METHODOLOGY ARTICLE

Open Access

A multitask transfer learning framework for the prediction of virus-human protein–protein interactions

Thi Ngan Dong[1]* , Graham Brogden[2,3], Gisa Gerold[2,3,4,5] and Megha Khosla[1]

*Correspondence:
dong@l3s.de
[1] L3S Research Center,
Leibniz University Hannover,
Hannover, Germany
Full list of author information
is available at the end of the
article

Abstract

Background: Viral infections are causing significant morbidity and mortality worldwide. Understanding the interaction patterns between a particular virus and human proteins plays a crucial role in unveiling the underlying mechanism of viral infection and pathogenesis. This could further help in prevention and treatment of virus-related diseases. However, the task of predicting protein–protein interactions between a new virus and human cells is extremely challenging due to scarce data on virus-human interactions and fast mutation rates of most viruses.

Results: We developed a *multitask transfer learning* approach that exploits the information of around 24 million protein sequences and the interaction patterns from the human interactome to counter the problem of small training datasets. Instead of using hand-crafted protein features, we utilize statistically rich protein representations learned by a deep language modeling approach from a massive source of protein sequences. Additionally, we employ an additional objective which aims to maximize the probability of observing human protein–protein interactions. This additional task objective acts as a regularizer and also allows to incorporate domain knowledge to inform the virus-human protein–protein interaction prediction model.

Conclusions: Our approach achieved competitive results on 13 benchmark datasets and the case study for the SARS-CoV-2 virus receptor. Experimental results show that our proposed model works effectively for both virus-human and bacteria-human protein–protein interaction prediction tasks. We share our code for reproducibility and future research at https://git.l3s.uni-hannover.de/dong/multitask-transfer.

Keywords: Protein–protein interaction, Human PPI, Virus-human PPI, Multitask, Transfer learning, Protein embedding

Introduction

Virus infections cause an enormous and ever increasing burden on healthcare systems worldwide. The ongoing COVID-19 pandemic caused by the zoonotic virus, SARS-CoV-2, has resulted in enormous socio-economic losses [1]. Viruses infect all life forms and require host cells to complete their replication cycle by utilizing the host cell machinery. Virus infection involves several types of protein–protein interactions (PPIs)

between the virus and its host. These interactions include the initial attachment of virus coat or envelope proteins to host membrane receptors, hijacking of the host translation and intracellular transport machineries resulting in replication, assembly and subsequent release of virus particles [2–4]. Besides providing mechanistic insights into the biology of infection, knowledge of virus-host interactions can point to essential events needed for virus entry, replication, or spread, which can be potential targets for the prevention, or treatment of virus-induced diseases [5].

In vitro experiments based on yeast-two hybrid (Y2H), ligand-based capture MS, proximity labeling MS, and protein arrays have identified tens of thousands of virus-human protein interactions [6–14]. These interaction data are deposited in publicly available databases including InAct [15], VirusMetha [16], VirusMINT [17], and HPIDB [18], and others. However, experimental approaches to unravel PPIs are limited by several factors, including the cost and time required, the generation, cultivation and purification of appropriate virus strains, the availability of recombinantly expressed proteins, generation of knock in or overexpression cell lines, availability of antibodies and cellular model systems. Computational approaches can assist *in vitro* experimentation by providing a list of most probable interactions, which actual biological experimentation techniques can falsify or verify.

In this work, we cast the problem of predicting virus-human protein interactions as a binary classification problem and focus specifically on emerging viruses that has limited experimentally verified interaction data.

Key challenges in learning to predict virus-human PPI

Limited interaction data. One of the main challenges in tackling the current task as a learning problem is the *limited training data*. Towards predicting virus-host PPI, some known interactions of other human viruses collected from wet-lab experiments are employed as training data. The number of known PPIs is usually too small and thus, not representative enough to ensure the generalizability of trained models. In effect, the trained models might overfit the training data and would give inaccurate predictions for any given new virus.

Difference to other pathogens. A natural strategy to overcome the limitation posed by scarce virus protein interaction data is to employ transfer learning from available intra-species PPI or PPI data for other types of pathogens. This may, in its simplest fashion, not be a viable strategy as virus proteins can differ substantially from human or bacterial proteins. Typically, they are highly structurally and functionally dynamic. Virus proteins often have multiple independent functions so that they cannot be easily detected by common sequence-structure comparison [19–21]. Besides, virus protein sequences of different species are highly diverse [22]. Consequently, models trained for intra-species human PPI [23–27] or for other pathogen-human PPI [28–33] cannot be directly used to predict virus-human protein interactions.

Limited information on structure and function of virus proteins. While for human proteins, researchers can retrieve information from many publicly available databases to extract features related to their function, semantic annotation, domains, structure, pathway association, and intercellular localization, such information is not readily available for most virus proteins. Protein crystal structures are available for some virus proteins.

However, for many, predictive structures based on the amino acid sequence must be used. Thus, for the majority of virus proteins, currently, the only reliable source of virus protein information is its amino acid sequence. *Learning effective representations* of the virus proteins, therefore, is an important step towards building prediction models. Heuristics such as K-mer amino acid composition are bound to fail as it is known that virus proteins with completely different sequences might show similar interaction patterns.

Our contributions

In this work, we develop a machine learning model which overcomes the above limitations in two main steps, which are described below.

Transfer Learning via protein sequence representations. Though the training data on interactions as well as the input information on protein features are limited, a large number of unannotated protein sequences are available in public databases like UniProt. Inspired by advancements in Natural Language Processing, Alley et al. [34] trained a deep learning model on more than 24 million protein sequences to extract statistically meaningful representations. These representations have been shown to advance the state-of-the-art in protein structure and function prediction tasks. Rather than using hand-crafted protein sequence features, we use the pre-trained model by [34] (referred to as UNIREP) to extract protein representations. The idea here is to exploit transfer learning from several million sequences to our scant training data.

Incorporating domain information. We further fine-tune UNIREP's globally trained protein representations using a simple neural network whose parameters are learned using a multitask objective. In particular, besides the main task, our model is additionally regularized by another objective, namely predicting interactions among human proteins. The additional objective allows us to encode (human) protein similarities dictated by their interaction patterns. The rationale behind encoding such knowledge in the learnt representation is that the human proteins sharing similar biological properties and functions would also exhibit similar interacting patterns with viral proteins. Using a simpler model and an additional side task helps us overcome overfitting, which is usually associated with models trained with small amounts of training data.

We refer to our model as MULTITASK TRANSFER (MTT) and is further illustrated in "Method" section. To sum up, we make the following contributions.

- We propose a new model that employs a transfer learning-based approach to first obtain the statistically rich protein representations and then further refines them using a multitask objective.
- We evaluated our approach on several benchmark datasets of different types for virus-human and bacteria-human protein interaction prediction. Our experimental results (c.f. "Result analysis" section) show that MTT outperforms several baselines even on datasets with rich feature information.
- Experimental results on the SARS-CoV-2 virus receptor shows that our model can help researchers to reduce the search space for yet unknown virus receptors effectively.
- We release our code for reproducibility and further development at https://git.l3s. uni-hannover.de/dong/multitask-transfer.

Related work

Existing work mainly casts PPI prediction task as a supervised machine learning problem. Nevertheless, the information about non-interacting protein pairs is usually not available in public databases. Therefore, researchers can only either adapt models to learn from only positive samples or employ certain negative sampling strategy to generate negative examples for training data. Since the quality and quantity of the generated negative samples would significantly affect the outcome of the learned models, the authors in [31, 35, 36] proposed models that only learned from the available known positive interactions. Nourani et al. [36] and Li et al. [31] treated the virus-human PPI problem as a matrix completion problem in which the goal was to predict the missing entries in the interaction matrix. Nouretdinov et al. [35] use a conformal method to calculate p-values/confidence level related to the hypothesis that two proteins interact based on similarity measures between proteins.

Another line of work which casts the problem as a binary classification task focussed on proposing new negative sampling techniques. For instance, Eid et al [22] proposed Denovo—a negative sampling technique based on virus sequence dissimilarity. Mei et al. [37] proposed a negative sampling technique based on one class SVM. Basit et al. [33] offered a modification to the Denovo technique by assigning sample weights to negative examples inversely proportional to their similarity to known positive examples during training.

Dick et al. [30] utilizes the interaction pattern from intra-species PPI networks to predict the inter-species PPI between human-HIV-1 virus and human. Though the results are promising, this cannot be directly applied to completely new viruses where information about closely-related species is not available or to viruses whose intra-species PPI information is not available.

