J. Orton, and Daniel G. Streicker. "Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes". In: *Science* 362.6414 (Nov. 2018). Publisher: American Association for the Advancement of Science, pp. 577–580. DOI: 10.1126/science.aap9072. URL: https://www.science.org/doi/full/10.1126/science.aap9072.

[32] Andrew Kitchen, Laura A Shackelton, and Edward C Holmes. "Family level phylogenies reveal modes of macroevolution in RNA viruses". In: *Proceedings of the National Academy of Sciences* 108.1 (2011), pp. 238–243.

[33] Irene Gil-Farina and Manfred Schmidt. "Interaction of vectors and parental viruses with the host genome". In: *Current opinion in virology* 21 (2016), pp. 35–40.

[34] Janis Doss et al. "A review of phage therapy against bacterial pathogens of aquatic and terrestrial organisms". In: *Viruses* 9.3 (2017), p. 50.

[35] Dimiter S Dimitrov. "Virus entry: molecular mechanisms and biomedical applications". In: *Nature Reviews Microbiology* 2.2 (2004), pp. 109–122.

[36] F Yu and S Mizushima. “Roles of lipopolysaccharide and outer membrane protein OmpC of Escherichia coli K-12 in the receptor function for bacteriophage T4.” In: *Journal of Bacteriology* 151.2 (Aug. 1982), pp. 718–722. ISSN: 0021-9193. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC220313/.

[37] Jun Lan et al. “Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor”. In: *nature* 581.7807 (2020), pp. 215–220.

[38] John P Moore and Robert W Doms. “The entry of entry inhibitors: a fusion of science and medicine”. In: *Proceedings of the National Academy of Sciences* 100.19 (2003), pp. 10598–10602.

[39] Tom Melby and Mike Westby. “Inhibitors of viral entry”. In: *Antiviral Strategies* (2009), pp. 177–202.

[40] John C Tilton and Robert W Doms. “Entry inhibitors in the treatment of HIV-1 infection”. In: *Antiviral research* 85.1 (2010), pp. 91–100.

[41] Francis Crick. “On protein synthesis”. en. In: *Symposia of the Society for Experimental Biology* 12 (1958), pp. 138–63. URL: https://wellcomecollection.org/works/z3d5fnyg/items.

[42] Luis P Villarreal. “Evolution of viruses”. In: *Encyclopedia of virology* (2008), p. 174.

[43] Adi Stern and Raul Andino. “Viral evolution: it is all about mutations”. In: *Viral pathogenesis*. Elsevier, 2016, pp. 233–240.

[44] Katherine LaTourrette and Hernan Garcia-Ruiz. “Determinants of virus variation, evolution, and host adaptation”. In: *Pathogens* 11.9 (2022), p. 1039.

[45] Joanna Masel. “Genetic drift”. In: *Current Biology* 21.20 (2011), R837–R838.

[46] Elinor K Karlsson, Dominic P Kwiatkowski, and Pardis C Sabeti. “Natural selection and infectious disease in human populations”. In: *Nature Reviews Genetics* 15.6 (2014), pp. 379–393.

[47] Magdalena Gayà-Vidal and M Mar Albà. “Uncovering adaptive evolution in the human lineage”. In: *BMC genomics* 15 (2014), pp. 1–12.

[48] Wenhan Shao et al. “Evolution of influenza a virus by mutation and re-assortment”. In: *International journal of molecular sciences* 18.8 (2017), p. 1650.

[49] Alexandros A Irwin Nicholas AT . Pittis, Thomas A Richards, and Patrick J Keeling. “Systematic evaluation of horizontal gene transfer between eukaryotes and viruses”. In: *Nature microbiology* 7.2 (2022), pp. 327–336.

[50] Eric A Franzosa and Yu Xia. “Structural principles within the human-virus protein-protein interaction network”. In: *Proceedings of the National Academy of Sciences* 108.26 (2011), pp. 10538–10543.

[51] Carl Graham et al. “Neutralization potency of monoclonal antibodies recognizing dominant and subdominant epitopes on SARS-CoV-2 Spike is impacted by the B. 1.1. 7 variant”. In: *Immunity* 54.6 (2021), pp. 1276–1289.

[52] Nicholas Magazine et al. “Mutations and evolution of the SARS-CoV-2 spike protein”. In: *Viruses* 14.3 (2022), p. 640.

[53] Antoni G Wrobel. “Mechanism and evolution of human ACE2 binding by SARS-CoV-2 spike”. In: *Current Opinion in Structural Biology* 81 (2023), p. 102619.

[54] David S Schneider and Janelle S Ayres. “Two ways to survive infection: what resistance and tolerance can teach us about treating infectious diseases”. In: *Nature Reviews Immunology* 8.11 (2008), pp. 889–895.

[55] Nisha K Duggal and Michael Emerman. “Evolutionary conflicts between viruses and restriction factors shape immunity”. In: *Nature Reviews Immunology* 12.10 (2012), pp. 687–695.

[56] Norman E Davey, Gilles Travé, and Toby J Gibson. “How viruses hijack cell regulation”. In: *Trends in biochemical sciences* 36.3 (2011), pp. 159–169.

[57] Mattia Ficarelli, Stuart JD Neil, and Chad M Swanson. “Targeted restriction of viral gene expression and replication by the ZAP antiviral system”. In: *Annual Review of Virology* 8.1 (2021), pp. 265–283.

