Briefings in Bioinformatics, 00(00), 2021, 1–16

doi: 10.1093/bib/bbab029
Method Review

Current status and future perspectives of computational studies on human–virus protein–protein interactions

Xianyi Lian, Xiaodi Yang, Shiping Yang and Ziding Zhang

Corresponding author: Ziding Zhang, State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. Tel.: +86 10 62734376, E-mail: zidingzhang@cau.edu.cn; Shiping Yang, State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193, China. Tel.: +86 10 62733780, E-mail: shi_ping_yang@163.com

Abstract

The protein–protein interactions (PPIs) between human and viruses mediate viral infection and host immunity processes. Therefore, the study of human–virus PPIs can help us understand the principles of human–virus relationships and can thus guide the development of highly effective drugs to break the transmission of viral infectious diseases. Recent years have witnessed the rapid accumulation of experimentally identified human–virus PPI data, which provides an unprecedented opportunity for bioinformatics studies revolving around human–virus PPIs. In this article, we provide a comprehensive overview of computational studies on human–virus PPIs, especially focusing on the method development for human–virus PPI predictions. We briefly introduce the experimental detection methods and existing database resources of human–virus PPIs, and then discuss the research progress in the development of computational prediction methods. In particular, we elaborate the machine learning-based prediction methods and highlight the need to embrace state-of-the-art deep-learning algorithms and new feature engineering techniques (e.g. the protein embedding technique derived from natural language processing). To further advance the understanding in this research topic, we also outline the practical applications of the human–virus interactome in fundamental biological discovery and new antiviral therapy development.

Key words: human–virus relationship; database; prediction; machine learning; network analysis; drug development

Introduction

Viruses are extremely tiny microorganisms that contain a core of genetic material, either RNA or DNA, wrapped in protein capsids [1]. They can only reproduce themselves by attaching and entering host cells and then hijacking the host cells' metabolic machinery [2]. Viral infections cause many human diseases and can even become serious threats to global health. In modern times, viruses have triggered three major pandemics: the 1918–1919 'Spanish flu' epidemic [3], which killed 20–40 million people, the acquired immune deficiency syndrome (AIDS) epidemic [4], which killed an estimated 1.5 million people worldwide in 2013, and the ongoing high-transmissible Coronavirus Disease-2019 (COVID-19) epidemic [5] with more than 82 million cases including 1 818 849 deaths by 1 February 2021 (https://covid19.who.int/).

Xianyi Lian is a PhD student at the State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. Her current research interests include protein bioinformatics and machine learning.
Xiaodi Yang is a PhD student at the State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. Her current research interests include protein bioinformatics and machine learning.
Shiping Yang is a postdoctoral fellow at the State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University. His current research interests include protein bioinformatics and plant genomics.
Ziding Zhang is a professor at the State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. His research interests are protein bioinformatics and systems biology.
Submitted: 10 November 2020; **Received (in revised form):** 14 January 2021

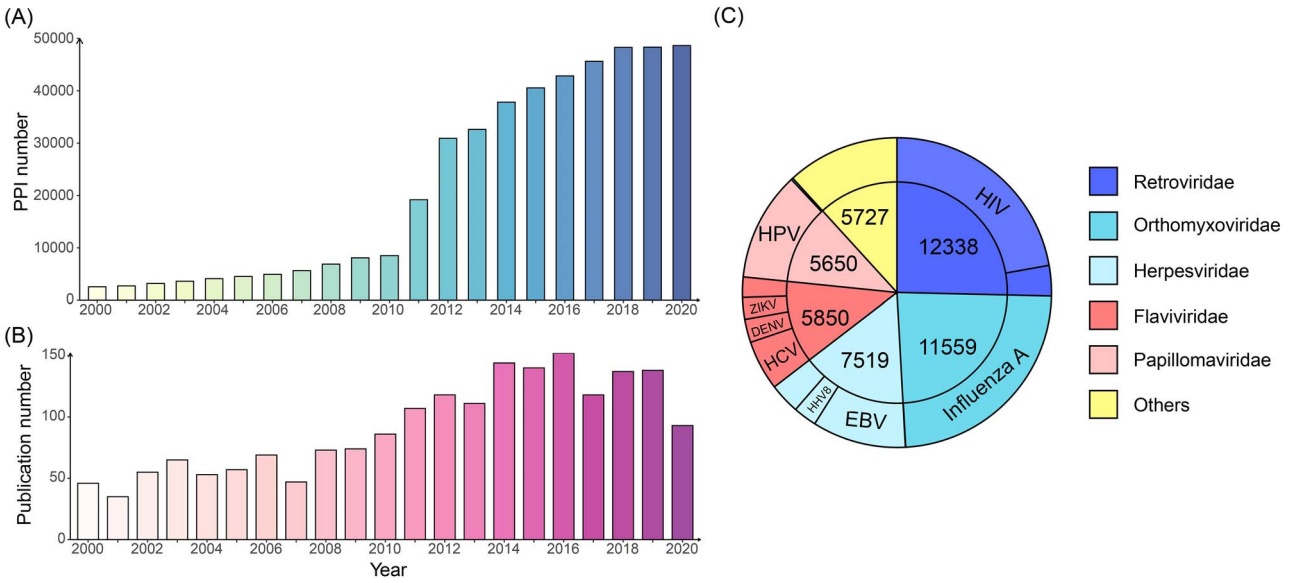

Figure 1. Numbers of experimentally identified human–virus PPIs and publications associated with human–virus PPIs in the past 20 years. (A) The human–virus PPIs were downloaded and integrated from four databases (HPIDB, PHISTO, VirHostNet and VirusMentha; version 2020-03) for statistical analysis. (B) The number of publications was counted by searching the keywords associated with human–virus PPIs in the PubMed database (https://pubmed.ncbi.nlm.nih.gov/, version 2020-08). (C) The numbers of known human–virus PPIs of the top five virus families in 2020 and one or several species with the largest number of PPIs in each virus family are shown.

The protein–protein interactions (PPIs) between human and viruses are a crucial entry point for deciphering complicated human–virus relationships. Recent advances in high-throughput technologies have fueled large-scale mapping of human–virus PPIs (Figure 1), but the current data remain far from sufficient for establishing a complete human–virus PPI network (also termed the human–virus interactome). Computational biologists are also working hard to develop predictive models to dramatically accelerate the completeness and soundness of the human–virus PPI network, which contributes to an improved understanding of viral infection mechanisms, receptor discovery, host cell tropism investigation and drug target discovery. Cost-effective prediction methods are also increasingly useful for fighting against sudden outbreaks of new viruses, such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The potential PPIs between human and new viruses can be quickly predicted to aid the initiation of drug design and vaccine development without any delay [6, 7].

In this review, the current research status of the human–virus interactome is summarized based on experimental methods, public databases, computational prediction approaches and biological applications (Figure 2A). In particular, human–virus PPI prediction methods based on machine learning (ML) are emphasized and elaborated. Moreover, future trends and perspectives in this research area are outlined to provide more hints for researchers in the future.

Experimental methods for detecting human–virus PPIs

In general, experimental approaches for the determination of PPIs are mainly based on biochemical assays, genetic assays and structural biology approaches [8], which can be simply divided into two categories according to the experimental scale, i.e. low- and high-throughput detection methods [8–10]. Different experimental methods have their own advantages and limitations. Low-throughput experiments (e.g. isothermal titration calorimetry [11], pulldown assays [12] and surface plasmon resonance [13]) can detect few but high-quality PPIs, whereas high-throughput methods identify PPIs in large quantities but with a higher false positive rate. Among the high-throughput assays, the classical yeast two-hybrid (Y2H) system [14, 15] has become the most widely used method for detecting PPIs. Another popular high-throughput technique is affinity purification coupled with mass spectrometry (AP-MS) [16, 17]. In parallel with the development of high-throughput detection techniques, large-scale human–virus PPI studies using Y2H or AP-MS approaches have been performed for Epstein-Barr virus (EBV) [18], hepatitis C virus (HCV) [19–22], influenza A (H1N1) virus [23], dengue virus (DENV) [24, 25], herpes simplex virus type-1 (HSV-1) [26, 27], human papillomavirus (HPV) [28–30], human immunodeficiency virus type 1 (HIV-1) [31], Ebola virus [32], Zika virus (ZIKV) [25] and SARS-CoV-2 [33,34] (Table 1).

In view of the inevitability of false positives and false negatives in experimental detection, certain data filtering methods should be adopted to obtain relatively high-quality data and thus ensure that subsequent applications will be more meaningful. Although several scoring methods have been developed to assess the reliability of experimental PPIs [35–38], these methods were mainly proposed for intraspecies PPIs, and their usability in the quality assessment of human–virus PPIs remains unclear.