The works presented in [38–44] employed different feature extraction strategies to represent a virus-human protein pair as a fixed-length vector of features extracted from their protein sequences. Instead of hard-coding sequence feature, Yang et al. [45] and Lanchantin et al. [46] proposed embedding models to learn the virus and human proteins' feature representations from their sequences. However, their training data was limited to around 500,000 protein sequences. Though not very common, other types of information/features were also used in some proposed models besides sequence-based features. Those include protein functional information (or GO annotation) as in [47], proteins domain-domain associations information as in [48], protein structure information as in [32, 49], and the disease phenotype of clinical symptoms as in [47]. One limitation of these approaches is that they cannot be generalized to novel viruses where such kind of information is not available.

Among the network-based approaches, Liu et al. and Wang et al. [50, 51] constructed heterogeneous networks to compute virus and human proteins features. Nodes of the same type were connected by either weighted edges based on their sequence similarity or a combination of sequence similarity and Gaussian Interaction Profile kernel similarity. Deng et al. [43] proposed a deep-learning-based model with a complex architecture of convolutional and LSTM layers to learn the hidden representation of virus and human proteins from their input sequence features along with

the classification problem. Despite the promising performance, those studies still have the limitation posed by hand-crafted protein features.

Method

We first provide a formal problem statement.

Problem statement. We are given protein sequences corresponding to infectious viruses and their known interactions with human proteins. Given a completely new (novel) virus, its set of protein(s) V along with its (their) sequence(s), we are interested in predicting potential interactions between V and the human proteins.

We cast the above problem as that of binary classification. The positive samples consist of pairs of virus and human proteins whose interaction has been verified experimentally. All other pairs are considered to be non-interacting and constitute the negative samples. In "Data description and experimental set up" section, we add details on positive and negative samples corresponding to each dataset.

Summary of the approach. The schematic diagram of our proposed model is presented in Fig. 1. As shown in the diagram, the input to the model is the raw human and virus protein sequences which are passed through the UniRep model to extract low dimensional vector representations of the corresponding proteins. The extracted embeddings are then passed as initialization values for the embedding layers. These representations are further fine-tuned using the Multilayer Perceptron (MLP) modules (shown in blue). The fine-tuning is performed while learning to predict an
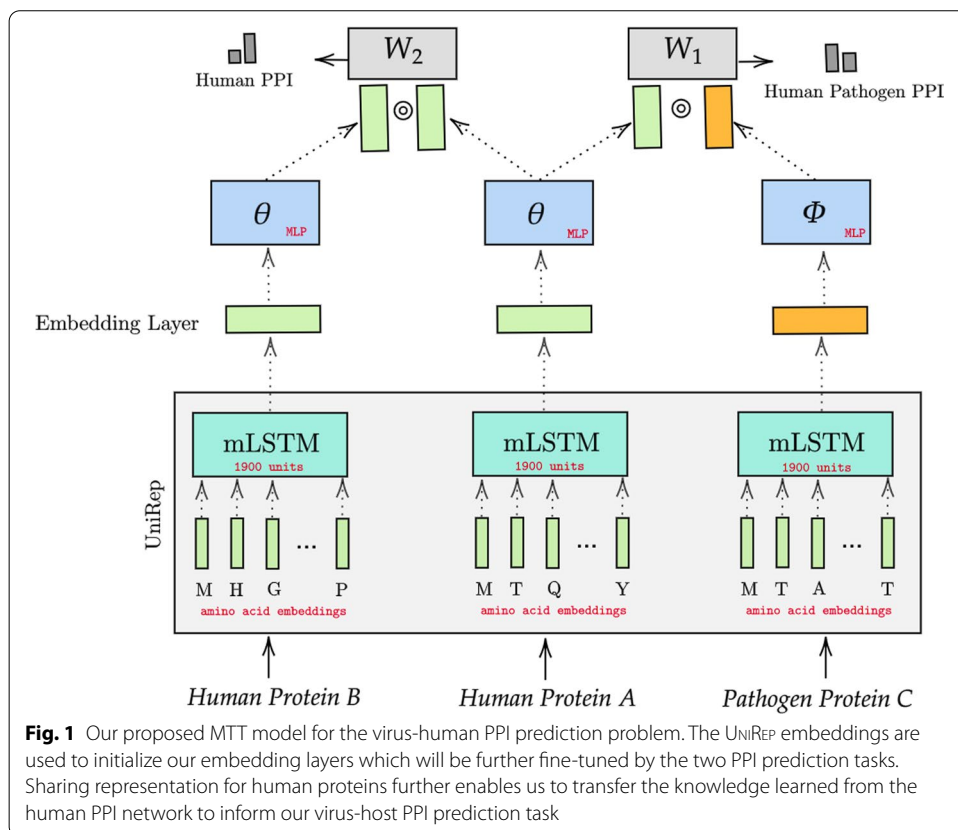

Fig. 1 Our proposed MTT model for the virus-human PPI prediction problem. The UɴɪRᴇᴘ embeddings are used to initialize our embedding layers which will be further fine-tuned by the two PPI prediction tasks. Sharing representation for human proteins further enables us to transfer the knowledge learned from the human PPI network to inform our virus-host PPI prediction task

interaction between two human proteins (between proteins A and B in the figure) as well as the interaction between human and virus proteins (between proteins B and C). In the following, we describe in detail the main components of our approach.

Extracting protein representations

Significance of using protein sequence as input. We note that the protein sequence determines the protein's structural conformation (fold), which further determines its function and its interaction pattern with other proteins. However, the underlying mechanism of the sequence-to-structure matching process is very complex and cannot be easily specified by hand-crafted rules. Therefore, rather than using hand-crafted features extracted from amino acid sequences, we employ the pre-trained UNIREP model [34] to generate latent representations or protein embeddings. The protein representations extracted from UNIREP model are empirically shown to preserve fundamental properties of the proteins and are hypothesized to be statistically more robust and generalizable than hand-crafted sequence features.

UNIREP for extracting sequence representations. In particular, UNIREP consists of an embedding layer that serves as a lookup table for each amino acid representation. Each amino acid is represented as an embedding vector of 10 dimensions. Each input protein sequence of length N will be denoted as a two-dimensional matrix of size Nx10. That two-dimensional matrix will then feed as input to a Multiplicative Long Short Term Memory (mLSTM) network of 1900 units. The 1900 dimension is selected experimentally from a pool of architectures that require different numbers of parameters as described in [52], namely, a 1900-dimensional single layer multiplicative LSTM ($\sim$ 18.2 million parameters), a 4-layer stacked mLSTM of 256 dimensions per layer ($\sim$ 1.8 million parameters), and a 4-layer stacked mLSTM with 64 dimensions per layer ($\sim$ 0.15 million parameters). The output from mLSTM is a 1900 dimensional embedding vector that serves as the pre-trained protein embedding for the input protein sequence. We use the calculated pre-trained virus and human protein embeddings to initialize our embedding layers. The two supervised PPI prediction tasks will further fine-tune those embeddings during training.

Learning framework

We further fine-tune these representations by training two simple neural networks (single layer MLP with ReLu activation) using an additional objective of predicting human PPI in addition to the main task. More precisely, the UNIREP representations will be passed through one hidden layer MLPs with ReLU activations to extract the latent representations. Let $\mathbf{X}$ denote the embedding lookup matrix. The ith row corresponds to the embedding vector of node i. The final output from MLP layers for an input v is then given by $\mathbf{hid}(v) = \mathrm{MLP}(\mathbf{X}(v))$. To predict the likelihood of interaction between a pair (v_1, v_2) we first perform an element-wise product of the corresponding hidden vectors (output of MLPs) and pass it through a linear layer followed by sigmoid activation. In the following we provide a detailed description of our multi-task objective.

Training using a multi-task objective

Let Θ, Φ denote the set of learnable parameters corresponding to fine-tuning components (as shown in Fig. 1 in green and blue boxes), i.e., the Multilayer Perceptrons (MLP) corresponding to the virus and human proteins, respectively. Let $\mathbf{W}_1, \mathbf{W}_2$ denote the two learnable weight matrices (parameters) for the linear layers (as depicted in gray boxes in the Figure). We use *VH*, and *HH* to denote the training set of virus-human, human-human PPI, correspondingly. We use binary cross entropy loss for predicting virus-human PPI predictions, as given below:

$$\mathcal{L}_1 = \sum_{(v,h)\in VH} -z_{vh} \log y_{vh}(\Theta, \Phi, \mathbf{W}_1) - (1 - z_{vh}) \log(1 - y_{vh}(\Theta, \Phi, \mathbf{W}_1)), \tag{1}$$

where variables z_{vh} is the corresponding binary target variable and y_{vh} is the predicted likelihood of observing virus-human protein interaction, i.e.,

$$y_{vh}(\Theta, \Phi, \mathbf{W}_1) = \sigma((\mathbf{hid}(v) \odot \mathbf{hid}(h))\mathbf{W}_1), \tag{2}$$

where $\sigma(x) = 1/1 + e^{-x}$ is the sigmoid activation and $\odot$ denotes the element-wise product.