[58] KQ de Andrade and CC Cirne-Santos. *Antiviral activity of zinc finger antiviral protein (ZAP) in different virus families. Pathogens 12: 1461*. 2023.

[59] Jennifer L Meagher et al. “Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences”. In: *Proceedings of the National Academy of Sciences* 116.48 (2019), pp. 24303–24309.

[60] Benjamin D Greenbaum et al. “Patterns of evolution and host gene mimicry in influenza and other RNA viruses”. In: *PLoS pathogens* 4.6 (2008), e1000079.

[61] Dorota Kmiec et al. “CpG frequency in the 5 third of the env gene determines sensitivity of primary HIV-1 strains to the zinc-finger antiviral protein”. In: *MBio* 11.1 (2020), pp. 10–1128.

[62] Christine Mordstein et al. “Transcription, mRNA export, and immune evasion shape the codon usage of viruses”. In: *Genome Biology and Evolution* 13.9 (2021), evab106.

[63] Neetu Tyagi, Rahila Sardar, and Dinesh Gupta. “Natural selection plays a significant role in governing the codon usage bias in the novel SARS-CoV-2 variants of concern (VOC)”. In: *PeerJ* 10 (2022), e13562.

[64] Iris Bahir et al. “Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences”. In: *Molecular systems biology* 5.1 (2009), p. 311.

[65] Azeem Mehmood Butt et al. “Evolution of codon usage in Zika virus genomes is host and vector specific”. In: *Emerging microbes & infections* 5.1 (2016), pp. 1–14.

[66] Cole Maguire et al. “Molecular Mimicry as a Mechanism of Viral Immune Evasion and Autoimmunity”. In: *Nature Communications* 15.1 (Oct. 2024), p. 9403. ISSN: 2041-1723. DOI: 10.1038/s41467-024-53658-8.

[67] Allegra Via et al. “How pathogens use linear motifs to perturb host cell networks”. In: *Trends in biochemical sciences* 40.1 (2015), pp. 36–48.

[68] Tzachi Hagai et al. “Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions”. In: *Cell reports* 7.5 (2014), pp. 1729–1739.

[69] Kim Davey Norman E. Van Roey et al. “Attributes of short linear motifs”. In: *Molecular BioSystems* 8.1 (2012), pp. 268–281.

[70] Anjana Soorajkumar et al. “Computational analysis of short linear motifs in the spike protein of SARS-CoV-2 variants provides possible clues into the immune hijack and evasion mechanisms of omicron variant”. In: *International journal of molecular sciences* 23.15 (2022), p. 8822.

[71] Filip Mihali et al. “Large-scale phage-based screening reveals extensive pan-viral mimicry of host short linear motifs”. In: *Nature Communications* 14.1 (2023), p. 2409.

[72] Leandro Simonetti et al. “SLiM-binding pockets: an attractive target for broad-spectrum antivirals”. In: *Trends in Biochemical Sciences* 48.5 (2023), pp. 420–427.

[73] Michel J Tremblay, Jean-François Fortin, and Réjean Cantin. “The acquisition of host-encoded proteins by nascent HIV-1”. In: *Immunology today* 19.8 (1998), pp. 346–351.

[74] Aspiro Nayim Lin Chih-Yen. Urbina et al. *Virus hijacks host proteins and machinery for assembly and budding, with HIV-1 as an example*. 2022.

[75] Gorka Lasso. Barry Honig and Sagi D. Shapira. “A Sweep of Earths Virome Reveals Host-Guided Viral Protein Structural Mimicry and Points to Determinants of Human Disease”. In: *Cell Systems* 12.1 (2021), 82–91.e3. ISSN: 2405-4712. URL: https://www.sciencedirect.com/science/article/pii/S240547122030363X.

[76] Haiying Lu et al. “Recent advances in the development of proteinprotein interactions modulators: mechanisms and clinical trials”. en. In: *Signal Transduction and Targeted Therapy* 5.1 (Sept. 2020). Publisher: Nature Publishing Group, pp. 1–23. ISSN: 2059-3635. DOI: 10.1038/s41392-020-00315-3. URL: https://www.nature.com/articles/s41392-020-00315-3.

[77] Alex W White, Andrew D Westwell, and Ghali Brahemi. “Protein–protein interactions as targets for small-molecule therapeutics in cancer”. In: *Expert reviews in molecular medicine* 10 (2008), e8.

[78] Levi L Blazer and Richard R Neubig. “Small molecule protein–protein interaction inhibitors as CNS therapeutic agents: current progress and future hurdles”. In: *Neuropsychopharmacology* 34.1 (2009), pp. 126–141.

[79] Tugba G Kucukkal et al. “Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins”. In: *Current opinion in structural biology* 32 (2015), pp. 18–24.

[80] Dapeng Xiong et al. “Implications of disease-related mutations at protein–protein interfaces”. In: *Current opinion in structural biology* 72 (2022), pp. 219–225.

[81] Binh T Vassilev Lyubomir T. Vu et al. “In vivo activation of the p53 pathway by small-molecule antagonists of MDM2”. In: *Science* 303.5659 (2004), pp. 844–848.

[82] Bernhard Suter, Saranya Kittanakom, and Igor Stagljar. “Two-hybrid technologies in proteomics research”. In: *Current Opinion in Biotechnology*. Protein technologies / Systems biology 19.4 (Aug. 2008), pp. 316–323. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2008.06.005. URL: https://www.sciencedirect.com/science/article/pii/S095816690800075X.