Current public databases of human–virus PPIs

With the accumulation of experimentally determined human–virus PPI data, several public database resources have been established to store and manage human–virus PPI data for the community. According to the coverage of virus species, the databases can be divided into two categories (Table 2). The first category consists of species-specific databases covering PPI data from only one specific viral species and includes HCVpro [39], NCBI HIV-1 Human Interaction Database [40], DenHunt [41], DenvInt [42] and ZikaBase [43]. The other includes pan-species databases based on a wider range of viral species, such

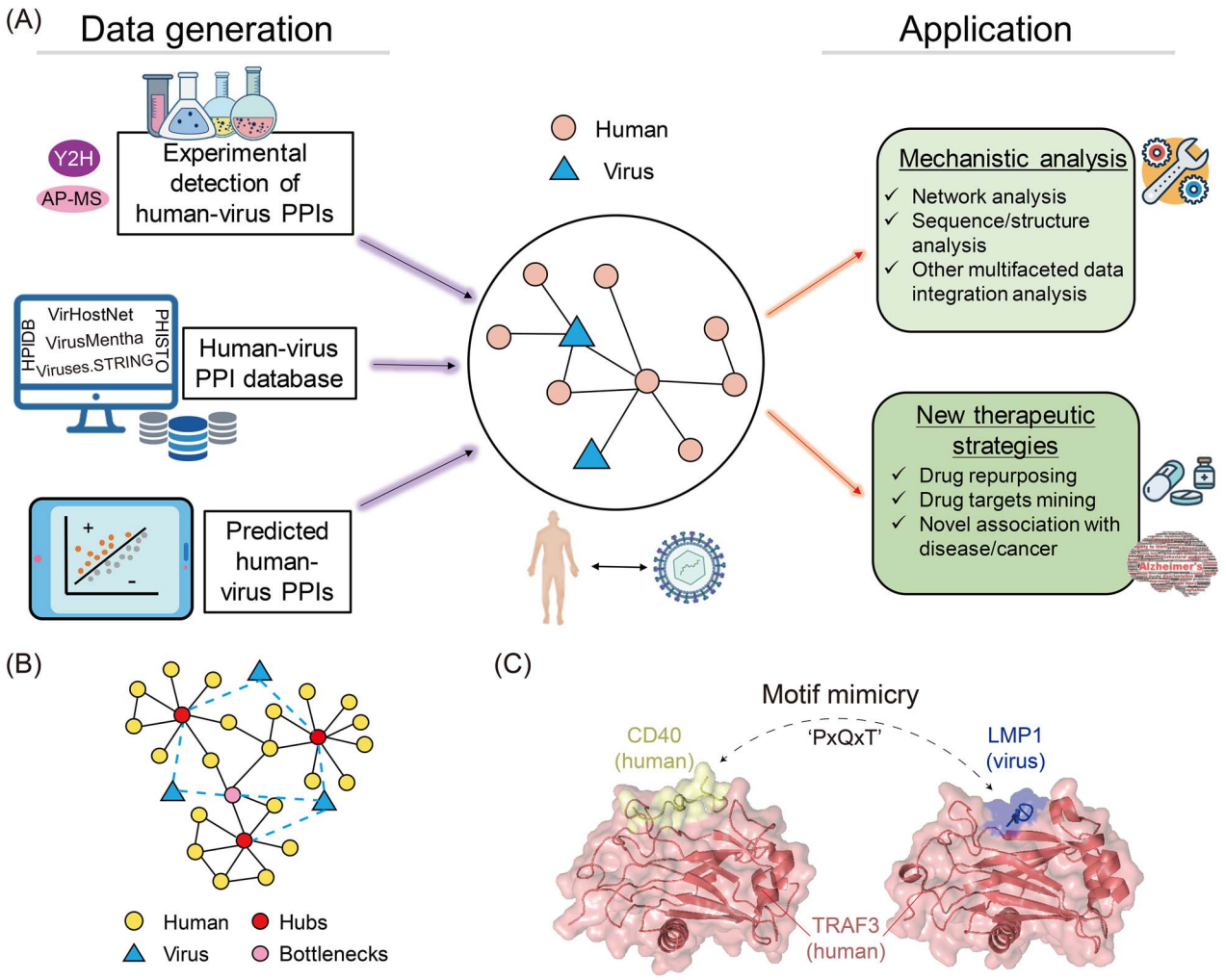

Figure 2. Overview of research on human–virus PPIs. (A) The left panel describes the generation and management of human–virus PPI data, including experimental identifications, database resources and computational predictions, and the right panel shows the applications of human–virus PPI data related to obtaining a mechanistic understanding of human–virus relationships and the development of new therapeutic strategies. (B) Viral proteins tend to target hub or bottleneck proteins in the human PPI network. (C) Simplified schema representing examples of viral mimicry of human motifs. The LMP1 protein of EBV mimics the motif ('PxQxT') of CD40 to interact with TRAF3; CD40 is the endogenous interacting partner of TRAF3.

Table 1. Large-scale experimental detection of human–virus PPIs

Virus	Viral type	Technique	Reference
EBV	DNA virus	Y2H	Calderwood *et al.*, 2007 [18]
HCV	RNA virus	Y2H, AP-MS	de Chassey *et al.*, 2008 [19]; Dolan *et al.*, 2013 [20]; Germain *et al.*, 2014 [21]; Ramage *et al.*, 2015 [22]
H1N1	RNA virus	Y2H	Shapira *et al.*, 2009 [23]
DENV	RNA virus	Y2H, AP-MS	Khadka *et al.*, 2011 [24]; Shah *et al.*, 2018 [25]
HSV-1	DNA virus	Y2H, AP-MS	Pichlmair *et al.*, 2012 [26]; Griffiths *et al.*, 2013 [27]
HPV	DNA virus	Y2H, AP-MS	Rozenblatt-Rosen *et al.*, 2012 [28]; White *et al.*, 2012 [29]; Eckhardt *et al.*, 2018 [30]
HIV-1	RNA virus	AP-MS	Jager *et al.*, 2012 [31]
EBOV	RNA virus	AP-MS	Batra *et al.*, 2018 [32]
ZIKV	RNA virus	AP-MS	Shah *et al.*, 2018 [25]
SARS-CoV-2	RNA virus	AP-MS	Gordon *et al.*, 2020 [33]; Li *et al.*, 2020 [34]

as VirHostNet [44], VirusMentha [45], HPIDB [46], PHISTO [47] and Viruses.STRING [48]. In general, the human–virus PPIs deposited in these public databases were mainly integrated from other comprehensive PPI databases using automatic integration tools (e.g. PSICQUIC [49]) or manually collected from the published literature.

Table 2. Overview of host-pathogen PPI databases containing human–virus PPIs

Database	URL	Pathogen	Host
HCVpro	https://www.cbrc.kaust.edu.sa/hcvpro/	HCV	Human
NCBI HIV-1	http://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions	HIV-1	Human
DenHunt	http://proline.biochem.iisc.ernet.in/DenHunt/	DENV	Human
DenvInt	https://denvint.000webhostapp.com/index.html	DENV	Human, mosquito
HVIDB	http://zzdlab.com/hvidb/	Viruses	Human
VirHostNet	http://virhostnet.prabi.fr	Viruses	Human, animal, plant
Viruses.STRING	http://viruses.string-db.org/	Viruses	All hosts (including human)
VirusMentha	http://virusmentha.uniroma2.it/	Viruses	All hosts (including human)
HPIDB	https://hpidb.igbb.msstate.edu/index.html	Viruses, bacteria, fungi	Human, animal, plant
PHISTO	http://www.phisto.org	Viruses, bacteria, fungi, protozoa	Human

Regarding the species-specific databases, HCVpro [39] is an HCV-specific knowledge base devoted to housing HCV intraspecies PPIs and human–HCV interspecies PPIs. Additionally, this database provides abundant annotation information on PPIs from a variety of cross-referenced data resources. The NCBI HIV-1 Human Interaction Database [40] deposits all known interaction information between HIV-1 and the human host, including human–HIV-1 PPIs, human genes that have been reported to affect viral replication and infectivity, and proteins from disease organisms associated with HIV/AIDS. DenHunt [41] is a database designed for human–DENV PPIs, which not only contains PPIs but also includes human genes that are differentially expressed under DENV infections. DenvInt [42] focuses on storing PPIs between DENV and its hosts (human and mosquitoes), and this database also stores DENV intraspecies PPIs. ZikaBase [43] curates the human–ZIKV PPIs from the published literature and stores some attributes, such as differentially expressed genes, pathway information, and available 3D structures of ZIKV proteins.

Among the pan-species databases, VirHostNet [44] is one of the earliest databases containing human–virus PPIs. This database focuses on host–virus, virus–virus and host–host PPI networks, and provides visualization of these PPI networks. VirusMentha [45], which is an extension of VirusMINT [50], stores virus–virus and host–virus PPI data. As a comprehensive host–virus PPI resource, VirusMentha is regularly updated each week. HPIDB [46] is a resource for host-pathogen interactions, including human–virus PPIs. In addition to providing more annotation information for each PPI, this database provides an online BLAST tool for searching homologous human–virus PPIs. It is worth noting that only HPIDB provides PPI data in standardized PSI-MI format [51] among the aforementioned databases. PHISTO [47] is also a comprehensive platform with information about host-pathogen interactions and contains a wealth of analysis tools, such as visualizing the PPI networks and analyzing the network properties of virally targeted human proteins. Viruses.STRING [48] is a derivative version of the popular intraspecies PPI database STRING [52] that only focuses on the intravirus and host-virus PPIs. In this database, multiple lines of evidence (including experiments and predictions) to infer a host–virus PPI are combined to obtain a confidence score that ultimately represents the possibility of the PPI.

Although the human–virus PPI data in these databases are increasingly available, the majority of the experimental PPIs are only from a few virus species (Figure 1C). Indeed, the current human–virus databases still have room for improvement. First, the databases should not be designed only for specific virus species/strains; otherwise, they will not be able to meet the needs of a wider user group. Second, these databases lack comprehensive multidimensional data to facilitate further analysis. Third, they do not provide online predictors of human–virus PPIs. For these reasons, we have recently developed a new human–virus PPI database called HVIDB (http://zzdlab.com/hvidb) [53] with the purpose of providing more comprehensive annotations associated with known human–virus PPIs.

Computational methods of predicting human–virus PPIs

Despite the increasing number of experimentally identified human–virus PPI data, the current human–virus interactome remains incomplete. In this context, computational prediction methods are becoming increasingly important to supplement experimental efforts. The existing prediction methods include interolog mapping [54, 55], domain–domain interaction (DDI)-based inference [56], domain–motif interaction (DMI)-based inference [57], structure-based method [58] and ML-based method [59] (Figure 3A). Briefly, the main idea of interolog mapping is to infer unknown PPIs from known homologous PPIs (termed interologs); the DDI-based method relies on the detection of the interacting domain pairs in the query protein pair to infer the potential interaction [60,61]. Here, we only focus on describing the DMI-, structure- and ML-based methods for predicting human–virus PPIs.