For human PPI, we predict the confidence score of observing an interaction between two human proteins. More specifically, we directly predict $z_{hh'}$—the normalized confidence scores for interaction between two human proteins as collected from STRING [53] database. Predicting the normalized confidence scores helps us overcome the issues with defining negative interactions. We use mean square error loss to compute the loss for the human PPI prediction task as below where $y_{hh'}$ is computed similar to (2) for human proteins and N is the number of (h, h') pairs.

$$\mathcal{L}_2 = \frac{1}{N} \sum_{(h,h')\in HH} (y_{hh'}(\Theta, \mathbf{W}_2) - z_{hh'})^2 \tag{3}$$

We use a linear combination of the two loss functions to train our model.

$$\mathcal{L} = \mathcal{L}_1 + \alpha \cdot \mathcal{L}_2 \tag{4}$$

where α is the human PPI weight factor.

Data description and experimental set up

We commence by describing the 13 datasets used in this work to evaluate our approach.

Benchmark datasets

The realistic host cell-virus testing datasets

The Novel H1N1 and Novel Ebola datasets. We retrieve the curated or experimentally verified PPIs between virus and human from four databases: APID [54], IntAct [15], VirusMetha [16], and UniProt [55] using the PSICQUIC web service [56]. In total, there are 11,491 known PPIs between 246 viruses and humans. From this source of data, we generate new training and testing data for the two viruses: the human H1N1 Influenza virus and Ebola virus. We name the two datasets Novel H1N1 and Novel Ebola

according to the virus present in the testing set. The positive training data for the NOVEL H1N1 dataset includes PPIs between human and all viruses except H1N1. Similarly, the positive training data for the NOVEL EBOLA dataset includes PPIs between human and all viruses except Ebola. The positive testing data for the human-H1N1 dataset contains PPIs between human and 11 H1N1 virus proteins. Likewise, the positive testing data for the human-Ebola dataset contains PPIs between human and three of the eight Ebola virus proteins (VP24, VP35, and VP40).

Negative sampling techniques such as the dissimilarity-based method [22], the exclusive co-localization method [57, 58] are usually biased as they restrict the number of tested human proteins. It is also unrealistic for a new virus because information about such restricted human protein set, generated from filtering criteria based on the positive instances, is typically unavailable. For those reasons, we argue that random negative sampling is the most appropriate, unbiased approach to generate negative training/testing samples. Since the exact ratio of positive:negative is unknown, we conducted experiments with different negative sample rates. In our new virus-human PPI experiments, we try four negative sample rates: [1,2,5,10]. In addition, to reduce the bias of negative samples, the negative sampling in the training and testing set is repeated ten times. In the end, for each dataset, we test each method with 4x4x10 = 160 different combinations of negative training and negative testing sets (with fixed positive training and test samples). The statistics for our new testing datasets are given in Table 1.

The DEEPVIRAL [47] **Leave-One-Species-Out (LOSO) benchmark datasets**. The data was retrieved from the HPIDB database [18] to include all *Pathogen-Host* interactions that have confidence scores available and are associated with an existing virus family in the NCBI taxonomy [59]. After filtering, the dataset includes 24,678 positive interactions and 1,066 virus proteins from 14 virus families. We follow the same procedure as mentioned in [47] to generate the training and testing data corresponding to four virus species with taxon IDs: 644788 (Influenza A), 333761 (HPV 18), 2697049 (SARS-CoV-2), 2043570 (Zika virus). From now on, we will use the NCBI taxon ID of the virus species in the testing set as the dataset name. For each dataset, the positive testing data consists of all known interactions between the test virus and the human proteins. The negative testing data consists of all possible combinations of virus and 16,627 human

Table 1 The virus-human PPI realistic benchmark datasets' statistics

	TRAINING DATA				TESTING DATA			
	$|E^+|$	$|E^-|$	$|V^h|$	$|V^v|$	$|E^+|$	$|E^-|$	$|V^h|$	$|V^v|$
NOVEL H1N1	10,858	Varies	7636	641	381	Varies	622	11
NOVEL EBOLA	11,341	Varies	7816	659	150	Varies	290	3
ZHOU'S H1N1	10,858	10,858	7636	641	381	381	622	11
ZHOU'S EBOLA	11,341	11,341	7816	659	150	150	290	3
2697049	24,698	246,980	16,638	1066	278	448,651	16,627	27
333761	23,892	238,920	16,638	1070	534	132,482	16,627	8
2043570	24,372	243,720	16,638	1085	309	66,199	16,627	4
644788	24,825	248,250	16,638	1090	54	33,200	16,627	2

$|E^+|$ and $|E^-|$ refer to the number of positive and negative interactions, respectively. $|V^h|$ and $|V^v|$ are the number of human proteins and virus proteins

proteins in Uniprot (with a length limit of 1000 amino acids) that do not appear in the positive testing set. Similarly, the positive training data consists of all known interactions between human protein and any virus protein, except for the one which is in the testing set. The negative training data is generated randomly with the positive:negative rate of 1:10 from the pool of all possible combinations of virus and 16,627 human proteins that do not appear in the positive training set. Statistics of the datasets are presented in Table 1. Though performing a search on the set of 16,627 human proteins might not be a fruitful realistic strategy, we still keep the same training and testing data as released in the DEEPVIRAL study in our experiments to have a direct and fair comparison with the DEEPVIRAL method.

The widely used new virus-human PPI prediction benchmarked datasets

The two datasets released by Zhou et al. [41] are widely used by recent papers to evaluate state-of-the-art models on new virus-human PPI prediction tasks. We refer to them as ZHOU's H1N1 and ZHOU's EBOLA where each dataset was named after the viruses in the testing sets. ZHOU's H1N1 and ZHOU's EBOLA share similar positive training and testing samples with the NOVEL H1N1 and NOVEL EBOLA datasets. However, they differ in the negative training and testing samples sets. While the negative samples in NOVEL H1N1 and NOVEL EBOLA were generated randomly from the pool of all possible pairs, the negative training/testing samples in ZHOU's H1N1 and ZHOU's EBOLA were generated based on the protein sequence dissimilarity score. Therefore, ZHOU's H1N1 and ZHOU's EBOLA have the limitations as mentioned in "The realistic host cell-virus testing datasets" section and are not ideal for evaluating the new virus-human PPI prediction task. The data statistics for these two datasets are shown in Table 1.

The specialized testing datasets

The dataset with protein motif information (DENOVO SLIM [22]). The DENOVO SLIM dataset Virus-human PPIs were collected from VirusMentha database [16]. The presence of Short Linear Motif (SLiM) in virus sequences was used as a criterion for data filtering. SLiMs are short, recurring patterns of protein sequences that are believed to mediate protein–protein interaction [60, 61]. Therefore, sequence motifs can be a rich feature set for virus-human PPI prediction tasks. The test set [22] contained 425 positives and 425 negative PPIs (Supplementary file S12 used in DeNovo's study ST6). The training data consisted of the remaining PPI records and comprised of 1590 positive and 1515 negative records for which virus SLiM sequence is known and 3430 positives and 3219 negatives without virus SLiM sequences information. DENOVO_SLIM negative samples were also generated using the Denovo negative sampling strategy (based on sequence dissimilarity).

   The BARMAN's dataset [48] **with protein domain information**. The dataset was retrieved from VirusMINT database [17]. Interacting protein pairs that did not have any "InterPro" domain hit were removed. In the end, the dataset contained 1035 positives and 1035 negative interactions between 160 virus proteins of 65 types and 667 human proteins. 5-Fold cross-validation was then employed to test each method's performance.

The bacteria human PPI prediction task

We evaluate our method on three datasets for three human pathogenic bacteria: Bacillus anthracis (B1), Yersinia pestis (B2), and Francisella tularensis (B3), which were shared by Fatma et al. [22].