[83] Henna Iqbal, Darrin R Akins, and Melisha R Kenedy. “Co-immunoprecipitation for identifying protein-protein interactions in Borrelia burgdorferi”. In: *Borrelia burgdorferi: Methods and Protocols* (2018), pp. 47–55.

[84] Yuen Ho et al. “Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry”. In: *Nature* 415.6868 (2002), pp. 180–183.

[85] Damian Szklarczyk et al. “STRING v11: proteinprotein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D607–D613. ISSN: 0305-1048. DOI: 10.1093/nar/gky1131. URL: https://doi.org/10.1093/nar/gky1131.

[86] Samuel Kerrien et al. “The IntAct molecular interaction database in 2012”. In: *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D841–D846. ISSN: 0305-1048. DOI: 10.1093/nar/gkr1088. URL: https://doi.org/10.1093/nar/gkr1088.

[87] Rose Oughtred et al. “The BioGRID interaction database: 2019 update”. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D529–D541. ISSN: 0305-1048. DOI: 10.1093/nar/gky1079. URL: https://doi.org/10.1093/nar/gky1079.

[88] Samuel Sledzieski et al. “D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions”. en. In: *Cell Systems* 12.10 (Oct. 2021), 969–982.e6. ISSN: 2405-4712. DOI: 10.1016/j.cels.2021.08.010. URL: https://www.sciencedirect.com/science/article/pii/S2405471221003331.

[89] Josh Abramson et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3". en. In: *Nature* 630.8016 (June 2024). Publisher: Nature Publishing Group, pp. 493–500. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07487-w. URL: https://www.nature.com/articles/s41586-024-07487-w.

[90] Yize Li et al. "Pan-cancer proteogenomics connects oncogenic drivers to functional states". In: *Cell* 186.18 (2023), pp. 3921–3944.

[91] Kathrin Muda et al. "Parkinson-related LRRK2 mutation R1441C/G/H impairs PKA phosphorylation of LRRK2 and disrupts its interaction with 14-3-3". In: *Proceedings of the National Academy of Sciences* 111.1 (2014), E34–E43.

[92] Nahid Safari-Alighiarloo et al. "Protein-protein interaction networks (PPI) and complex diseases". In: *Gastroenterology and Hepatology from bed to bench* 7.1 (2014), p. 17.

[93] Katja Luck et al. "A reference map of the human binary protein interactome". In: *Nature* 580.7803 (2020), pp. 402–408.

[94] Adolfo García-Sastre. "Ten strategies of interferon evasion by viruses". In: *Cell host & microbe* 22.2 (2017), pp. 176–184.

[95] Alick Isaacs and Jean Lindenmann. "Virus interference. I. The interferon". In: *Proceedings of the Royal Society of London. Series B-Biological Sciences* 147.927 (1957), pp. 258–267.

[96] William M Schneider, Meike Dittmann Chevillotte, and Charles M Rice. "Interferon-stimulated genes: a complex web of host defenses". In: *Annual review of immunology* 32.1 (2014), pp. 513–545.

[97] Gijs A Versteeg and Adolfo García-Sastre. "Viral tricks to grid-lock the type I interferon system". In: *Current opinion in microbiology* 13.4 (2010), pp. 508–516.

[98] Michael G Katze, Yupeng He, and Michael Gale. "Viruses and interferon: a fight for supremacy". In: *Nature Reviews Immunology* 2.9 (2002), pp. 675–687.

[99] Tomas Doyle, Caroline Goujon, and Michael H Malim. "HIV-1 and interferons: who's interfering with whom?" In: *Nature Reviews Microbiology* 13.7 (2015), pp. 403–413.

[100] Sophia Davidson et al. "IFN λ is a potent anti-influenza therapeutic without the inflammatory side effects of IFN α treatment". In: *EMBO molecular medicine* 8.9 (2016), pp. 1099–1112.

[101] Stephen J Gaudino and Pawan Kumar. "Cross-talk between antigen presenting cells and T cells impacts intestinal homeostasis, bacterial infections, and tumorigenesis". In: *Frontiers in immunology* 10 (2019), p. 360.

[102] Zhongfang Wang et al. "Exposure to SARS-CoV-2 generates T-cell memory in the absence of a detectable viral infection". In: *Nature Communications* 12.1 (2021), p. 1724.

[103] Paul Moss. “The T cell immune response against SARS-CoV-2”. In: *Nature immunology* 23.2 (2022), pp. 186–193.

[104] Haian Fu, Xiulei Mo, and Andrey A Ivanov. “Decoding the functional impact of the cancer genome through protein–protein interactions”. In: *Nature Reviews Cancer* (2025), pp. 1–20.

[105] Derrick E Fouts. “Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences”. In: *Nucleic acids research* 34.20 (2006), pp. 5839–5851.

[106] Raymond HJ Staals and Stan JJ Brouns. “Distribution and mechanism of the type I CRISPR-Cas systems”. In: *CRISPR-Cas systems*. Springer, 2013, pp. 145–169.

[107] Philippe Horvath and Rodolphe Barrangou. “CRISPR/Cas, the immune system of bacteria and archaea”. en. In: *Science* 327.5962 (Jan. 2010), pp. 167–170.