DMI-based method

Compared with domains, motifs, which are short functional sequence segments, also play a role in regulating PPIs. Indeed, many PPIs are mediated by DMIs, where a domain in one protein binds to a short linear motif in the other protein [62]. Increasing lines of evidence shows that viral proteins commonly bind to domains in human proteins via motifs [63], which mimic the motifs in the endogenous interacting partners of virally binding human proteins [64–66] (Figure 2C). Motif mimicry is a tactic used by viruses to rapidly utilize human proteins to achieve self-replication or reduce the triggering of host immune responses [67, 68]. Therefore, the DMI-based method has been more widely applied for human–virus PPI predictions than the DDI-based method. The basic principle of this method is to first identify the domains of the query human protein and the motifs of the query viral protein, and then determine the interaction probability based on the occurrence of known DMIs between

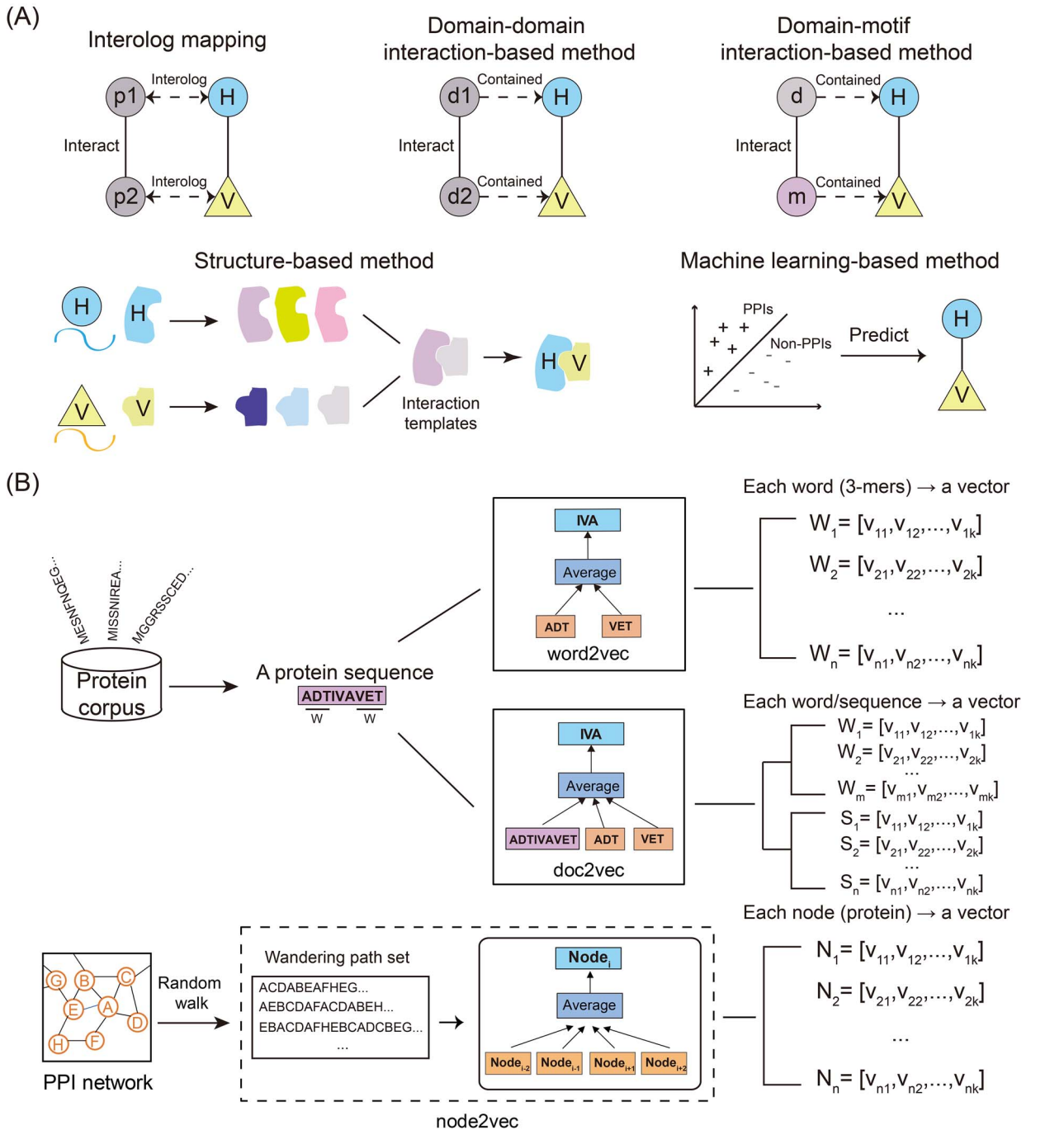

Figure 3. Graphical illustrations of human–virus PPI prediction methods and protein embedding techniques. (A) Schematic diagram of human–virus PPI prediction methods, including interolog mapping, DDI-based method, DMI-based method, structure-based method and ML-based method. The blue circle represents a human protein, and the grass green triangle represents a viral protein. (B) Three protein embedding algorithms. The CBOW architecture is used as an example to illustrate the use of word2vec, doc2vec and node2vec for inferring word vectors. For sequence embedding, the protein sequence is first broken into small k-mers (here $k = 3$). The word2vec model learns to predict the vectors for center 3-mers from their context 3-mers, whereas the doc2vec model learns to predict the vectors for center 3-mers not only from their context 3-mers but also from the whole protein sequence. In terms of node2vec, the node paths generated by random walks in the human PPI network compose sequences that will be further inputted into the word2vec model.

the protein pair. The ELM [69] and 3did databases [70] are two public resources that store known DMIs. The motifs in viral proteins are usually searched programmatically based on the motif patterns appearing in these known DMIs. The domains of human proteins are assigned by searching against domain databases (e.g. PROSITE [71] or Pfam [72]). The published studies using the DMI-based method are detailed in Table 3.

It is worth noting that some filtering approaches are necessary based on the consideration that only a few motifs present in the proteins are actually involved in the interactions. Evans *et al.* [57] first detected conserved motifs that are conserved at a rate of at least 70% in different HIV-1 subtypes and then predicted which human proteins can be targeted by viral proteins based on the domains that bind to these motifs. This conservative

Table 3. Summary of existing studies using the DMI-based method

Authors	Year	Identification of domains in human proteins	**Source of DMI**	**Application**
Evans *et al.* [57]	2009	The PROSITE scan tool with default parameters was used to annotate the PROSITE domains.	ELM database	Human–HIV-1
Becerra *et al.* [75]	2017	The Pfam database was downloaded to annotate the Pfam domains.	ELM database	Human–HIV-1
Chiang *et al.* [73]	2017	Not described.	ELM database	Human–HCV
García-Pérez *et al.* [74]	2018	The Pfam database was downloaded to annotate the Pfam domains.	3did database	Human–influenza A virus
Lian *et al.* [76]	2020	The hmmscan tool was used to annotate the Pfam domains.	3did database	Human–HSV-1

motif assignment strategy has also been used in other studies [73,74]. Becerra *et al.* [75] developed three filtering methods to obtain linear motif sets that are (i) conserved in viral proteins (C), (ii) located in disordered regions (D) and (iii) rare or scarce in a set of randomized viral sequences (R). The performance based on the union and intersection sets of these three sets (C, D and R) was further examined. In our previous work [76], the highly frequently occurring motifs were also filtered. One should bear in mind that the DMI-based method can only capture human–virus PPI types mediated by DMI; thus, the coverage of the predicted PPIs is somehow limited. Moreover, if the motifs are not effectively filtered, false positives are also easily generated. Currently, various motif identification algorithms have been developed [77], and these makes reliable sequence motif detection more convenient for researchers.

Structure-based method

The 3D structural information of proteins provides an intuitive understanding of protein functions, which has been applied in two main types of structure-based PPI prediction methods. The first type of method called protein docking predicts the interaction details between two interacting proteins whose 3D structures are available and selects the most likely binding mode as the predicted protein complex structure based on the binding energy score [78]. Although the biological applications of protein docking are important, this method is beyond the scope of the current review. The aim of the second type is to employ the structural properties for PPI prediction, and the results are expected to provide additional information in comparison to conventional protein sequence-based predictions.

A commonly used prediction strategy is the so-called 'interaction redundancy'. The main idea is that two structurally similar proteins tend to share the same interaction partners. Doolittle *et al.* [79] employed this approach to predict human–HIV-1 PPIs between nine HIV-1 proteins and human proteins with known PPIs in the human PPI network. The PPI templates used in that paper are known human PPIs with structural complexes. First, these researchers identified human proteins sharing regions of high structural similarity to an HIV-1 protein as 'HIV-like' proteins. If these 'HIV-like' human proteins interact with other human proteins ('targets'), the corresponding HIV-1 protein was predicted to interact with the 'targets'. Based on the same idea, the researchers then predicted PPIs between human and DENV [80]. de Chassey *et al.* [81] further implemented this idea using not only known human PPIs but also human–virus PPIs as templates to determine novel human–virus PPIs. In addition to this 'interaction redundancy' evidence, the P-HIPSTer model developed by Lasso *et al.* [82] also integrated the predicted human–virus PPIs mediated by structural DDIs or DMIs. Notably, all the above methods require the available structural information of human and virus protein pairs. Although the 3D structures of many proteins can be determined by homology modeling, accurate structural prediction of all proteins remains an unsolved issue, which limits the applications of structure-informed human–virus PPI prediction methods.

ML-based method

Although some rules and patterns governing interactions between human and viruses have been captured from known human–virus PPIs, numerous hidden features cannot be discovered through simple statistical analysis. ML-based prediction methods have advantages in addressing this issue and have also been flourishing in human–virus PPI prediction over the past decade (Table 4). As an important branch of artificial intelligence, ML can automatically generate models that can analyze large-scale complicated data and provide more accurate prediction results. From the algorithmic point of view, the human–virus PPI prediction task can be regarded as a binary classification problem for which supervised learning is commonly employed. In other words, ML-based approaches build predictors based on training data consisting of interacting protein pairs (PPIs) and noninteracting protein pairs (non-PPIs) (Figure 4A). Several major factors, such as data sufficiency, data quality, negative sample selection, feature extraction methods and ML algorithms, would affect the performance of ML models. We mainly summarize and discuss these issues as follows.

Sample selection

Data sufficiency and data quality. A sufficient number of training samples are a prerequisite to acquiring a reasonable ML model. Compared with intraspecies PPIs, data scarcity is more serious in the prediction of human–virus PPIs. Initially, ML-based human–virus PPI predictions mainly focused on some viruses of high concern (e.g. HIV) due to the abundant experimental PPIs. In general, a sufficient amount of human–virus PPI data should be collected to train the predictor. For viruses with a small amount of data, transfer learning can be used to borrow useful data, information or models from homologous species [83–85].