The data was first collected from HPIDB [18]. B1 belongs to a bacterial phylum different from that of B2 and B3, while B2 and B3 share the same class but differ in their taxonomic order. B1 has 3057 PPIs, B2 has 4020, and B3 has 1346 known PPIs. A sequence-dissimilarity-based negative sampling method was employed to generate negative samples. For each bacteria protein, ten negative samples were generated randomly. Each of the bacteria was then set aside for testing, while the interactions from the other two bacteria were used for training. For simplicity, we use the name of the bacteria in the testing set as the name of the dataset. The statistics for those three datasets are presented in Table 2.

Description of compared methods

We compare our method with the following seven baseline methods and two simper variants of our model.

- Generalized [41]: It is a generalized SVM model trained on hand-crafted features extracted from protein sequence for the novel virus-human PPI task. Each virus-human pair is represented as a vector of 1175 dimensions extracted from the two protein sequences.
- Hybrid [43]: It is a complex deep model with convolutional and LSTM layers for extracting latent representation of virus and human proteins from their input sequence features and is trained using L1 regularized Logistic regression.
- doc2vec [45]: It employs the doc2vec [62] approach to generate protein embeddings from the corpus of protein sequences. A random forest model is then trained for the PPI prediction.
- MotifTransformer [46]: It is a transformer-based deep neural network that pretrains protein sequence representations using unsupervised language modeling tasks and supervised protein structure and function prediction tasks. These representations are used as input to an order-independent classifier for the PPI prediction task.
- DeNovo [22]: This model trained an SVM classifier on a hand-crafted feature set extracted from the K-mer amino acid composition information using a novel negative sampling strategy. Each protein pair is represented as a vector of 686 dimensions.

Table 2 Our bacteria-human PPI benchmark datasets' statistics

	Training data				Testing data			
	$|E^+|$	$|E^-|$	$|V^h|$	$|V^b|$	$|E^+|$	$|E^-|$	$|V^h|$	$|V^b|$
Bacillus anthracis	5366	15,590	1559	2674	3057	9440	944	1705
Yersinia pestis	4403	12,880	1288	2278	4020	12,150	1215	2147
Francisella tularensis	7077	21,590	2159	3041	1346	3440	344	1023

$|E^+|$ and $|E^-|$ refer to the number of positive and negative interactions, respectively. $|V^h|$ and $|V^b|$ are the number of human proteins and bacteria proteins

- DEEPVIRAL [47]: It is a deep learning-based method that combines information from various sources, namely, the disease phenotypes, virus taxonomic tree, protein GO annotation, and proteins sequences for intra- and inter-species PPI prediction.
- BARMAN [48]: It used an SVM model trained on a feature set consisting of the protein domain-domain association and methionine, serine, and valine amino acid composition of viral proteins.
- 2 simpler variants of MTT: Towards ablation study, we evaluate two simpler variants: (i) SINGLETASK TRANSFER (STT), which is trained on a single objective of predicting pathogen-human PPI. STT is basically the MTT without the human PPI prediction side task and (ii) NAIVE BASELINE, which is a Logistic regression model using concatenated human and pathogen protein UNIREP representations as input.

Implementation details and parameter set up

We use Pytorch [63] to implement our model and run it on an Nvidia GTX 1080-Ti with 11GB memory. We use Adam optimizer for the model parameter optimization. For all datasets, we left out 10% of the training data for validation and performed a grid search for the best combination of parameters on that validation set. For datasets other than NOVEL H1N1 and NOVEL EBOLA, we perform parameter grid searching with the MLP hidden dimension *hid* in [8, 16,32, 64], α in $[10^{-3}, 10^{-2}, 10^{-1}, 1]$, the number of *epochs* from 0 to 200 with a step of 2 and the learning rate *lr* in $[10^{-3}, 10^{-2}]$. For the NOVEL H1N1 and NOVEL EBOLA datasets, we test each with 160 different combinations of negative training and negative testing. Therefore, we fix the hidden dimension to 16, $\alpha = 10^{-3}$, $lr = 10^{-3}$ and only perform grid searching on the number of epochs. The reported results for each dataset are the results corresponding to the best-performed model on the validation set.

For the DOC2VEC model, we use the released code shared by the authors with the given parameters. For the GENERALIZED and DENOVO models, we re-implement the methods in Python using all the parameters and feature set as described in the original papers. For BARMAN and DEEPVIRAL, the results are taken from the original papers or calculated from the given model prediction scores.

Evaluation metrics

For all benchmark datasets except the case study, we report five metrics: the Area under Receiver Operating Characteristic curve (AUC) and the area under the precision-recall curve (AP), the PRECISION, RECALL, and F1 scores.

For the case study, we report the topK score with *K* from 1 to 10. TopK is equal to 1 if the human receptor for SARS-CoV-2 virus appears in the top *K* proteins that have the highest scores predicted by the model and 0 otherwise.

Result analysis

In the following four subsections, we provide a detailed comparison of MTT with (i) methods employing hand-crafted input features, (ii) sequence embedding-based methods, (iii) an approach that uses protein domain information, (iv) simpler variants of MTT as ablation studies respectively. All statistical test results present in this section are
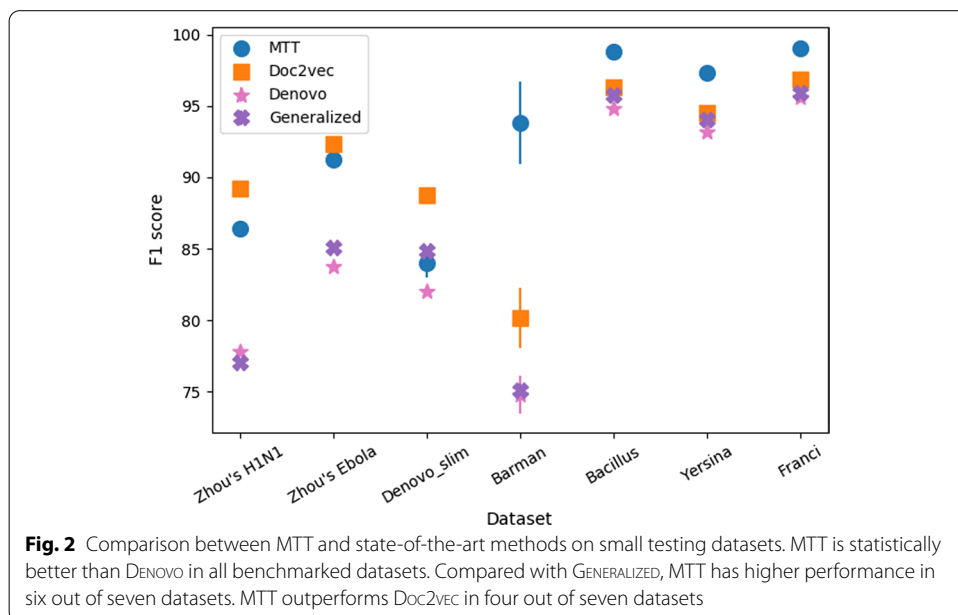
those from the pair-wise t-test [64] on the F1 scores attained from multiple runs on the same dataset.

Comparison with methods employing hand-crafted features

GENERALIZED [41] and DENOVO [22] are the two traditional methods relying on hand-crafted features extracted from the protein sequences. The number of hand-crafted features employed by DENOVO and GENERALIZED are 686 and 1175, respectively. They both employ SVM for the classification task. Since SVM scales quadratically with the number of data points, DENOVO and GENERALIZED are not scalable to larger datasets.

Figure 2 presents their comparison between MTT on small testing datasets. Detailed scores are given in Table 6 in the Appendix. Results from the two-tailed t-test [65, 66] support that MTT significantly outperforms DENOVO in all benchmarked datasets with a confidence score of at least 95%. Compared with GENERALIZED, MTT has higher performance in six out of seven datasets (except DENOVO_SLIM). The difference is the most significant on the BARMAN, ZHOU's H1N1, and ZHOU's EBOLA datasets. On DENOVO_SLIM dataset, MTT 's F1 score is lower than GENERALIZED and only 2% higher than DENOVO. This is expected since DENOVO_SLIM is a specialized dataset favoring methods using local sequence motif features, which are exploited by DENOVO and GENERALIZED.

HYBRID is one recently proposed, deep learning-based method. Despite that, the input features are still manually extracted from the protein sequence. Since the code is not publicly available, we only have the AUC score corresponding to the ZHOU's H1N1 dataset, which is also taken from the original paper as listed in Table 6. Compared with HYBRID, MTT has higher AUC score. Though comparison on the AUC for one dataset does not bring much insight, we include this method here for completeness.