[108] Francesca Young, Simon Rogers, and David L Robertson. “Predicting host taxonomic information from viral genomes: A comparison of feature representations”. In: *PLoS computational biology* 16.5 (2020), e1007894.

[109] Juwen Shen et al. “Predicting protein–protein interactions based only on sequences information”. In: *Proceedings of the National Academy of Sciences* 104.11 (2007), pp. 4337–4341.

[110] Mudita Singhal and Haluk Resat. “A domain-based approach to predict protein-protein interactions”. In: *BMC bioinformatics* 8 (2007), pp. 1–19.

[111] Erli Pang and Kui Lin. “Yeast protein–protein interaction binding sites: prediction from the motif–motif, motif–domain and domain–domain levels”. In: *Molecular BioSystems* 6.11 (2010), pp. 2164–2173.

[112] Alexander Rives et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. In: *Proceedings of the National Academy of Sciences of the United States of America* 118.15 (Apr. 2021), e2016239118. ISSN: 0027-8424. DOI: 10.1073/pnas.2016239118. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8053943/.

[113] Ahmed Elnaggar et al. “Prottrans: Toward understanding the language of life through self-supervised learning”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.10 (2021), pp. 7112–7127.

[114] Zhihan Zhou et al. “Dnabert-2: Efficient foundation model and benchmark for multi-species genome”. In: *arXiv preprint arXiv:2306.15006* (2023).

[115] Jiayang Chen et al. “Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions”. In: *arXiv preprint arXiv:2204.00300* (2022).

[116] Michel Van Kempen et al. “Fast and accurate protein structure search with Foldseek”. In: *Nature biotechnology* 42.2 (2024), pp. 243–246.

[117] KR Chowdhary. “Natural language processing”. In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[118] Marinka Zitnik et al. “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities”. In: *Information Fusion* 50 (2019), pp. 71–91.

[119] Frank Nielsen. “Hierarchical clustering”. In: *Introduction to HPC with MPI for Data Science* (2016), pp. 195–211.

[120] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. “The global k-means clustering algorithm”. In: *Pattern recognition* 36.2 (2003), pp. 451–461.

[121] Andrew P Bradley. “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[122] Jesse Davis and Mark Goadrich. “The relationship between Precision-Recall and ROC curves”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 233–240.

[123] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[124] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.

[125] Oded Maron and Tomás Lozano-Pérez. “A framework for multiple-instance learning”. In: *Advances in neural information processing systems* 10 (1997).

[126] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. “Solving the multiple instance problem with axis-parallel rectangles”. In: *Artificial intelligence* 89.1-2 (1997), pp. 31–71.

[127] Gwenolé Quellec et al. “Multiple-instance learning for medical image and video analysis”. In: *IEEE reviews in biomedical engineering* 10 (2017), pp. 213–234.

[128] PJ Sudharshan et al. “Multiple instance learning for histopathological breast cancer image classification”. In: *Expert Systems with Applications* 117 (2019), pp. 103–111.

[129] Fayyaz Ul Amir Afsar Minhas, Eric D Ross, and Asa Ben-Hur. “Amino acid composition predicts prion activity”. In: *PLoS computational biology* 13.4 (2017), e1005465.

[130] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386.

[131] Franco Scarselli et al. “The graph neural network model”. In: *IEEE transactions on neural networks* 20.1 (2008), pp. 61–80.

[132] Lijia Ma et al. "Multi-view attention graph convolutional networks for the host prediction of phages". In: *Knowledge-Based Systems* 308 (2025), p. 112755.

[133] Jiayu Shang and Yanni Sun. "CHERRY: a Computational metHod for accuratE pRediction of virus–pRokarYotic interactions using a graph encoder–decoder model". In: *Briefings in Bioinformatics* 23.5 (2022), bbac182.

[134] Jonathan M Stokes et al. "A deep learning approach to antibiotic discovery". In: *Cell* 180.4 (2020), pp. 688–702.

[135] S Hochreiter. "Long Short-term Memory". In: *Neural Computation MIT-Press* (1997).

[136] Sho Tsukiyama et al. "LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec". In: *Briefings in bioinformatics* 22.6 (2021), bbab228.

[137] Sumit Madan et al. "Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings". en. In: *Patterns* 3.9 (Sept. 2022), p. 100551. ISSN: 2666-3899. DOI: 10.1016/j.patter.2022.100551. URL: https://www.sciencedirect.com/science/article/pii/S2666389922001568.

[138] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265.

[139] Joseph L Watson et al. "De novo design of protein structure and function with RFdiffusion". In: *Nature* 620.7976 (2023), pp. 1089–1100.

[140] Dzmitry Bahdanau. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[141] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[142] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023), pp. 1123–1130.

[143] Hugo Dalla-Torre et al. "Nucleotide Transformer: building and evaluating robust foundation models for human genomics". In: *Nature Methods* (2024), pp. 1–11.

[144] Kenneth Ward Church. "Word2Vec | Natural Language Engineering". en. In: *Cambridge Core* (). DOI: 10.1017/S1351324916000334. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/word2vec/B84AE4446BD47F48847B4904F0B36

[145] Rico Sennrich. "Neural machine translation of rare words with subword units". In: *arXiv preprint arXiv:1508.07909* (2015).

[146] Mike Schuster and Kaisuke Nakajima. "Japanese and korean voice search". In: *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2012, pp. 5149–5152.