Table 4. Summary of existing ML-based human–virus PPI prediction methods

Authors	Features	Species	ML method	Negative sampling	Year
Tastan *et al.* [59]	Sequence (sequence similarity), network, biological function (GO, domain-motif, post-translational modification) and expression (gene expression, tissue expression) features	Human–HIV-1	RF	Random sampling	2009
Dyer *et al.* [93]	Sequence, network and biological function (domain) features	Human–HIV	SVM	Random sampling	2011
Cui *et al.* [94]	Sequence features	Human–HPV/HCV	SVM	Random sampling	2012
Mei *et al.* [85]	Biological function (GO) features	Human–HIV-1	SVM + probability weighted ensemble transfer learning model	Combination of random sampling and exclusiveness of subcellular colocalized proteins	2013
Barman *et al.* [96]	Sequence, network and biological function (domain–domain association) features	Human–virus	SVM	Random sampling	2014
Emamjomeh *et al.* [97]	Sequence, network, biological function (post-translational modification), expression (tissue expression) and evolutionary information features	Human–HCV	Ensemble learning based on SVM, RF, NB and MLP	Random sampling	2014
Eid *et al.* [87]	Sequence features	Human–virus	SVM	Dissimilarity-based negative sampling	2016
Yang *et al.* [112]	Protein embeddings	Human–virus	RF	Dissimilarity-based negative sampling	2020
Lian *et al.* [76]	Sequence and network features	Human–HSV-1	RF	Random sampling	2020
Dey *et al.* [6]	Sequence features	Human–SARS-CoV-2	Ensemble learning based on RF, SVM-polynomial, and SVM-radial	Degree-based negative sampling	2020

However, this approach is only a stopgap measure and would be less effective if the target viruses are far from the source viruses. The data quality also cannot be ignored and might have a greater impact on model quality. The limitations of the experimental methods result in the inevitable noise of human–virus PPI data in the databases. Thus, some measures need to be taken to filter the data. For example, PPIs with multiple reports or identified through low-throughput experiments are preferred.

Negative sample selection. In addition to positive samples, negative samples are also required in training a supervised learning-based model. Due to the difficulty of obtaining experimental evidence that two proteins are noninteracting, the so-called "gold standard" of non-PPIs is hard to establish. The common method for negative sample selection is random sampling [86], which randomly selects a certain number of human–virus protein pair combinations without interaction evidence. The underlying assumption is that the total number of negative samples is markedly larger than that of the positive samples, and random sampling can thus identify real noninteraction samples with a high probability. The shortcoming of this simple negative selection method lies in the inclusion of PPIs, which might yield biased prediction performance to some extent. To overcome this shortcoming, other negative sample selection strategies have been proposed. Mei *et al.* [85] eliminated the subcellular colocalized protein pairs based on randomly selected negative samples and found that the performance was better. However, the resulting negative sample data cannot represent those protein pairs that share subcellular colocalization but do not interact. The dissimilarity-based negative sampling approach, which was developed by Eid *et al.* [87], appears to make more biological sense. The core idea is that if two viral proteins share sequence similarity (i.e. sequence identity > 20%), a human protein that interacts with one of the viral proteins will not be able to pair with the other as a negative sample. Very recently, Dey *et al.* proposed a new negative sampling method for predicting

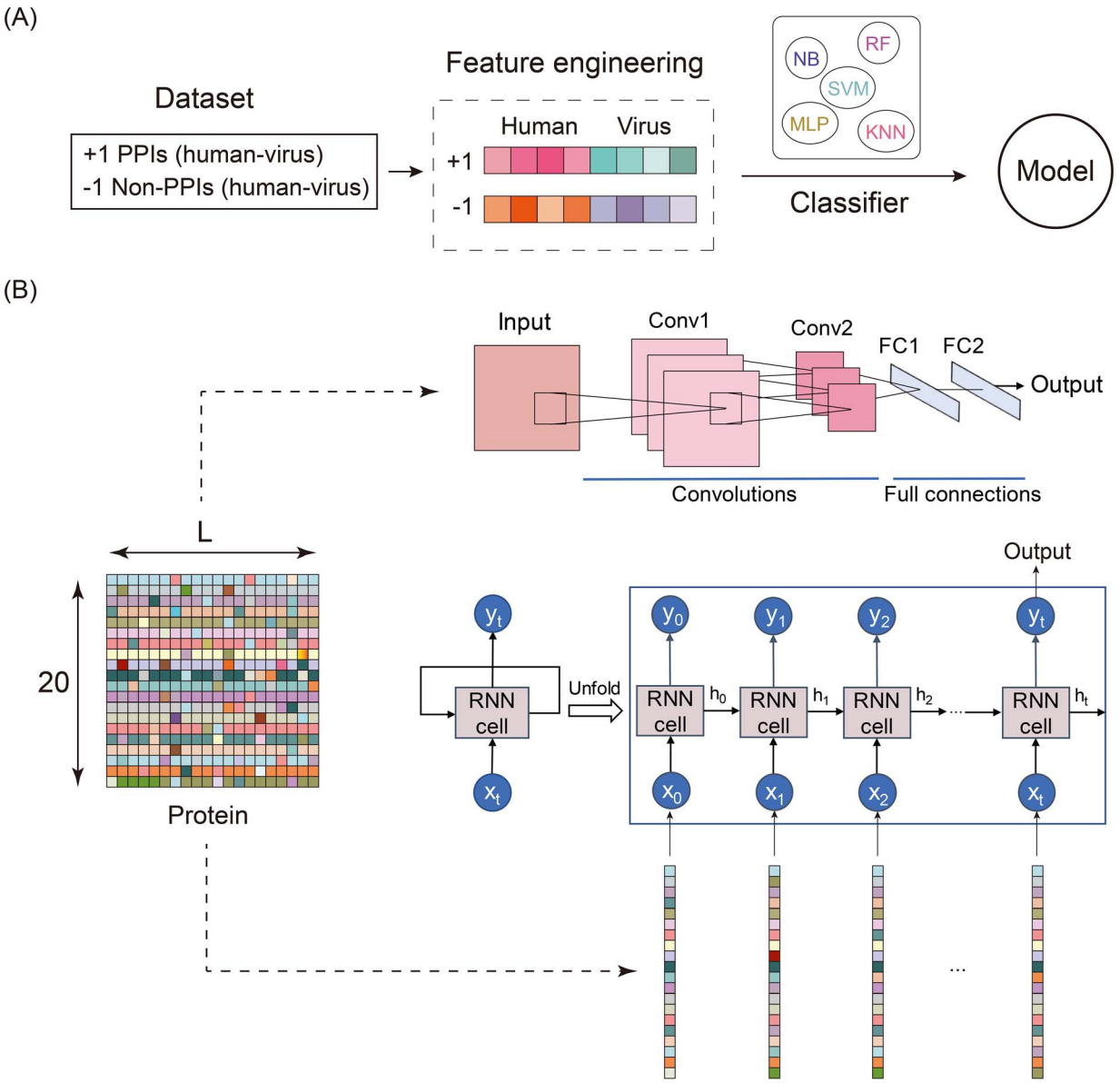

Figure 4. Flowcharts of training human–virus PPI predictive models by traditional ML algorithms and two DL algorithms. (A) Traditional ML methods first perform complicated feature engineering on the dataset and then input the feature vectors into the classifier (e.g. KNN, NB, RF, SVM and MLP) for model training. (B) A simple diagram showing the model training process of CNN or RNN. In general, CNNs consist of one or more convolutional layers and multiple fully connected layers. Here, a CNN containing two convolutional layers ('Conv1' and 'Conv2') and two fully connected layers ('FC1' and 'FC2') is shown. RNN establishes weight connections between neurons in the same layer, in which each input is dependent on the previous input in a time series.

human–SARS-CoV-2 PPIs [6] that considers the degree of human proteins in the human PPI network. More specifically, these researchers randomly selected lower degree human proteins to pair with virus proteins as negative samples. The above strategy was inspired by the observation that virus proteins tend to target higher degree human proteins [63, 88]. To avoid the construction of negative samples, some methods have been developed to learn the interaction patterns only from positive samples [89–91]. However, these methods inevitably have a higher risk of yielding false positives without learning noninteractive patterns.

Another open issue is the ratio of positive-to-negative samples. The simplest way is to use a balanced ratio of positive-to-negative samples (i.e. 1:1). However, training the model with a balanced ratio will overestimate the performance of the model because the actual number of negative samples is markedly larger than that of positive samples. In contrast, selecting an extremely unbalanced positive-to-negative sample ratio will cause the model to be biased toward learning the characteristics of negative samples, which will reduce the generalization ability of the model. To address this issue, we propose a possible strategy to choose an optimal sample ratio in model training. First, the ratio of positives to negatives in the test set should be fixed at the real ratio of positives to negatives. This test set should then be used to assess the performance of models trained on the training set with different ratios of positives to negatives. By doing so, an optimal ratio in the training set can be determined.

Feature engineering

As an important step in ML, feature engineering involves the establishing an encoding scheme that converts training data into machine-recognizable data representations (i.e. feature vectors) [92]. Seeking suitable encoding schemes is an effective approach for developing a powerful ML predictor. Currently, the classical encoding schemes for predicting human–virus PPIs mainly include sequence-based features [59, 76, 87, 93–95], network-based features [59, 76, 93, 96, 97], biological function-based features [59, 85, 93, 96, 97], expression features [59, 97, 98] and evolutionary information [97]. Among these features, sequence-based features and network-based features are frequently used and often result in good performance. Compared with the above-mentioned traditional feature engineering methods, more advanced feature extraction methods have been proposed with the rise of deep learning (DL) [99]. These new encoding schemes are collectively referred to as protein embedding methods [100], which are derived from natural language processing. Some traditional feature extraction methods and novel feature embedding methods are further elaborated as follows.