Fig. 2 Comparison between MTT and state-of-the-art methods on small testing datasets. MTT is statistically better than DENOVO in all benchmarked datasets. Compared with GENERALIZED, MTT has higher performance in six out of seven datasets. MTT outperforms DOC2VEC in four out of seven datasets

Comparison with sequence embedding based methods

Doc2vec and MotifTransformer are state-of-the-art methods based on sequence embeddings or representations. Doc2vec utilizes the embeddings learned from the extracted k-mer features while MTT and MotifTransformer employ the embedding directly learned from the amino acid sequences. In addition, MTT is a multitask-based approach that incorporates additional information on human protein–protein interaction into the learning process.

Figure 3 shows a comparison in F1 score of MTT and Doc2vec over all benchmarked datasets. Detailed scores are presented in Table 7 in the Appendix. Since the code for the MotifTransformer model is not publicly available, we only have the corresponding results available for the Zhou's H1N1 and Zhou's Ebola datasets, which are also taken from the original paper. '-' denotes the score is not available. Compared with MotifTransformer, MTT has a slightly worse F1 score on Zhou's H1N1 and significantly better F1 score on Zhou's Ebola datasets.

Comparison with Doc2vec. MTT out-performs Doc2vec in 5 out of 9 benchmark datasets, and the performance gap is statistically significant with a p-value smaller than 0.05. MTT is significantly better than Doc2vec on the Novel Ebola dataset, while on the Novel H1N1 dataset, the reverse holds true. Doc2vec outperforms MTT in three testing datasets whose negative samples were drawn from a sequence dissimilarity method. We also note that these datasets might be biased since in the ideal testing scenario, we do not have knowledge about the set of human proteins that interacted with the virus. Therefore, such dissimilarity-based negative sampling is infeasible.

(a) Novel Ebola

(b) Novel H1N1

Fig. 3 Comparison between MTT and state-of-the-art methods on the Novel Ebola and Novel H1N1 datasets over different combinations of negative training and testing sets. MTT is significantly better than Doc2vec on the Novel Ebola dataset (**a**), while on the Novel H1N1 dataset (**b**), the reverse holds true. MTT is statistically better than Denovo on both datasets. For the Generalized model, we can only have results up to the negative training rate of 2 because, for larger negative training rates, the model took days to finish one run
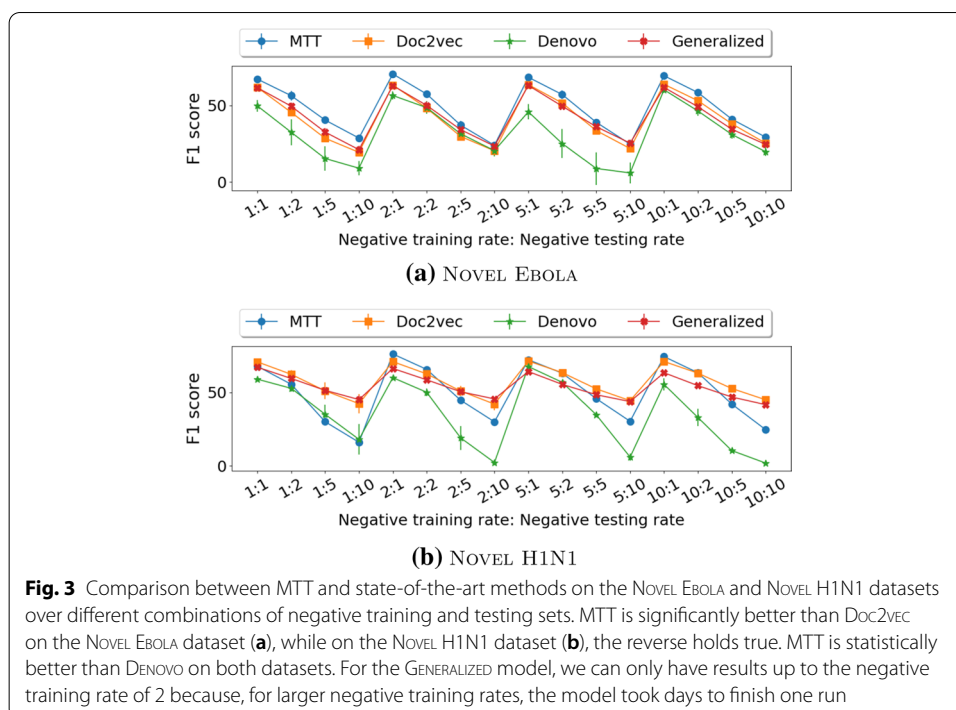
Table 3 Comparison between MTT and Barman—a method that relies on the protein domain information

Model	AUC	AP	Precision	Recall	F1
Barman	0.7300	–	–	67.00	69.41
MTT	0.9804	0.9802	93.53	94.05	93.79

Due to data and code availability issues, for the Barman method, we only have results for the Barman's dataset, which are also taken from the original paper. '−' indicates that the result is not available

Table 4 Comparison with DeepViral—a method that can utilize knowledge from the disease phenotype, virus taxonomy, the human PPI network, and the protein GO annotation

Dataset	Model	AUC	AP	Precision	Recall	F1
2697049	DeepViral	0.7288	0.0015	0.07	0.07	0.07
	MTT	**0.7566**	**0.0021**	**0.97**	**0.97**	**0.97**
333761	DeepViral	0.8009	0.0147	1.72	1.72	1.72
	MTT	**0.8160**	**0.0262**	**6.35**	**6.35**	**6.35**
2043570	DeepViral	**0.7708**	**0.0116**	0.52	0.52	0.52
	MTT	0.6956	0.0096	**1.89**	**1.91**	**1.90**
644788	DeepViral	0.9325	**0.0357**	**3.70**	3.70	3.70
	MTT	**0.9537**	0.0302	3.54	**22.04**	**5.46**

The bold font is used to highlight highest scores corresponding to each dataset

Results from the pair-wise t-test indicate that MTT is significantly better than DeepViral on three datasets (2697049, 333761, and 2043570) with a p-value smaller than 0.05. On the 644788 dataset, the difference is not statistically significant

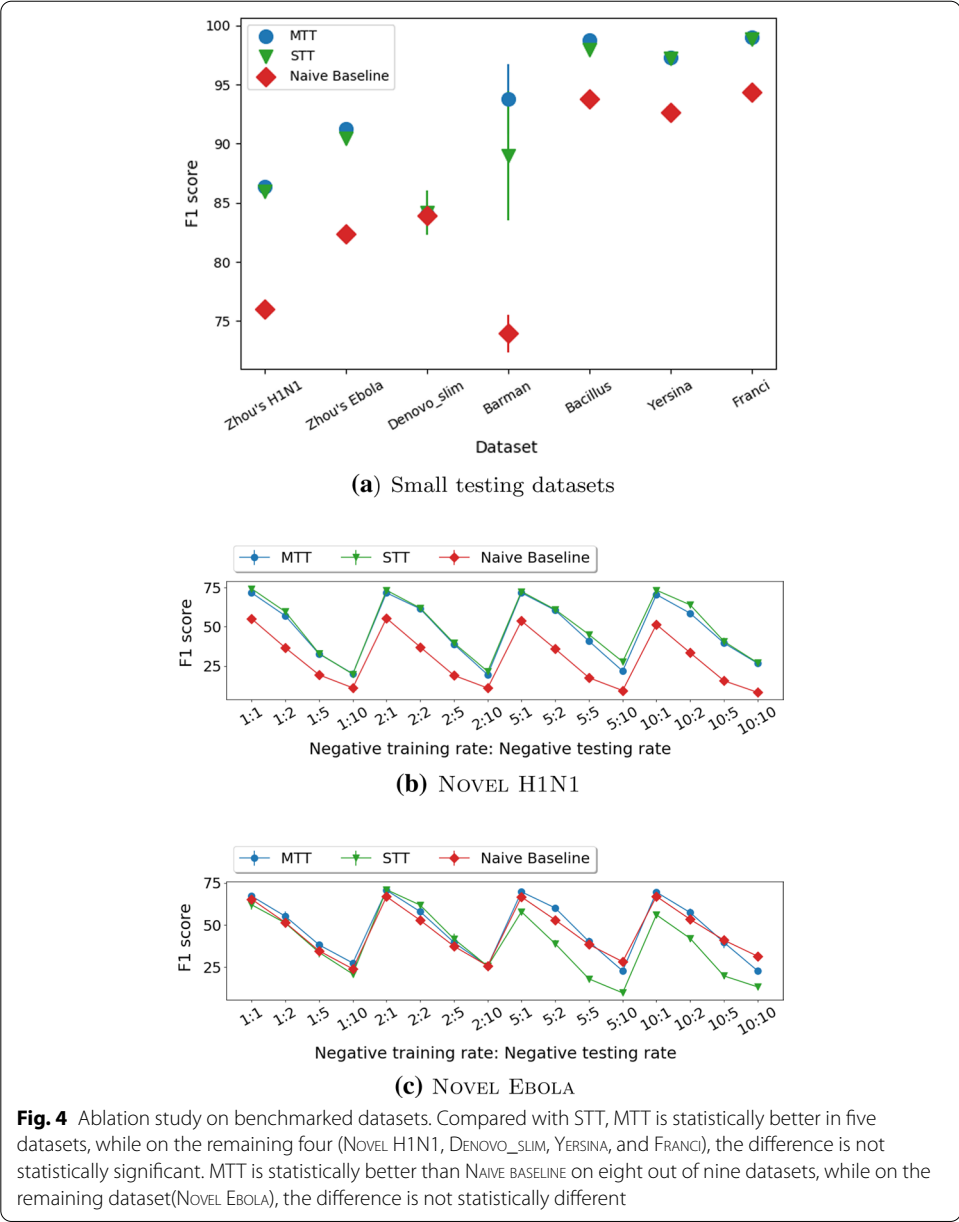### Comparison with methods that use domain information