[147] T Kudo. "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing". In: *arXiv preprint arXiv:1808.06226* (2018).

[148] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[149] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[150] Zhenzhong Lan. "Albert: A lite bert for self-supervised learning of language representations". In: *arXiv preprint arXiv:1909.11942* (2019).

[151] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140 (2020), pp. 1–67.

[152] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[153] Yinhan Liu. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* 364 (2019).

[154] Thomas Hayes et al. *Simulating 500 million years of evolution with a language model*. en. July 2024. DOI: 10.1101/2024.07.01.600583. URL: http://biorxiv.org/lookup/doi/10.1101/2024.07.01.600583.

[155] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *nature* 596.7873 (2021), pp. 583–589.

[156] Brian L Hie, Kevin K Yang, and Peter S Kim. "Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins". In: *Cell Systems* 13.4 (2022), pp. 274–285.

[157] Rasko Leinonen et al. "UniProt archive". In: *Bioinformatics* 20.17 (2004), pp. 3236–3237.

[158] Urvashi Khandelwal et al. "Sharp nearby, fuzzy far away: How neural language models use context". In: *arXiv preprint arXiv:1805.04623* (2018).

[159] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv:1908.10084 [cs]. Aug. 2019. URL: http://arxiv.org/abs/1908.10084.

[160] Evelien M Roux Simon. Adriaenssens et al. "Minimum information about an uncultivated virus genome (MIUViG)". In: *Nature biotechnology* 37.1 (2019), pp. 29–37.

[161] Francois Roux Simon. Enault, Bonnie L Hurwitz, and Matthew B Sullivan. "VirSorter: mining viral signal from microbial genomic data". In: *PeerJ* 3 (2015), e985.

[162] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136.

[163] Tomoko Mihara et al. *Linking Virus Genomes with Host Taxonomy*. 2016.

[164] Tatiana Tatusova et al. "RefSeq microbial genomes database: new representation and annotation strategy". In: *Nucleic acids research* 42.D1 (2014), pp. D553–D559.

[165] Cesar O Flores et al. "Statistical structure of host–phage interactions". In: *Proceedings of the National Academy of Sciences* 108.28 (2011), E288–E297.

[166] Asa Ben-Hur and William Stafford Noble. "Choosing negative examples for the prediction of protein-protein interactions". In: *BMC bioinformatics*. Vol. 7. 1. Springer. 2006, pp. 1–6.

[167] Diogo Manuel Carvalho Leite et al. "Computational prediction of inter-species relationships through omics data analysis and machine learning". In: *BMC bioinformatics* 19.14 (2018), pp. 151–159.

[168] Juan Fernando López et al. "Applying one-class learning algorithms to predict phage-bacteria interactions". In: *2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. 2019, pp. 1–6. DOI: 10.1109/LA-CCI47412.2019.9037032.

[169] Siamak Ravanbakhsh Barnabás Póczos Ruslan Salakhutdinov, Alexander J Smola Manzil Zaheer, and Satwik Kottur. "Deep Sets". In: *Advances in Neural Information Processing (NIPS)* (2017).

[170] Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D258–D261.

[171] Philip Jones et al. "InterProScan 5: genome-scale protein function classification". In: *Bioinformatics* 30.9 (2014), pp. 1236–1240.

[172] Guangchuang Yu et al. "ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data". In: *Methods in Ecology and Evolution* 8.1 (2017), pp. 28–36.

[173] Simon Roux et al. "iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes". In: *bioRxiv* (2022).

[174] Ruoshi Zhang et al. "SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts". In: *Bioinformatics* 37.19 (2021), pp. 3364–3366.

[175] Clovis Galiez et al. "WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs". In: *Bioinformatics* 33.19 (2017), pp. 3113–3114.

[176] Congyu Lu et al. "Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics". en. In: *BMC Biol.* 19.1 (Jan. 2021), p. 5.

[177] Nathan A Ahlgren et al. "Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences". en. In: *Nucleic Acids Res.* 45.1 (Jan. 2017), pp. 39–53.

[178] Weili Wang et al. "A network-based integrated framework for predicting virus–prokaryote interactions". In: *NAR genomics and bioinformatics* 2.2 (2020), lqaa044.

[179] Deyvid Amgarten et al. "vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks". Dec. 2020.

[180] Martin Steinegger and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". In: *Nature biotechnology* 35.11 (2017), pp. 1026–1028.

[181] Fan Wu et al. "A new coronavirus associated with human respiratory disease in China". In: *Nature* 579.7798 (2020), pp. 265–269.

[182] Icaro Putinhon Caruso et al. "Insights into the specificity for the interaction of the promiscuous SARS-CoV-2 nucleocapsid protein N-terminal domain with deoxyribonucleic acids". In: *International journal of biological macromolecules* 203 (2022), pp. 466–480.

[183] Tord Berggård, Sara Linse, and Peter James. "Methods for the detection and analysis of proteinprotein interactions". en. In: *PROTEOMICS* 7.16 (2007). _eprint: https://onlinelibrary.wile pp. 2833–2842. ISSN: 1615-9861. DOI: 10.1002/pmic.200700131. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200700131.