Sequence-based features. The secret of the interaction between two proteins might lie in their sequences because the sequence of a protein determines its structure and thus its function. The simplest sequence-based encoding scheme for predicting human–virus PPIs involves calculating the amino acid composition [96]. In addition, some methods first classified amino acids into several groups (six or seven groups) according to their physicochemical properties and then calculated the *K* consecutive amino acid group composition [87, 93, 94]. The frequency of triplet amino acid groups (i.e. $K = 3$, also called 'conjoint triad' [101]) is often calculated for each protein. To predict host–virus PPIs, Zhou *et al.* [95] employed another encoding scheme based on dividing amino acids into seven categories that calculates the composition, transition and distribution of amino acid groups.

Network-based features. This type of feature is derived from the topological characteristics of the human PPI network and was inspired by the previous finding that viruses evolve to target human proteins with unique topological properties in the human PPI network [59]. For instance, virally targeted human proteins often appear as hubs (proteins with many interacting partners) and bottlenecks (proteins that are central to many paths in the network) in the human PPI network [18, 63, 88] (Figure 2B). In addition to the degree and betweenness centrality, other network properties, such as the clustering coefficient, have also been used [59, 76, 97]. In short, the centrality of a human protein in the human PPI network is an indicator of whether the human protein is targeted by virus proteins. In general, these network topology parameters can be easily calculated from some well-known network analysis tools or packages, such as Cytoscape [102] and the R package igraph [103]. However, the network properties of human proteins cannot be obtained if they do not appear in the current human interactome. To address this issue, a usual imputation method is to replace the missing parameters with mean values or with the values of homologous/similar proteins.

Biological function-based features. This feature type includes gene ontology (GO) features [85], pathway features [91], domain features [59, 93, 96] and post-translational modification features [59, 97, 98]. Among these, GO features provide the most comprehensive display of protein functional features. GO [104] is a widely used gene/protein functional annotation system that provides a defined vocabulary of gene/protein attributes from three biological aspects (i.e. cell components, molecular functions and biological processes). Each protein can be annotated with one or more descriptive GO terms. The GO features in human–virus PPI prediction mainly consider the following two points [105]: (i) proteins located in the same cell compartment are more likely to interact than proteins residing in spatially separated compartments and (ii) proteins that participate in similar biological processes or perform similar molecular functions are more likely to interact. Tastan *et al.* [59] developed two GO-based features named 'pairwise GO similarity' and 'neighboring GO similarity', which measure the GO similarity between the HIV-1 protein and the human protein in a protein pair and the GO similarity between the HIV-1 protein and the human protein's partners, respectively. Given that some proteins still lack GO information for extracting this type of feature, Mei *et al.* [85] proposed a transfer learning method to transfer the GO information of homologs to enrich or substitute for the GO information of targets. However, the lack of GO information remains a major drawback of the GO-based encoding scheme.

Protein embeddings. As a new type of feature engineering, protein embeddings were derived from the word embedding technique developed in natural language processing [106]. Briefly, word embedding is actually the process of converting a word in a sentence, a paragraph or an article into a distributed representation [106]. The main idea is to map each word in the corpus to a unique, continuous and low-dimensional vector in the vector space, in which the direction and position of this vector can measure the meaning and emotional color of the word to some extent. The neural network can be modeled according to the context of the word and the relationship between the context and the target word, and each word can then finally derive its corresponding word vector from the model [107]. Among the several neural network models developed for word embedding, the two most famous are the continuous bag-of-words (CBOW) model and skip-gram model proposed by Mikolov *et al.* [108]. The word2vec [107] approach was then developed to implement these two models of word embedding. The doc2vec [109] approach, which is an extension of word2vec, learns representations from entire sentences, paragraphs or documents not just surrounding context words. In recent years, these embeddings have also been rapidly applied in the representation of protein sequences [92, 110, 111]; in this method, the protein sequence is regarded as a long sentence, and the k-mers derived from the protein sequence are treated as words (Figure 3B). In our previous work [112], a large number of protein sequences from the Swiss-Prot database [113] were used to train the doc2vec model, and the protein sequences from the human–virus PPIs were then fed into the model to obtain their vector representations. The embedding inferred from doc2vec was found to be superior to that obtained with the traditional sequence-based encoding schemes. Similarly, the node2vec [114] approach is an upgraded version of the traditional network parameter encoding, which can learn the feature representations for nodes in the graph/network (Figure 3B). In fact, node2vec first uses random walks to generate many node wandering path sequences and thus to cleverly convert the node embedding into word embedding. Expanding on this idea, GO2vec[115] utilizes the GO hierarchy graph. The node2vec and GO2vec approaches have been tentatively applied for intraspecies PPI prediction [115, 116] and have achieved promising results. Therefore, these methods deserve more attention in the future prediction of human–virus PPIs.

ML algorithms and predictive frameworks

Traditional ML algorithms, such as k-nearest neighbor (KNN), naïve Bayes (NB), random forest (RF), support vector machine (SVM) and multilayer perceptron (MLP), have been widely employed for the development of bioinformatics prediction methods. The choice of ML algorithms depends on the classification tasks or training data. The optimal combination among different ML algorithms and encoding schemes also needs to be selected according to the actual performance. To the best of our knowledge, SVM and RF are two frequently used ML algorithms for predicting human–virus PPI in recent decades and are generally superior to other popular ML algorithms. Figure 4A shows the process of training the model with traditional ML algorithms.

In addition to traditional ML methods, DL algorithms [99] have also been successfully applied to solve various biological prediction issues, including PPI prediction. As a branch of ML, DL has blossomed and grown in popularity over the last decade. Due to the limited amount of data and computing resources, the earliest neural network architecture was too simple to demonstrate its full advantages. Due to the improvements in computing power and the emergence of massive data, particularly after 2010, various deep neural network frameworks have emerged and have exhibited powerful performance in a series of biological prediction and classification tasks [117, 118]. Regarding the issue of PPI prediction, several DL frameworks [119–123] have also been developed for predicting intraspecies PPIs. These studies commonly use convolutional neural networks (CNNs) [124, 125], which are widely used for modeling images to automatically extract local features, or recurrent neural networks (RNNs) [126], which are designed to preserve context and long-term memory information for sequential data. The input of CNN or RNN does not require the complicated encoding schemes described above because the deep neural network itself is equivalent to a process of deep extraction of features. CNN mainly consists of one or more convolutional layers and pooling layers (Figure 4B), whereas the core idea of RNN is that a sequential relationship exists between inputs at different times in the network (Figure 4B). The endpoint of CNN or RNN is usually one or more fully connected layers. As far as future computational prediction of human–virus PPIs is concerned, these frameworks can also be applied to build more powerful prediction models.

It is generally acknowledged that integrating various classifiers or combining multiple features to conduct model training will yield better results than a single classifier or a single feature [127–129]. For instance, Emamjomeh *et al.* [97] employed stacking to combine four component learners (i.e. SVM, RF, NB and MLP) to predict human–HCV PPIs, and in this approach, the output from each classifier was considered as the input of a meta-learner (i.e. MLP) to produce the final prediction results. In terms of feature integration, the simplest way to combine multiple features is to concatenate them into a long feature vector. However, the resulting high-dimensional vectors might cause dimensional explosion and the contributions of some small feature types to be ignored. In other words, the optimal combination requires intensive computational experiments.

Model evaluation

When building the model, the collected dataset should be partitioned into a training set and an independent test set, and the ratio of partitions is related to the abundance of the available data. Note that the independent test set should not be used for training at all but rather only used to assess the final performance of the trained model. To avoid overfitting in the training process, k-fold cross-validation ($k = 5$ or 10 are often adopted) is commonly conducted with the training set. The parameters of the model can be optimized according to the performance of the k-fold cross-validation. Eventually, the performance of the k-fold cross-validation and the independent test will be jointly used to evaluate the overall performance of the developed prediction method.

Several performance evaluation indicators commonly used in binary classification models include Accuracy, Precision, Sensitivity (i.e. Recall) and Specificity, which are derived from the confusion matrix [130]. These indicators are calculated for a given classification threshold, whereas two types of curves called the receiver operating characteristic curve (ROC curve) and the precision-recall curve (PR curve) are plotted by gradually changing the thresholds. The area under the ROC or PR curve is usually used as a more accurate indicator to further measure the model performance. To properly evaluate the performance, it is worth emphasizing the following issues. The performance of a predictive model is highly relevant to the collection of the dataset (e.g. the ratio of positive to negative samples, the filtering of positive samples and the construction of negative samples). To avoid the generation of biased benchmarking results when comparing different human–virus PPI prediction methods, the model must be trained and assessed with the same datasets. Moreover, the PR curve is more suitable for evaluating the model performance when the ratio of positive and negative samples is unbalanced. Due to the lack of 'gold standard' human–virus PPI datasets, fair performance comparison among different prediction methods remains a challenging task. Due to the increasing availability of data regarding human–virus PPIs, we hope that some standard training and testing datasets can be constructed in the future, and these benchmark datasets will definitely facilitate the reliable comparison of different prediction methods.

Biological applications of human–virus PPI networks

Based on experimentally verified or predicted human–virus PPIs, the corresponding human–virus PPI networks can be constructed. Further exploration of the networks can capture more biologically meaningful knowledge. On the one hand, a deeper understanding of the human–virus interaction mechanism could be gained through network analysis, structure analysis and integration analysis with other multifaceted data. On the other hand, human–virus PPIs could be applied to develop new antiviral therapeutic strategies, such as drug development.

Applications in mechanistic analysis of the human–virus relationship

By analyzing the topological properties of each protein node in the human PPI network, the key patterns of virally targeted human proteins in the PPI network can be captured. As mentioned earlier, Cytoscape [102] and igraph [103] can analyze multiple topology parameter properties for a network. It is a widely proven consensus that viruses tend to attack hub proteins or bottleneck proteins in the human PPI network [63, 88] (Figure 2B).