Barman features set is constructed from the domain-domain association and the hand-crafted feature extracted from the protein sequences. Since the protein domain information is not available for all viral proteins, the Barman method has restricted application. A comparison between Barman and MTT is presented in Table 3. Due to data and code availability, we only have the results for the Barman model on one dataset. From reported results, we could clearly see that MTT outperforms its competitor for a large margin in all available metrics.

Comparison with methods that used GO, taxonomy and phenotype information

DeepViral exploited that disease phenotypes, the viral taxonomies, and proteins' GO annotation to enrich its protein embeddings. Table 4 presents a comparison between MTT and DeepViral on the four datasets released by DeepViral's authors. The reported results on each dataset are the average after five experimental runs for DeepViral and ten experimental runs for MTT. We observe MTT and STT significantly supersede their competitor regarding the averaged F1 score. The gain is more significant on smaller datasets (644788 and 333761)

Ablation studies

We compare our method with two of its simpler variants: the STT and the Naive baseline baseline models. STT is the MTT model without the human PPI prediction task. Naive baseline concatenates the learned embeddings for the virus and human proteins to form the input to a Logistic Regression model. Figure 4 presents

(a) Small testing datasets

(b) Novel H1N1

(c) Novel Ebola

Fig. 4 Ablation study on benchmarked datasets. Compared with STT, MTT is statistically better in five datasets, while on the remaining four (Novel H1N1, Denovo_slim, Yersina, and Franci), the difference is not statistically significant. MTT is statistically better than Naive baseline on eight out of nine datasets, while on the remaining dataset(Novel Ebola), the difference is not statistically different

a comparison between the F1 score of MTT and its variants on our benchmarked datasets. Table 8 show all reported scores over all datasets. MTT is significantly better than STT in five out of nine benchmarked and the four DeepViral datasets with a p-value smaller than 0.05. While in the remaining four datasets, the difference is not statistically significant. This confirms that the learned patterns from the human PPI network bring additional benefits to the virus-human PPI prediction task.

Compare with Naive baseline, MTT wins in eight out of nine benchmarked and the four DeepViral datasets. On the remaining dataset (Novel H1N1), the difference is not statistically different. STT significantly outperforms Naive baseline in eight out of nine datasets. This claims the effectiveness of our chosen architecture.

Case study for SARS-CoV-2 binding prediction

The virus binding to cells or the interaction between viral attachment proteins and host cell receptors is the first and decisive step in the virus replication cycle. Identifying the host receptor(s) for a particular virus is often fundamental in unveiling the virus pathogenesis and its species tropism.

Here we present a case study for detecting the human protein binding partners for SARS-CoV-2. Our virus-human PPI dataset is retrieved from the InAct Molecular Interaction database [15] (the latest update is 07.05.2021). We retrieve the protein sequences from Uniprot [55]. In the next section, we describe the construction of the training and testing dataset to predict SARS-CoV-2 binding partners.

Training, validation and test sets for virus-human PPI

The statistics for our SARS-CoV-2 binding prediction dataset are presented in Table 5. We construct the corresponding datasets as follows.

Training set. As positive interaction samples, we include in the training data only *direct* interactions between the human proteins and any virus except the SARS-CoV and SARS-CoV-2. *Direct* interaction requires two proteins to directly bind to each other, i.e. without an additional bridging protein. Moreover, the interacting human protein should be on the cell surface. Without loss of generality, we perform our search for the binding receptor on the set of all human proteins that have a *KNOWN direct interaction* with any virus and *locate* to the cell surface. Our surface human protein list consists of all reviewed Uniprot proteins that meet at least one of the following criteria: (i) appears in the human surfacetome [67] list or (ii) has at least one of the following GO annotations [68, 69]:{*CC-plasma membrane, CC-cell junction*}.

The negative samples for training data contain *indirect* (interactions that are not marked as direct in the database) between the human proteins and any virus except SARS-CoV and SARS-CoV-2. The *indirect* interactions can be a physical association (two proteins are detected in the same protein complex at the same point of time) or an association in which two proteins that may participate in the formation of one or more physical complexes without additional evidence whether the proteins are directly binding to specific members of such a complex).

Validation and test sets. As established in studies [70–72], angiotensin-converting enzyme 2 (ACE2) is the human receptor for both SARS-CoV [73] and SARS-CoV-2 viruses [72]. The positive validation and testing set consist of interaction between the known human receptor (ACE2) and the corresponding spike proteins of SARS-CoV and SARS-CoV-2, respectively. Our negative validation and testing set encapsulate of all

Table 5 The case study statistics

$|V^h|$	$|V^v|$	TRAINING		VALIDATION		TESTING		HUMAN PPI
		$|E^+|$	$|E^-|$	$|E^+|$	$|E^-|$	$|E^+|$	$|E^-|$	$|E|$
5563	834	554	17,418	1	51	1	51	96,459

$|E^+|$ and $|E^-|$ refer to the number of positive and negative interactions, respectively. $|V^h|$ and $|V^v|$ are the number of human proteins and virus proteins

possible combinations the two viral spike proteins and 52 human proteins that meet our filtering criteria.
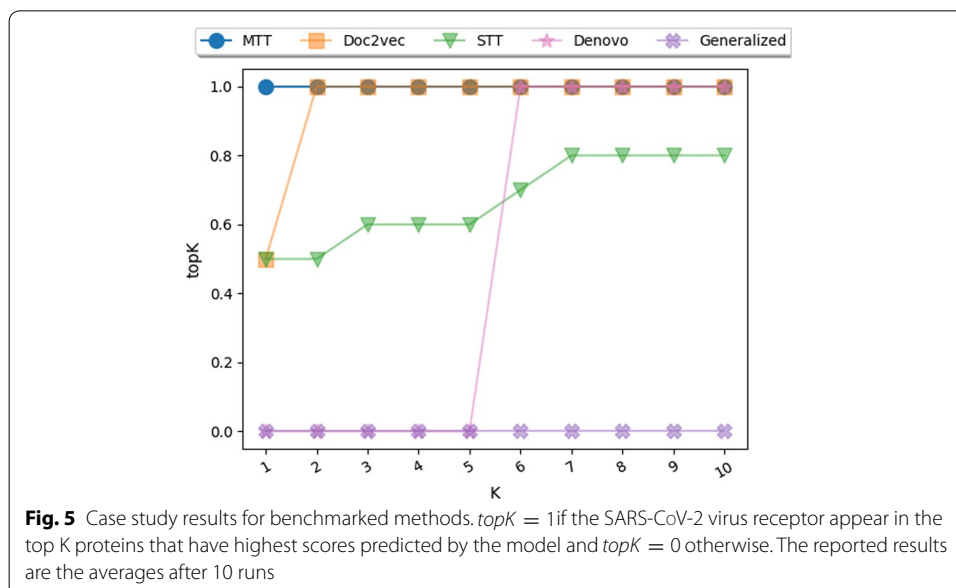
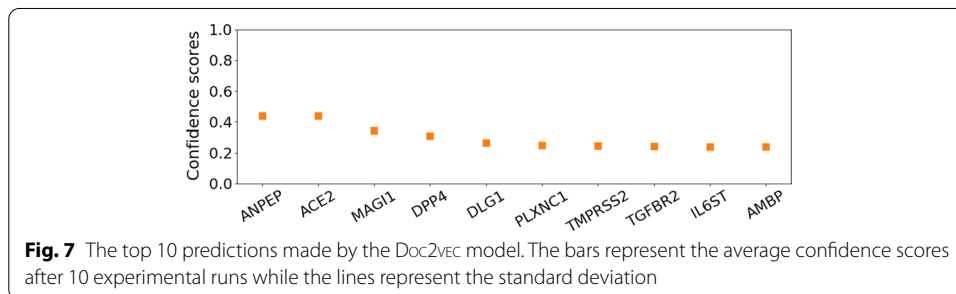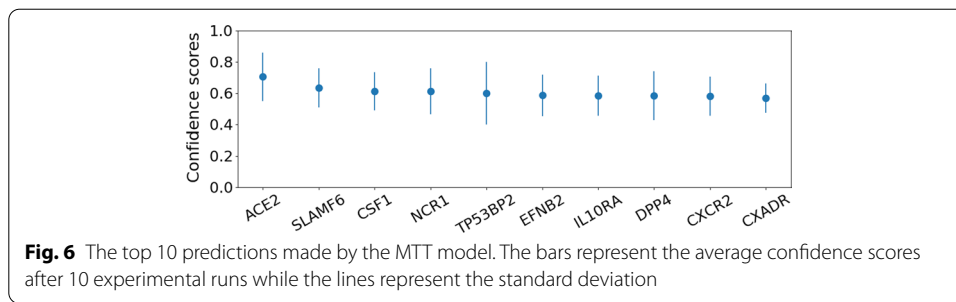### The intra human PPI for the side task