[184] Alessia David and Michael J. E. Sternberg. "The Contribution of Missense Mutations in Core and Rim Residues of ProteinProtein Interfaces to Human Disease". In: *Journal of Molecular Biology* 427.17 (Aug. 2015), pp. 2886–2898. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2015.07.004. URL: https://www.sciencedirect.com/science/article/pii/S0022283615003824.

[185] Asa Ben-Hur and William Stafford Noble. "Kernel methods for predicting protein–protein interactions". In: *Bioinformatics* 21.suppl_1 (2005), pp. i38–i46.

[186] Somaye Hashemifar et al. "Predicting proteinprotein interactions through sequence-based deep learning". In: *Bioinformatics* 34.17 (Sept. 2018), pp. i802–i810. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty573. URL: https://doi.org/10.1093/bioinformatics/bty573.

[187] Yan Huang et al. "SGPPI: structure-aware prediction of proteinprotein interactions in rigorous conditions with graph convolutional network". In: *Briefings in Bioinformatics* 24.2 (Mar. 2023), bbad020. ISSN: 1477-4054. DOI: 10.1093/bib/bbad020. URL: https://doi.org/10.1093/bib/bbad020.

[188] Muhao Chen et al. "Multifaceted proteinprotein interaction prediction based on Siamese residual RCNN". In: *Bioinformatics* 35.14 (July 2019), pp. i305–i314. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz328. URL: https://doi.org/10.1093/bioinformatics/btz328.

[189] Logan Hallee and Jason P. Gleghorn. *Protein-Protein Interaction Prediction is Achievable with Large Language Models*. en. Pages: 2023.06.07.544109 Section: New Results. June 2023. DOI: 10.1101/2023.06.07.544109. URL: https://www.biorxiv.org/content/10.1101/2023.06.07.544109v1.

[190] Tristan Bepler and Bonnie Berger. "Learning the protein language: Evolution, structure, and function". In: *Cell systems* 12.6 (2021), pp. 654–669.

[191] Zhidian Zhang et al. *Protein language models learn evolutionary statistics of interacting sequence motifs*. en. Pages: 2024.01.30.577970 Section: New Results. Jan. 2024. DOI: 10.1101/2024.01.30.577970. URL: https://www.biorxiv.org/content/10.1101/2024.01.30.577970v1.

[192] Chai Discovery et al. "Chai-1: Decoding the molecular interactions of life". In: *bioRxiv* (2024), pp. 2024–10.

[193] Damian Szklarczyk et al. "The STRING database in 2023: proteinprotein association networks and functional enrichment analyses for any sequenced genome of interest". In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D638–D646. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1000. URL: https://doi.org/10.1093/nar/gkac1000.

[194] Weizhong Li and Adam Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". eng. In: *Bioinformatics (Oxford, England)* 22.13 (July 2006), pp. 1658–1659. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl158.

[195] Rohit Singh et al. "Topsy-Turvy: integrating a global view into sequence-based PPI prediction". In: *Bioinformatics* 38.Supplement_1 (June 2022), pp. i264–i272. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac258. URL: https://doi.org/10.1093/bioinformatics/btac258.

[196] Florian Richoux et al. *Comparing two deep learning sequence-based models for protein-protein interaction prediction*. arXiv:1901.06268. Jan. 2019. DOI: 10.48550/arXiv.1901.06268. URL: http://arxiv.org/abs/1901.06268.

[197] Eric F. Pettersen et al. "UCSF ChimeraX: Structure visualization for researchers, educators, and developers". In: *Protein Science* 30.1 (Jan. 2021). Publisher: John Wiley & Sons, Ltd, pp. 70–82. ISSN: 0961-8368. DOI: 10.1002/pro.3943. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/pro.3943.

[198] F. Jelinek et al. “Perplexitya measure of the difficulty of speech recognition tasks”. In: *The Journal of the Acoustical Society of America* 62.S1 (Aug. 2005), S63. ISSN: 0001-4966. DOI: 10.1121/1.2016299. URL: https://doi.org/10.1121/1.2016299.

[199] Quinn McNemar. “Note on the sampling error of the difference between correlated proportions or percentages”. en. In: *Psychometrika* 12.2 (June 1947), pp. 153–157. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02295996. URL: http://link.springer.com/10.1007/BF02295996.

[200] Yang Xu Jinrui . Zhang. “How significant is a protein structure similarity with TM-score= 0.5?” In: *Bioinformatics* 26.7 (2010), pp. 889–895.

[201] Jeffrey Zhang Yang . Skolnick. “Scoring function for automated assessment of protein structure template quality”. In: *PROTEINS-NEW YORK-* 68.4 (2007), p. 1020.

[202] Haenig C et al. “Interactome Mapping Provides a Network of Neurodegenerative Disease Proteins and Uncovers Widespread Protein Aggregation in Affected Brains”. en. In: *PubMed* (2020). URL: https://pubmed.ncbi.nlm.nih.gov/32814053/.

[203] Tomohiro Kabuta et al. “Ubiquitin C-terminal hydrolase L1 (UCH-L1) acts as a novel potentiator of cyclin-dependent kinases to enhance cell proliferation independently of its hydrolase activity”. eng. In: *The Journal of biological chemistry* 288.18 (May 2013), pp. 12615–12626. ISSN: 1083-351X. DOI: 10.1074/jbc.m112.435701. URL: https://europepmc.org/articles/PMC3642309.