The analysis of the human–virus interactome from different levels of protein structure enables an in-depth understanding of the interaction mechanism. At the protein domain level, Itzhaki [131] and Zheng *et al.* [132] analyzed the PPIs between viruses and human, and both of these research groups found that viral domains preferentially interact with human hub domains. From the perspective of shorter motif structure, motif

mimicry (Figure 2C) was found to be a more commonly used strategy for viruses to more efficiently interact with human proteins [64, 133]. Due to the limitation of genome size, viruses evolve multiple short linear motifs to effectively mimic, hijack and manipulate complex host processes for survival [64]. Furthermore, from the perspective of the 3D structure of the protein, a systems understanding of interface mimicry can be achieved. Franzosa and Xia [67] depicted the structural principles within the human–virus PPI network and reconstructed the human–virus structural interaction network by mapping curated and predicted 3D structural models of human–virus and human–human protein complexes. Because human–virus PPIs are transient and regulatory in nature, researchers have found that viral proteins are more inclined to use interface mimicry to achieve efficient interactions without any sequence or structural similarity to virally binding human proteins' partners.

Human–virus interactome analyses in the context of human protein complexes have also been performed. It has been well established that human protein complexes are heavily involved in viral infection [31, 134]. Jäger *et al.* unveiled a number of host complexes targeted by viral proteins [31] and these included eIF3d (a subunit of eukaryotic translation initiation factor 3) cleaved by HIV protease. Our previous study revealed that viral targets are enriched within human protein complexes and tend to have a high within-complex degree [134]. Moreover, we found that complexes necessary for viral replication are simultaneously targeted by multiple viruses [134].

Furthermore, by combining other data information, such as the pathway annotation [30, 63], GO [31, 33], gene expression [63, 134], protein abundance [33, 63] and evolutionary rate [63, 134] of proteins, the patterns mediating human–virus PPIs can be better deciphered. Through integrative analysis, virally targeted human proteins were found to be highly conserved [31, 63], expressed abundantly across multiple tissues [63], or play specific roles in multiple pathways or processes, such as cell cycle regulation, nuclear transport and immune response [63, 88].

Taken together, the results from large-scale human–virus interactome analyses involving the integration of multifaceted data have provided a comprehensive and increasingly clear landscape of human–virus relationships. The above observations have provided some fundamental hints for understanding how viruses evolve to interact with human proteins that might control critical human cellular processes. It should be emphasized that the computational prediction and mechanistic analysis of human–virus PPIs are mutually reinforcing. On the one hand, the reliable prediction results facilitate biological analysis from a more complete human–virus interactome. By employing structural information to infer the human–virus PPI network, Lasso *et al.* [82] not only rediscovered the existing biological knowledge but also obtained a series of new biological findings. For instance, these researchers observed the shared and unique infection mechanisms employed by different viruses [82]. On the other hand, the patterns inferred from biological analysis can be rapidly used for designing new strategies to predict human–virus PPIs. For example, the network-based encoding scheme is often derived from the topological analysis of the experimentally identified human–virus PPI network [63,88].

Applications in the development of new therapeutic strategies

The available human–virus interactome can play a crucial role in antiviral drug discovery. The conventional treatments for viral diseases mainly attack the components of the virus, for example, viral enzymes, to break the virus structure and functionality [135, 136]. However, the rapid mutation of the viral genome makes the drug quickly ineffective. To develop more efficient antiviral drugs, host-oriented drug target discovery [135, 136] is an important direction in which human–virus PPIs have been a crucial resource. Human proteins that are indispensable for viruses during infections but are not essential for human cells could serve as potential antiviral targets. In addition, drug development is a time-consuming and costly process. Even if the drug has been developed for a suitable target, it remains far from being officially approved for marketing. Therefore, finding new uses for traditional medicines (i.e. drug repositioning or drug repurposing) is a very promising approach for saving time and cost in drug development. Some known drug–target interactions can be obtained from databases, such as DrugBank [137] (http://www.drugbank.ca) and Therapeutic Target Database [138] (http://db.idrblab.net/ttd/). By integrating drug–target interactions, human–virus PPIs, human PPI networks and other information, drug repurposing can be performed. It is worth mentioning that in response to the sudden outbreak of COVID-19 caused by SARS-CoV-2, several research teams have rapidly found druggable human targets or drug repurposing candidates through human–virus PPI analysis [33, 139–141].

The established human–virus PPI networks can also provide new evidence on diseases associated with the viruses under investigation. By combining disease-gene associations and human–virus PPIs, a link can be constructed to find the potential relationship between viruses and disease occurrence or development. As a public database to store associations between diseases and genes, DisGeNET [142] (http://www.disgenet.org/) can be employed for such tasks. Zheng *et al.* [132] performed an analysis that links human–virus PPIs to diseases by integrating human–virus PPI, DDI and disease-related gene information, and their results uncovered several unknown virus–disease relationships. More recently, we also reconstructed the PPI network between human and HSV-1 and obtained potential new evidence of the association between HSV-1 infection and Alzheimer's disease [76].

It should also be mentioned that a causal relationship has been demonstrated between 12% cancers and seven different viruses (e.g. cervical cancer caused by HPV) [143, 144]. Therefore, the analysis of human–virus PPI networks might shed new light on the therapeutic strategies for cancers [30, 145]. Eckhardt *et al.* [30] developed an integrated strategy based on the human–HPV PPI network and tumor genome analysis for in-depth exploration, and their approach led to the identification of multiple carcinogenic pathways promoted by human–HPV PPIs. Methodologically, the integration of human–virus PPI mapping and tumor genome analysis defines a pipeline for other virally induced cancers to further analyze the relationship between human–virus PPIs and oncogenesis.

Summary and future perspectives

To explore the interplay between human and viruses, establishing a global view of the human–virus interactome is critical. Recent years have witnessed the rapid accumulation of human–virus PPI data, which have been summarized in many databases for the convenience of the research community. A large number of human–virus PPI prediction methods, particularly ML-based approaches, have been elegantly developed. Moreover, many large-scale analyses based on experimental or predicted human–virus PPI networks have been performed. Due to the joint efforts of experimental and computational biologists, we

have obtained an increasingly complete human–virus interactome and have deciphered fundamental mechanisms governing human–virus relationships.

Despite two decades of progress in human–virus PPI research, the current computational studies of the human–virus interactome are still subjected to the following three major limitations or challenges. First, the performance of existing human–virus PPI prediction methods in real applications remains unsatisfactory. Second, the predicted human–virus PPI data are not easily accessible to the community. Third, the speed of converting experimental or predicted human–virus PPIs into new scientific discoveries or practical applications should be accelerated.

Several studies have attempted to address the above limitations or challenges. First, we still need sufficient high-quality human–virus PPI data for conducting large-scale interactome analysis and constructing reliable predictive models. Second, in terms of prediction algorithms, DL algorithms should be rapidly applied to the prediction of human–virus PPIs because DL has recently exhibited powerful performance in solving a series of protein bioinformatics prediction tasks, including the prediction of intraspecies PPIs [121, 123]. Moreover, only a few existing methods provide online predictors, and thus, more useful prediction software programs or web servers are needed to allow experimental scientists to take full advantage of the progress in prediction methods. Third, the establishment of tissue- or spatiotemporal-specific human–virus PPI networks should be considered. Given the potential tissue preference of viruses, human–virus PPIs might differ among different tissues of the human host. In addition, the establishment of PPI networks with spatial and temporal characteristics can better demonstrate the dynamics of the infection process. Construction of the structural human–virus interactome and integration of multiomics data associated with the human–virus interactome are also two effective approaches for better elucidating a mechanistic understanding of human–virus relationships. Last but certainly not least, more attention should always be paid to converting the available human–virus interactome data into new therapeutic strategies for human diseases associated with viruses in the future.

Key Points

- In past decades, advances in experimental technology have led to the identification of an unprecedented number of human–virus PPIs and the establishment of a series of human–virus PPI databases.
- Cost-effective methods for predicting human–virus PPIs have been intensively developed to complement experimental efforts, in which machine learning-based approaches are playing an increasingly important role.
- The booming deep learning methods and new feature engineering approaches (e.g. the protein embedding techniques) are propelling the performance of machine learning-based methods to a new level.
- Computational prediction methods and large-scale analysis for human–virus PPIs are mutually reinforcing. On the one hand, the prediction results allow us to capture the global landscape of the human–virus interactome more rapidly, and on the other hand, the patterns/rules inferred from human–virus PPI analysis can be used to develop new prediction methods.
- The available human–virus PPI networks have been applied to understand human–virus relationships and to develop antiviral therapies.

Acknowledgments

We apologize that due to space limitations, we cannot cite all the related literature.

Funding

National Key Research and Development Program of China [2017YFC1200205].

References

1. Louten J. Virus structure and classification. *Essent Hum Virol* 2016;**21**:19–29.
2. Mayer KA, Stöckl J, Zlabinger GJ, *et al*. Hijacking the supplies: metabolism as a novel facet of virus–host interaction. *Front Immunol* 2019;**10**:1533.
3. Méthot P-O, Alizon S. Emerging disease and the evolution of virulence: the case of the 1918–1919 influenza pandemic. *Classif Dis Evid* 2015;**7**:93–130.
4. Kharsany ABM, Karim QA. HIV infection and AIDS in sub-Saharan Africa: current status, challenges and opportunities. *Open AIDS J* 2016;**10**:34–48.
5. Pfefferbaum B, North CS. Mental health and the Covid-19 pandemic. *N Engl J Med* 2020;**383**:510–2.
6. Dey L, Chakraborty S, Mukhopadhyay A. Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins. *Biom J* 2020;**43**:438–50.
7. Khorsand B, Savadi A, Naghibzadeh M. SARS-CoV-2–human protein–protein interaction network. *Inform Med Unlocked* 2020;**20**:100413.
8. Gul S, Hadian K. Protein–protein interaction modulator drug discovery: past efforts and future opportunities using a rich source of low- and high-throughput screening assays. *Expert Opin Drug Discov* 2014;**9**:1393–404.
9. Peng X, Wang J, Peng W, *et al*. Protein–protein interactions: detection, reliability assessment and applications. *Brief Bioinform* 2017;**18**:798–819.
10. Petschnigg J, Snider J, Stagljar I. Interactive proteomics research technologies: recent applications and advances. *Curr Opin Biotechnol* 2011;**22**:50–8.
11. Velazquez-Campoy A, Freire E. ITC in the post-genomic era…? Priceless. *Biophys Chem* 2005;**115**:115–24.
12. Brymora A, Valova VA, Robinson PJ. Protein–protein interactions identified by pull-down experiments and mass spectrometry. *Curr Protoc Cell Biol* 2004;**22**:17.5.1–17.5.51.
13. Jung SO, Ro HS, Kho BH, *et al*. Surface plasmon resonance imaging-based protein arrays for high-throughput screening of protein–protein interaction inhibitors. *Proteomics* 2005;**5**:4427–31.
14. Ito T, Chiba T, Ozawa R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci* 2001;**98**:4569–74.
15. Stynen B, Tournu H, Tavernier J, *et al*. Diversity in genetic in vivo methods for protein–protein interaction studies:

from the yeast two-hybrid system to the mammalian split-luciferase system. *Microbiol Mol Biol Rev* 2012;**76**:331–82.
16. Mak AB, Ni Z, Hewel JA, *et al*. A lentiviral functional proteomics approach identifies chromatin remodeling complexes important for the induction of pluripotency. *Mol Cell Proteomics* 2010;**9**:811–23.
17. Köcher T, Superti-Furga G. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat Methods* 2007;**4**:807–15.
18. Calderwood MA, Venkatesan K, Xing L, *et al*. Epstein–Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci* 2007;**104**:7606–11.
19. de Chassey B, Navratil V, Tafforeau L, *et al*. Hepatitis C virus infection protein network. *Mol Syst Biol* 2008;**4**:230.
20. Dolan PT, Zhang C, Khadka S, *et al*. Identification and comparative analysis of hepatitis C virus-host cell protein interactions. *Mol Biosyst* 2013;**9**:3199–209.
21. Germain MA, Chatel-Chaix L, Gagné B, *et al*. Elucidating novel hepatitis C virus–host interactions using combined mass spectrometry and functional genomics approaches. *Mol Cell Proteomics* 2014;**13**:184–203.
22. Ramage HR, Kumar GR, Verschueren E, *et al*. A combined proteomics/genomics approach links hepatitis C virus infection with nonsense-mediated mRNA decay. *Mol Cell* 2015;**57**:329–40.
23. Shapira SD, Gat-Viks I, Shum BOV, *et al*. A physical and regulatory map of host–influenza interactions reveals pathways in H1N1 infection. *Cell* 2009;**139**:1255–67.
24. Khadka S, Vangeloff AD, Zhang C, *et al*. A physical interaction network of dengue virus and human proteins. *Mol Cell Proteomics* 2011;**10**:M111.012187.
25. Shah PS, Link N, Jang GM, *et al*. Comparative flavivirus–host protein interaction mapping reveals mechanisms of dengue and Zika virus pathogenesis. *Cell* 2018;**175**:1931–45.
26. Pichlmair A, Kandasamy K, Alvisi G, *et al*. Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature* 2012;**487**:486–90.
27. Griffiths SJ, Koegl M, Boutell C, *et al*. A systematic analysis of host factors reveals a Med23-interferon-λ regulatory axis against herpes simplex virus type 1 replication. *PLoS Pathog* 2013;**9**:e1003514.
28. Rozenblatt-Rosen O, Deo RC, Padi M, *et al*. Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* 2012;**487**:491–5.
29. White EA, Sowa ME, Tan MJA, *et al*. Systematic identification of interactions between host cell proteins and E7 oncoproteins from diverse human papillomaviruses. *Proc Natl Acad Sci* 2012;**109**:E260–7.
30. Eckhardt M, Zhang W, Gross AM, *et al*. Multiple routes to oncogenesis are promoted by the human papillomavirus–host protein network. *Cancer Discov* 2018;**8**:1474–89.
31. Jäger S, Cimermancic P, Gulbahce N, *et al*. Global landscape of HIV–human protein complexes. *Nature* 2012;**481**:365–70.
32. Batra J, Hultquist JF, Liu D, *et al*. Protein interaction mapping identifies RBBP6 as a negative regulator of Ebola virus replication. *Cell* 2018;**175**:1917–30.
33. Gordon DE, Jang GM, Bouhaddou M, *et al*. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020;**583**:459–68.
34. Li J, Guo M, Tian X, *et al*. Virus–host interactome and proteomic survey reveal potential virulence factors influencing SARS-CoV-2 pathogenesis. *Med* 2020 in press. doi: 10.1016/j.medj.2020.07.002.
35. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res* 2017;**45**:D408–14.
36. Li T, Wernersson R, Hansen RB, *et al*. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat Methods* 2017;**14**:61–4.
37. Braun P, Tasan M, Dreze M, *et al*. An experimentally derived confidence score for binary protein–protein interactions. *Nat Methods* 2009;**6**:91–7.
38. Villaveces JM, Jimenez RC, Porras P, *et al*. Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database* 2015;**2015**:bau131.
39. Kwofie SK, Schaefer U, Sundararajan VS, *et al*. HCVpro: hepatitis C virus protein interaction database. *Infect Genet Evol* 2011;**11**:1971–7.
40. Ako-Adjei D, Fu W, Wallin C, *et al*. HIV-1, human interaction database: current status and new features. *Nucleic Acids Res* 2015;**43**:D566–70.
41. Karyala P, Metri R, Bathula C, *et al*. DenHunt - a comprehensive database of the intricate network of dengue–human interactions. *PLoS Negl Trop Dis* 2016;**10**:e0004965.
42. Dey L, Mukhopadhyay A. DenvInt: a database of protein–protein interactions between dengue virus and its hosts. *PLoS Negl Trop Dis* 2017;**11**:e0005879.
43. Gurumayum S, Brahma R, Naorem LD, *et al*. ZikaBase: an integrated ZIKV–human interactome map database. *Virology* 2018;**514**:203–10.
44. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res* 2015;**43**:D583–7.
45. Calderone A, Licata L, Cesareni G. VirusMentha: a new resource for virus–host protein interactions. *Nucleic Acids Res* 2015;**43**:D588–92.
46. Ammari MG, Gresham CR, McCarthy FM, *et al*. HPIDB 2.0: a curated database for host–pathogen interactions. *Database* 2016;**2016**:baw103.
47. Durmuş Tekir S, Çakir T, Ardiç E, *et al*. PHISTO: pathogen–host interaction search tool. *Bioinformatics* 2013;**29**:1357–8.
48. Cook H, Doncheva N, Szklarczyk D, *et al*. Viruses.STRING: a virus–host protein–protein interaction database. *Viruses* 2018;**10**:519.
49. Aranda B, Blankenburg H, Kerrien S, *et al*. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods* 2011;**8**:528–9.
50. Chatr-aryamontri A, Ceol A, Peluso D, *et al*. VirusMINT: a viral protein interaction database. *Nucleic Acids Res* 2009;**37**:D669–73.
51. Orchard S, Kerrien S, Abbani S, *et al*. Protein interaction data curation: the international molecular exchange (IMEx) consortium. *Nat Methods* 2012;**9**:345–50.
52. Szklarczyk D, Gable AL, Lyon D, *et al*. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–13.
53. Yang X, Lian X, Fu C, *et al*. HVIDB: a comprehensive database for human–virus protein–protein interactions. *Brief Bioinform* 2021 in press. doi: 10.1093/bib/bbaa425.
54. Matthews LR, Vaglio P, Reboul J, *et al*. Identificaton of potential interaction networks using sequence-based searches for conserved protein–protein interactions or interologs. *Genome Res* 2001;**11**:2120–6.