Since we are interested in only the direct interaction between virus and human proteins, we also customize our intra human PPI training set. Our intra human PPI dataset is also retrieved from the InAct [15] database (the latest update is 07.05.2021). We retain only interactions between two human proteins that appear in the virus-human PPI dataset constructed above. The confidence scores are normalized into the [0, 1] ranges. All confidence scores corresponding to "indirect" interactions are set to 0. In the end, our intra-human PPI training set consists of 96,458 interactions between 5563 human proteins.

Results

Finally, we here evaluate the prediction methods on how effective they are in ranking human protein candidates for binding to an emerging virus envelope protein. Figure 5 presents the methods' performance after ten runs on the case study dataset. TopK is equal to 1 if the true human receptor appears in the top K proteins that correspond to the highest predicted scores by the model and is equal to 0 otherwise. The reported scores plotted in Fig. 5 are the average after ten experimental runs with random initialization.

Using this method we find that ACE2, the only SARS-CoV-2 receptor proven in in vivo and in vitro studies [72, 74, 75], consistently appears as the highest ranked prediction of MTT in each of the ten experimental runs. We observe a significant difference between the highest ranked performance of MTT and its competitors. The performance gain shown by MTT over STT is quite substantial after ten runs and supports the superiority of our multitask framework. The next highest nine hits presented in both models have not been shown to interact with SARS-CoV-2 in in vitro studies. Interestingly, dipeptidyl peptidase 4 (DDP4), a receptor for another betacoronavirus MERS-CoV [76]

Fig. 5 Case study results for benchmarked methods. *topK* = 1 if the SARS-CoV-2 virus receptor appear in the top K proteins that have highest scores predicted by the model and *topK* = 0 otherwise. The reported results are the averages after 10 runs

Fig. 6 The top 10 predictions made by the MTT model. The bars represent the average confidence scores after 10 experimental runs while the lines represent the standard deviation

Fig. 7 The top 10 predictions made by the Doc2vec model. The bars represent the average confidence scores after 10 experimental runs while the lines represent the standard deviation

also scored highly in the MTT method. However, although in silico analysis has speculated a possible interaction [77], it is yet to be shown experimentally. Similarly, the serine protease TMPRSS2, which is required for SARS-CoV-2 S protein priming during entry [72], appeared in position 7 using the Doc2vec model. Finally, aminopeptidase N (ANPEP) the receptor for the common cold coronavirus 229E appeared as first hit in the Doc2vec model [78].

In Figures 6 and 7 , we plot the average confidence scores (corresponding to predicted interaction probability) corresponding to top 10 predictions of MTT and Doc2vec models. Specifically, the proteins are ranked based on the average (over 10 runs) confidence scores as predicted by the two models. While for MTT, the receptor ACE2 always occurs at the top of the list with average confidence score of more than 0.70 (which is more than 11% higher than the confidence score assigned to the second hit), Doc2vec assigns it a score of less than 0.44 where ACE2 is ranked 2nd based on average scores. Moreover, there is negligible difference between the prediction scores for ACE2 and the first predicted hit ANPEP in case of Doc2vec.

These results indicate that MTT can provide high-quality prediction results and can help biologists to restrict the search space for the virus interaction partner effectively. This case study showcases the effectiveness of our method in solving virus-human PPI prediction problem and aims to convince biologists of the potential application of our prediction framework.

Conclusion

We presented a thorough overview of state-of-the-art models and their limitations for the task of virus-human PPI prediction. Our proposed approach exploits powerful statistical protein representations derived from a corpus of around 24 Million protein

sequences in a multitask framework. Noting the fact that virus proteins tend to mimic human proteins towards interacting with the host proteins, we use the prediction of human PPI as a side task to regularize our model and improve generalization. The comparison of our method with a variety of state-of-the-art models on several datasets showcase the superiority of our approach. Ablation study results suggest that the human PPI prediction side task brings additional benefits and helps boost the model performance. A case study on the interaction of the SARS-CoV-2 virus spike protein and its human receptor indicates that our model can be used as an effective tool to reduce the search space for evaluating host protein candidates as interacting partners for emerging viruses. In future work, we will enhance our multitask approach by incorporating more domain information including structural protein prediction tools [79] as well as exploiting more complex multitask model architectures.

Appendix

Detailed results

The following subsections provide detailed experimental results. For the Hybrid and MotifTransformer, the author's code is not available and results are taken from the original paper as the. '-' indicates that the score is not available. For other methods, the reported results are the average after 10 experimental runs. We perform pairwise t-test tests for statistical significance testing. Our presented results are statistically significant with a p-value less than 0.05.

Comparison with methods using hand crafted protein features

Table 6 provides a comparison between MTT and baselines which employ hand-crafted features. MTT outperfroms Denovo in all benchmarked datasets while MTT supersede Generalized in six out of the seven datasets. The performance gains are statistically significant with a p-value of 0.05.

Comparison with sequence embedding based methods

Table 7 provides a comparison between MTT and embedding-based methods on small testing datasets. MTT outperforms Doc2vec in 4 datasets. The performance gains are statistically significant with a p-value of 0.05. MTT is outperformed by Doc2vec in three datasets: Zhou's Ebola, Zhou's H1N1 and Denovo_slim. We point out that these datasets are quite specialized where the negative training and testing samples were drawn from a sequence dissimilarity negative sampling technique. In particular, the protein sequences for negative test set were already chosen based on their dissimilarity to those is positive set. This is not a realistic setting when the positive test set is in itself unknown. Nevertheless, the performance of MTT is comparable to the state of the art on these especially curated datasets too while it outperforms all methods on more general datasets.

Detailed results on ablation studies

In Table 8 we compare MTT with its simpler variants for 11 datasets. The reported results are average after 10 runs. Results from pair-wise t-test show that that (i)MTT

Table 6 Comparison with methods based on hand-crafted features

DATASET	MODEL	AUC	AP	PRECISION	RECALL	F1
ZHOU'S H1N1	DENOVO	0.8656	0.8619	77.75	77.95	77.85
	GENERALIZED	0.8600	0.8606	76.96	77.17	77.06
	HYBRID	0.937	–	–	–	–
	MTT	**0.9461**	**0.9589**	**86.28**	**86.51**	**86.40**
ZHOU'S EBOLA	DENOVO	0.8864	0.8366	83.44	84.00	83.72
	GENERALIZED	0.9154	0.9078	84.77	85.33	85.05
	MTT	**0.9680**	**0.9766**	**90.93**	**91.53**	**91.23**
DENOVO_SLIM	DENOVO	0.8701	0.8631	81.92	82.12	82.02
	GENERALIZED	0.8891	0.8851	**84.74**	**84.94**	**84.84**
	MTT	**0.9221**	**0.9324**	83.92	84.12	84.02
BARMAN	DENOVO	0.8217	0.8415	74.60	74.98	74.79
	GENERALIZED	0.8214	0.8458	74.90	75.27	75.08
	MTT	**0.9804**	**0.9802**	**93.53**	**94.05**	**93.79**
BACILLUS	DENOVO	0.9843	0.9650	94.80	94.83	94.83
	GENERALIZED	0.9833	0.9668	95.75	95.78	95.76
	MTT	**0.9997**	**0.9992**	**98.75**	**98.78**	**98.76**
YERSINA	DENOVO	0.9712	0.9302	93.14	93.16	93.15
	GENERALIZED	0.9758	0.9362	94.01	94.03	94.02
	MTT	**0.9988**	**0.9971**	**97.32**	**97.34**	**97.32**
FRANCI	DENOVO	0.9782	0.9584	95.55	95.62	95.58
	GENERALIZED	0.9799	0.9565	95.84	95.91	95.88
	MTT	**0.9998**	**0.9996**	**98.95**	**99.03**	**98.99**