[204] Arndt Grossmann et al. “Phosphotyrosine dependent proteinprotein interaction network”. In: *Molecular Systems Biology* 11.3 (Mar. 2015). Publisher: John Wiley & Sons, Ltd, p. 794. ISSN: 1744-4292. DOI: 10.15252/msb.20145968. URL: https://www.embopress.org/doi/full/10.15252/msb.20145968.

[205] Tyler M Aten et al. “Tyrosine phosphorylation of the orphan receptor ESDN/DCBLD2 serves as a scaffold for the signaling adaptor CrkL”. eng. In: *FEBS letters* 587.15 (Aug. 2013), pp. 2313–2318. ISSN: 1873-3468. DOI: 10.1016/j.febslet.2013.05.064. URL: https://europepmc.org/articles/PMC3759512.

[206] Janghoo Lim et al. “Opposing effects of polyglutamine expansion on native protein complexes contribute to SCA1”. en. In: (2008). DOI: 10.1038/nature06731. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2377396/.

[207] Daniel R. Rosen et al. “Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis”. en. In: *Nature* 362.6415 (Mar. 1993). Publisher: Nature Publishing Group, pp. 59–62. ISSN: 1476-4687. DOI: 10.1038/362059a0. URL: https://www.nature.com/articles/362059a0.

[208] Broom Wj et al. "SOD1A4V-mediated ALS: absence of a closely linked modifier gene and origination in Asia". en. In: *PubMed* (2008). URL: https://pubmed.ncbi.nlm.nih.gov/18055113/.

[209] R J Daly, M D Binder, and R L Sutherland. "Overexpression of the Grb2 gene in human breast cancer cell lines". eng. In: *Oncogene* 9.9 (Sept. 1994), pp. 2723–2727. ISSN: 1476-5594.

[210] Maro Ohanian et al. "Liposomal Grb2 antisense oligodeoxynucleotide (BP1001) in patients with refractory or relapsed haematological malignancies: a single-centre, open-label, dose-escalation, phase 1/1b trial". en. In: (2018). DOI: 10.1016/S2352-3026(18)30021-8. URL: https://www.thelancet.com/journals/lanhae/article/PIIS2352-3026(18)30021-8/fulltext.

[211] Katsumi Koshikawa et al. "Significant up-regulation of a novel gene, CLCP1, in a highly metastatic lung cancer subline as well as in lung cancers in vivo". en. In: *Oncogene* 21.18 (Apr. 2002). Publisher: Nature Publishing Group, pp. 2822–2828. ISSN: 1476-5594. DOI: 10.1038/sj.onc.1205405. URL: https://www.nature.com/articles/1205405.

[212] Ryan M. Ames et al. "Binding interface change and cryptic variation in the evolution of protein-protein interactions". In: *BMC Evolutionary Biology* 16 (Feb. 2016), p. 40. ISSN: 1471-2148. DOI: 10.1186/s12862-016-0608-1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4758157%20/.

[213] Jonathan Frazer et al. "Disease variant prediction with deep generative models of evolutionary data". en. In: *Nature* 599.7883 (Nov. 2021). Publisher: Nature Publishing Group, pp. 91–95. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04043-8. URL: https://www.nature.com/articles/s41586-021-04043-8.

[214] Jun Cheng et al. "Accurate proteome-wide missense variant effect prediction with AlphaMissense". In: *Science* 381.6664 (Sept. 2023). Publisher: American Association for the Advancement of Science, eadg7492. DOI: 10.1126/science.adg7492. URL: https://www.science.org/doi/10.1126/science.adg7492.

[215] Andre Cornman et al. "The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling". en. In: (Oct. 2024). Publisher: Cold Spring Harbor Laboratory Section: New Results. DOI: 10.1101/2024.08.14.607850. URL: https://www.biorxiv.org/content/10.1101/2024.08.14.607850v2.

[216] Bin Shao and Jiawei Yan. "A long-context language model for deciphering and generating bacteriophage genomes". en. In: *Nature Communications* 15.1 (Oct. 2024). Publisher: Nature Publishing Group, p. 9392. ISSN: 2041-1723. DOI: 10.1038/s41467-024-53759-4. URL: https://www.nature.com/articles/s41467-024-53759-4.

[217] Rohit K Jangra et al. *Influence of Protein-Protein Interactions (PPIs) on the Outcome of Viral Infections*. 2022.

[218] Max Kotlyar et al. "IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis". In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D640–D647. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1034. URL: https://doi.org/10.1093/nar/gkab1034.

[219] Anderson F. Brito and John W. Pinney. "ProteinProtein Interactions in VirusHost Systems". In: *Frontiers in Microbiology* 8 (2017). ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.01557. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.01557.

[220] Aidi Zhang, Libo He, and Yaping Wang. "Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions". In: *BMC bioinformatics* 18 (2017), pp. 1–13.

[221] Gorka Lasso et al. "A structure-informed atlas of human-virus interactions". In: *Cell* 178.6 (2019), pp. 1526–1541.

[222] Tobias Hamp and Burkhard Rost. "Evolutionary profiles improve protein–protein interaction prediction from sequence". In: *Bioinformatics* 31.12 (2015), pp. 1945–1950.

[223] Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents". In: *International conference on machine learning*. PMLR. 2014, pp. 1188–1196.

[224] Xiaodi Yang et al. "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method". In: *Computational and structural biotechnology journal* 18 (2020), pp. 153–161.