55. Yu H, Luscombe NM, Lu HX, *et al*. Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res* 2004;**14**:1107–18.
56. Dyer MD, Murali TM, Sobral BW. Computational prediction of host–pathogen protein–protein interactions. *Bioinformatics* 2007;**23**:i159–66.
57. Evans P, Dampier W, Ungar L, *et al*. Prediction of HIV-1 virus–host protein interactions using virus and host sequence motifs. *BMC Med Genomics* 2009;**2**:27.
58. Zhang QC, Petrey D, Deng L, *et al*. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 2012;**490**:556–60.
59. Tastan O, Qi Y, Carbonell JG, *et al*. Prediction of interactions between HIV-1 and human proteins by information integration. *Pacific Symp Biocomput* 2009;**527**:516–27.
60. Sen R, Nayak L, De RK. A review on host–pathogen interactions: classification and prediction. *Eur J Clin Microbiol Infect Dis* 2016;**35**:1581–99.
61. Nourani E, Khunjush F, DurmuÅŸ S. Computational approaches for prediction of pathogen–host protein–protein interactions. *Front Microbiol* 2015;**6**:94.
62. Akiva E, Friedlander G, Itzhaki Z, *et al*. A dynamic view of domain–motif interactions. *PLoS Comput Biol* 2012;**8**:e1002341.
63. Halehalli RR, Nagarajaram HA. Molecular principles of human virus protein–protein interactions. *Bioinformatics* 2015;**31**:1025–33.
64. Davey NE, Travé G, Gibson TJ. How viruses hijack cell regulation. *Trends Biochem Sci* 2011;**36**:159–69.
65. Elde NC, Malik HS. The evolutionary conundrum of pathogen mimicry. *Nat Rev Microbiol* 2009;**7**:787–97.
66. Chemes LB, de Prat-Gay G, Sánchez IE. Convergent evolution and mimicry of protein linear motifs in host–pathogen interactions. *Curr Opin Struct Biol* 2015;**32**:91–101.
67. Franzosa EA, Xia Y. Structural principles within the human–virus protein–protein interaction network. *Proc Natl Acad Sci* 2011;**108**:10538–43.
68. Hagai T, Azia A, Babu MM, *et al*. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host–virus interactions. *Cell Rep* 2014;**7**:1729–39.
69. Dinkel H, Michael S, Weatheritt RJ, *et al*. ELM - the database of eukaryotic linear motifs. *Nucleic Acids Res* 2012;**40**:242–51.
70. Stein A, Céol A, Aloy P. 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* 2011;**39**:D718–23.
71. Hulo N. The PROSITE database. *Nucleic Acids Res* 2006;**34**:D227–30.
72. Zhang Z, Kochhar S, Grigorov MG. Descriptor-based protein remote homology identification. *Protein Sci* 2005;**14**:431–44.
73. Chiang AWT, Wu WYL, Wang T, *et al*. Identification of entry factors involved in hepatitis C virus infection based on host-mimicking short linear motifs. *PLoS Comput Biol* 2017;**13**:e1005368.
74. García-Pérez CA, Guo X, Navarro JG, *et al*. Proteome-wide analysis of human motif–domain interactions mapped on influenza a virus. *BMC Bioinform* 2018;**19**:238.
75. Becerra A, Bucheli VA, Moreno PA. Prediction of virus–host protein–protein interactions mediated by short linear motifs. *BMC Bioinform* 2017;**18**:163.
76. Lian X, Yang X, Shao J, *et al*. Prediction and analysis of human–herpes simplex virus type 1 protein–protein interactions by integrating multiple methods. *Quant Biol* 2020;**8**:312–24.
77. Hashim FA, Mabrouk MS, Al-Atabany W. Review of different sequence motif finding algorithms. *Avicenna J Med Biotechnol* 2019;**11**:130–48.
78. Yan Y, Tao H, He J, *et al*. The HDOCK server for integrated protein–protein docking. *Nat Protoc* 2020;**15**:1829–52.
79. Doolittle JM, Gomez SM. Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. *Virol J* 2010;**7**:82.
80. Doolittle JM, Gomez SM. Mapping protein interactions between dengue virus and its human and insect hosts. *PLoS Negl Trop Dis* 2011;**5**:e954.
81. de Chassey B, Meyniel-Schicklin L, Aublin-Gex A, *et al*. Structure homology and interaction redundancy for discovering virus–host protein interactions. *EMBO Rep* 2013;**14**:938–44.
82. Lasso G, Mayer SV, Winkelmann ER, *et al*. A structure-informed atlas of human–virus interactions. *Cell* 2019;**178**:1526–41.
83. Kshirsagar M, Schleker S, Carbonell J, *et al*. Techniques for transferring host–pathogen protein interactions knowledge to new tasks. *Front Microbiol* 2015;**6**:36.
84. Kshirsagar M, Carbonell J, Klein-Seetharaman J. Multisource transfer learning for host–pathogen protein interaction prediction in unlabeled tasks. *NIPS Work Mach Learn Comput Biol* 2013;3–6.
85. Mei S. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PLoS One* 2013;**8**:e79606.
86. Gomez SM, Noble WS, Rzhetsky A. Learning to predict protein–protein interactions from protein sequences. *Bioinformatics* 2003;**19**:1875–81.
87. Eid F-E, ElHefnawi M, Heath LS. DeNovo: virus–host sequence-based protein–protein interaction prediction. *Bioinformatics* 2016;**32**:1144–50.
88. Dyer MD, Murali TM, Sobral BW. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog* 2008;**4**:e32.
89. Nouretdinov I, Gammerman A, Qi Y, *et al*. Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method. *Pac Symp Biocomput* 2012;311–22.
90. Mukhopadhyay A, Ray S, Maulik U. Incorporating the type and direction information in predicting novel regulatory interactions between HIV-1 and human proteins using a biclustering approach. *BMC Bioinform* 2014;**15**:26.
91. Nourani E, Khunjush F, Durmuş S. Computational prediction of virus–human protein–protein interactions using embedding kernelized heterogeneous data. *Mol Biosyst* 2016;**12**:1976–86.
92. Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* 2015;**10**:e0141287.
93. Dyer MD, Murali TM, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect Genet Evol* 2011;**11**:917–23.
94. Cui G, Fang C, Han K. Prediction of protein–protein interactions between viruses and human by an SVM model. *BMC Bioinform* 2012;**13**:S5.
95. Zhou X, Park B, Choi D, *et al*. A generalized approach to predicting protein–protein interactions between virus and host. *BMC Genomics* 2018;**19**:568.
96. Barman RK, Saha S, Das S. Prediction of interactions between viral and host proteins using supervised machine learning methods. *PLoS One* 2014;**9**:e112034.

97. Emamjomeh A, Goliaei B, Zahiri J, *et al*. Predicting protein–protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol Biosyst* 2014;**10**:3147–54.
98. Qi Y, Tastan O, Carbonell JG, *et al*. Semi-supervised multitask learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics* 2010;**26**:i645–52.
99. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**:436–44.
100. Yang KK, Wu Z, Bedbrook CN, *et al*. Learned protein embeddings for machine learning. *Bioinformatics* 2018;**34**:2642–8.
101. Shen J, Zhang J, Luo X, *et al*. Predicting protein–protein interactions based only on sequences information. *Proc Natl Acad Sci* 2007;**104**:4337–41.
102. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
103. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Syst* 2006; **1695**:1–9.
104. Schleker S, Garcia-Garcia J, Klein-Seetharaman J, *et al*. Prediction and comparison of *Salmonella*–human and *Salmonella–Arabidopsis* interactomes. *Chem Biodivers* 2012;**9**: 991–1018.
105. Maetschke SR, Simonsen M, Davis MJ, *et al*. Gene ontology-driven inference of protein–protein interactions using inducers. *Bioinformatics* 2012;**28**:69–75.
106. Bengio Y, Ducharme R, Vincent P, *et al*. A neural probabilistic language model. *J Mach Learn Res* 2003;**3**:1137–55.
107. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 2013;3111–9.
108. Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. In: *arXiv Prepr arXiv 1301.3781*, 2013.
109. Le QV, Mikolov T. Distributed representations of sentences and documents. *Int Conf Mach Learn ICML 2014* 2014;**32**:1188–96.
110. Kimothi D, Soni A, Biyani P, *et al*. Distributed representations for biological sequence analysis. In: *arXiv Prepr arXiv 1608.05949*, 2016.
111. Yang KK, Wu Z, Bedbrook CN, *et al*. Learned protein embeddings for machine learning. *Bioinformatics* 2018;**34**:2642–8.
112. Yang X, Yang S, Li Q, *et al*. Prediction of human–virus protein–protein interactions through a sequence embedding-based machine learning method. *Comput Struct Biotechnol J* 2020;**18**:153–61.
113. Boutet E, Lieberherr D, Tognolli M, *et al*. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Plant Bioinforma* 2016;**1374**:23–54.
114. Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp.855–864.
115. Zhong X, Kaalia R, Rajapakse JC. GO2Vec: transforming GO terms and proteins to vector representations via graph embeddings. *BMC Genomics* 2019;**20**:918.
116. Yue X, Wang Z, Huang J, *et al*. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 2019;**36**:1241–51.
117. Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.
118. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 2016;**44**:e107.
119. Sun T, Zhou B, Lai L, *et al*. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinform* 2017;**18**:277.
120. Zhang L, Yu G, Xia D, *et al*. Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* 2019;**324**:10–9.
121. Hashemifar S, Neyshabur B, Khan AA, *et al*. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* 2018;**34**:i802–10.
122. Du X, Sun S, Hu C, *et al*. DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. *J Chem Inf Model* 2017;**57**:1499–510.
123. Chen M, CJ-T J, Zhou G, *et al*. Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* 2019;**35**:i305–14.
124. Lawrence S, Giles CL, Tsoi AC, *et al*. Face recognition: a convolutional neural-network approach. *IEEE Trans Neural Netw* 1997;**8**:98–113.
125. Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput* 2017;**29**:2352–449.
126. Yu Y, Si X, Hu C, *et al*. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput* 2019;**31**:1235–70.
127. Kotlyar M, Pastrello C, Pivetta F, *et al*. In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat Methods* 2014;**12**:79–84.
128. Lian X, Yang S, Li H, *et al*. Machine-learning-based predictor of human–bacteria protein–protein interactions by incorporating comprehensive host-network properties. *J Proteome Res* 2019;**18**:2195–205.
129. Yang S, Li H, He H, *et al*. Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods. *Brief Bioinform* 2019;**20**:274–87.
130. Polat K, Güneş S, Arslan A. A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. *Expert Syst Appl* 2008;**34**:482–7.
131. Itzhaki Z. Domain–domain interactions underlying herpesvirus–human protein–protein interaction networks. *PLoS One* 2011;**6**:e21724.
132. Zheng L-L, Li C, Ping J, *et al*. The domain landscape of virus–host interactomes. *Biomed Res Int* 2014;**2014**:867235.
133. Garamszegi S, Franzosa EA, Xia Y. Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human–virus protein–protein interaction networks. *PLoS Pathog* 2013;**9**:e1003778.
134. Yang S, Fu C, Lian X, *et al*. Understanding human–virus protein–protein interactions using a human protein complex-based analysis framework. *mSystems* 2019;**4**: e00303-18.
135. de Chassey B, Meyniel-Schicklin L, Aublin-Gex A, *et al*. New horizons for antiviral drug discovery from virus–host protein interaction networks. *Curr Opin Virol* 2012;**2**:606–13.
136. de Chassey B, Meyniel-Schicklin L, Vonderscher J, *et al*. Virus–host interactomics: new insights and opportunities for antiviral drug discovery. *Genome Med* 2014;**6**:115.
137. Wishart DS, Feunang YD, Guo AC, *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.

138. Wang Y, Zhang S, Li F, *et al*. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res* 2020;**48**:D1031–41.
139. Khan AA, Khan Z. Comparative host–pathogen protein–protein interaction analysis of recent coronavirus outbreaks and important host targets identification. *Brief Bioinform* 2021. in press. doi: 10.1093/bib/bbaa207.
140. Sadegh S, Matschinske J, Blumenthal DB, *et al*. Exploring the SARS-CoV-2 virus–host–drug interactome for drug repurposing. *Nat Commun* 2020;**11**:3518.
141. Zhou Y, Hou Y, Shen J, *et al*. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* 2020;**6**:14.
142. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, *et al*. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020;**48**:D845–55.
143. White MK, Pagano JS, Khalili K. Viruses and human cancers: a long road of discovery of molecular paradigms. *Clin Microbiol Rev* 2014;**27**:463–81.
144. Morris JDH, Eddleston ALWF, Crook T. Viral infection and cancer. *Lancet* 1995;**346**:754–8.
145. Wu Z-J, Zhu Y, Huang D-R, *et al*. Constructing the HBV–human protein interaction network to understand the relationship between HBV and hepatocellular carcinoma. *J Exp Clin Cancer Res* 2010;**29**:146.