[225] Mais G. Ammari et al. "HPIDB 2.0: a curated database for hostpathogen interactions". In: *Database* 2016 (Jan. 2016), baw103. ISSN: 1758-0463. DOI: 10.1093/database/baw103. URL: https://doi.org/10.1093/database/baw103.

[226] Thibaut Guirimand, Stéphane Delmotte, and Vincent Navratil. "VirHostNet 2.0: surfing on the web of virus/host molecular interactions data". In: *Nucleic Acids Research* 43.Database issue (Jan. 2015), pp. D583–D587. ISSN: 0305-1048. DOI: 10.1093/nar/gku1121. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383936%20/.

[227] Andreas Zanzoni et al. "MINT: a Molecular INTeraction database". en. In: *FEBS Letters* 513.1 (2002). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/S0014-5793%2801%2903 8, pp. 135–140. ISSN: 1873-3468. DOI: 10.1016/S0014-5793(01)03293-8. URL: https://onlinelibrary.wiley.com/doi/abs/10.1016/S0014-5793%2801%2903293-8.

[228] Sandra Orchard et al. "The MIntAct projectIntAct as a common curation platform for 11 molecular interaction databases". In: *Nucleic acids research* 42.D1 (2014), pp. D358–D363.

[229] Xiaodi Yang et al. "HVIDB: a comprehensive database for humanvirus proteinprotein interactions". In: *Briefings in Bioinformatics* 22.2 (Mar. 2021), pp. 832–844. ISSN: 1477-4054. DOI: 10.1093/bib/bbaa425. URL: https://doi.org/10.1093/bib/bbaa425.

[230] Yungki Park and Edward M Marcotte. "Revisiting the negative example sampling problem for predicting protein–protein interactions". In: *Bioinformatics* 27.21 (2011), pp. 3024–3028.

[231] Yungki Park and Edward M. Marcotte. "Flaws in evaluation schemes for pair-input computational predictions". en. In: *Nature Methods* 9.12 (Dec. 2012). Number: 12 Publisher: Nature Publishing Group, pp. 1134–1136. ISSN: 1548-7105. DOI: 10.1038/nmeth.2259. URL: https://www.nature.com/articles/nmeth.2259.

[232] Tobias Hamp and Burkhard Rost. "More challenges for machine-learning protein interactions". In: *Bioinformatics* 31.10 (2015), pp. 1521–1525.

[233] Ivan Rep and Vladimir eperi. *Boosting the Performance of Transformer Architectures for Semantic Textual Similarity*. arXiv:2306.00708. June 2023. DOI: 10.48550/arXiv.2306.00708. URL: http://arxiv.org/abs/2306.00708.

[234] Baris E. Suzek et al. "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches". In: *Bioinformatics* 31.6 (Mar. 2015), pp. 926–932. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu739. URL: https://doi.org/10.1093/bioinformatics/btu739.

[235] Norman Goodacre et al. "Protein-protein interactions of human viruses". In: *Seminars in cell & developmental biology*. Vol. 99. Elsevier. 2020, pp. 31–39.

[236] Lakshminarayan M Iyer, L Aravind, and Eugene V Koonin. "Common origin of four diverse families of large eukaryotic DNA viruses". In: *Journal of virology* 75.23 (2001), pp. 11720–11734.

[237] Alexey Stukalov et al. "Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV". en. In: *Nature* 594.7862 (June 2021). Publisher: Nature Publishing Group, pp. 246–252. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03493-4. URL: https://www.nature.com/articles/s41586-021-03493-4.

[238] Li J et al. "Virus-Host Interactome and Proteomic Survey Reveal Potential Virulence Factors Influencing SARS-CoV-2 Pathogenesis". en. In: *PubMed* (2021). URL: https://pubmed.ncbi.nlm.nih.gov/32838362/.

[239] Edward C Holmes. "What does virus evolution tell us about virus origins?" In: *Journal of virology* 85.11 (2011), pp. 5247–5251.

[240] Nadav Brandes et al. "ProteinBERT: a universal deep-learning model of protein sequence and function". In: *Bioinformatics* 38.8 (2022), pp. 2102–2110.

[241] Eric Nguyen et al. “Sequence modeling and design from molecular to genome scale with Evo”. In: *Science* 386.6723 (2024), eado9336.

[242] Haopeng Yu et al. “An interpretable RNA foundation model for exploring functional RNA motifs in plants”. In: *Nature Machine Intelligence* 6.12 (2024), pp. 1616–1625.

[243] Ali Saberi et al. “A long-context RNA foundation model for predicting transcriptome architecture”. In: *bioRxiv* (2024), pp. 2024–08.

[244] Evolutionary Scale. “GESM Cambrian: Revealing the mysteries of proteins with unsupervised learning.” In: (2024).

[245] Alexey Strokach and Philip M Kim. “Deep generative modeling for protein design”. In: *Current opinion in structural biology* 72 (2022), pp. 226–236.

[246] Xiangxiang Zeng et al. “Deep generative molecular design reshapes drug discovery”. In: *Cell Reports Medicine* 3.12 (2022).

[247] Sarah Alamdari et al. “Protein generation with evolutionary diffusion: sequence is all you need”. In: *BioRxiv* (2023), pp. 2023–09.

[248] Minkyung Baek et al. “Accurate prediction of protein structures and interactions using a three-track neural network”. In: *Science* 373.6557 (2021), pp. 871–876.