The bold font is used to highlight highest scores corresponding to each dataset

Table 7 Comparison with embedding-based methods

DATASET	MODEL	AUC	AP	PRECISION	RECALL	F1
ZHOU'S H1N1	DOC2VEC	**0.9601**	**0.9674**	**89.04**	**89.34**	**89.19**
	MOTIFTRANSFORMER	0.945	–	–	–	86.50
	MTT	0.9461	0.9589	86.28	86.51	86.40
ZHOU'S EBOLA	DOC2VEC	**0.9781**	**0.9832**	**91.99**	**92.67**	**92.33**
	MOTIFTRANSFORMER	0.968	–	–	–	89.6
	MTT	0.9680	0.9766	90.93	91.53	91.23
DENOVO_SLIM	DOC2VEC	**0.9644**	**0.9681**	**88.60**	**88.87**	**88.73**
	MTT	0.9221	0.9324	83.92	84.12	84.02
BARMAN	DOC2VEC	0.8671	0.8922	79.95	80.37	80.16
	MTT	**0.9804**	**0.9802**	**93.53**	**94.05**	**93.79**
BACILLUS	DOC2VEC	0.9900	0.9739	96.29	96.32	96.31
	MTT	**0.9997**	**0.9992**	**98.75**	**98.78**	**98.76**
YERSINA	DOC2VEC	0.9814	0.9510	94.50	94.52	94.51
	MTT	**0.9988**	**0.9971**	**97.32**	**97.34**	**97.32**
FRANCI	DOC2VEC	0.9878	0.9606	96.77	96.84	96.81
	MTT	**0.9998**	**0.9996**	**98.95**	**99.03**	**98.99**

The bold font is used to highlight highest scores corresponding to each dataset

Table 8 Results for ablation studies

Dataset	Model	AUC	AP	Precision	Recall	F1
H1N1	Naive baseline	0.8310	0.8003	75.92	76.12	76.02
	STT	0.9472	0.9590	85.86	86.09	85.98
	MTT	**0.9461**	**0.9589**	**86.28**	**86.51**	**86.40**
Ebola	Naive baseline	0.8876	0.8665	82.12	82.67	82.39
	STT	0.9655	0.9749	90.13	90.73	90.43
	MTT	**0.9680**	**0.9766**	90.93	91.53	91.23
Denovo_slim	Naive baseline	0.8843	0.8673	83.80	84.00	83.90
	STT	0.9207	0.9343	84.04	84.24	84.14
	MTT	**0.9221**	**0.9324**	**83.92**	**84.12**	**84.02**
Barman's	Naive baseline	0.8084	0.8198	73.75	74.11	73.93
	MTT	0.9804	0.9802	93.53	94.05	93.79
	STT	0.9801	0.9802	93.83	94.29	94.06
Bacillus	Naive baseline	0.9842	0.9619	93.75	93.78	93.77
	STT	0.9995	0.9986	97.93	97.96	97.95
	MTT	**0.9997**	**0.9992**	**98.75**	**98.78**	**98.76**
Yersina	Naive baseline	0.9741	0.9277	92.61	92.64	92.63
	STT	**0.9987**	0.9970	97.18	97.30	97.24
	MTT	0.9988	**0.9971**	97.32	97.34	97.32
Franci	Naive baseline	0.9851	0.9680	94.36	94.43	94.39
	STT	0.9997	0.9993	98.84	98.92	98.88
	MTT	**0.9998**	**0.9996**	98.95	**99.03**	98.99
2697049	Naive baseline	0.5686	0.0010	0	0	0
	STT	0.7457	0.0017	0.07	0.07	0.07
	MTT	**0.7566**	**0.0021**	**0.97**	**0.97**	**0.97**
333761	Naive baseline	0.7002	0.0110	3.55	3.56	3.55
	STT	0.8114	0.0213	4.72	4.72	4.72
	MTT	0.8160	0.0262	**6.35**	6.35	**6.35**
2043570	Naive baseline	0.6624	0.0076	0.32	0.32	0.32
	STT	0.6706	0.0087	1.11	**3.01**	1.46
	MTT	**0.6956**	**0.0096**	**1.89**	1.91	**1.90**
644788	Naive baseline	0.8410	0.0089	1.82	1.85	1.83
	STT	**0.9705**	**0.0459**	3.97	9.26	4.65
	MTT	0.9537	0.0302	3.54	**22.04**	5.46

The bold font is used to highlight highest scores corresponding to each dataset

is significantly better than Naive baseline in all datasets with p-value smaller than 0.05, (ii) MTT is significantly better than STT in 4 out of 7 datasets with p-value smaller than 0.05 while for the remaining datasets, the difference is not statistically significant.

Abbreviations
ACE2: Angiotensin-converting enzyme 2; ANPEP: Aminopeptidase N; AP: The area under the precision-recall curve; AUC: The area under Receiver Operating Characteristic curve; DDP4: Dipeptidyl peptidase 4; HH: Human–human protein–protein interaction training set; LOSO: Leave-One-Species-Out; LSTM: Long Short Term Memory; mLSTM: Multiplicative Long Short Term Memory; MLP: Multilayer Perceptrons; MTT: Multitask Transfer; PPI: Protein–protein interaction; SLiM: Short Linear Motif; STT: Single task transfer; VH: Virus-human protein–protein interaction training set; Y2H: Yeast-two hybrid; 644788: The Influenza A virus taxon ID; 33761: The HPV 18 virus taxon ID; 697049: The SARS-CoV-2 virus taxon ID; 043570: The Zika virus taxon ID.

Dong *et al. BMC Bioinformatics*   (2021) 22:572

Page 22 of 24

Declarations

Author details
[1]L3S Research Center, Leibniz University Hannover, Hannover, Germany. [2]Institute for Biochemistry, University of Veterinary Medicine, Hannover, Germany. [3]Institute of Experimental Virology, TWINCORE, Center for Experimental and Clinical Infection Research Hannover, Hannover, Germany. [4]Department of Clinical Microbiology, Umeå University, Umeå, Sweden. [5]Wallenberg Centre for Molecular Medicine (WCMM), Umeå University, Umeå, Sweden.

References

1. Petersen E, Koopmans M, Go U, Hamer HH, Petrosillo N, Castelli F, Storgaard M, Al Khalili S, Simonsen L. Comparing SARS-COV-2 with SARS-COV and influenza pandemics. Lancet Infect Dis. 2020;20(9):238–2244.
2. Smith GA, Enquist LW. Break ins and break outs: viral interactions with the cytoskeleton of mammalian cells. Annu Rev Cell Dev Biol. 2002;18:135–61.
3. Beltran PMJ, Cook KC, Cristea IM. Exploring and exploiting proteome organization during viral infection. J Virol. 2017;91(18):00268–17.
4. Gerold G, Bruening J, Weigel B, Pietschmann T. Protein interactions during the flavivirus and hepacivirus life cycle. Mol Cell Proteomics. 2017;16(4 suppl 1):75–91.
5. Sadegh S, Matschinske J, Blumenthal DB, Galindez G, Kacprowski T, List M, Nasirigerdeh R, Oubounyt M, Pichlmair A, Rose TD, et al. Exploring the SARS-COV-2 virus-host-drug interactome for drug repurposing. Nat Commun. 2020;11(1):1–9.
6. Wendt F, Milani ES, Wollscheid B. Elucidation of host-virus surfaceome interactions using spatial proteotyping. Adv Virus Res. 2021;109:105–34.
7. Zapatero-Belinchón FJ, Carriquí-Madroñal B, Gerold G. Proximity labeling approaches to study protein complexes during virus infection. Adv Virus Res. 2021;109:63–104.
8. Lasswitz L, Chandra N, Arnberg N, Gerold G. Glycomics and proteomics approaches to investigate early adenovirus-host cell interactions. J Mol Biol. 2018;430(13):1863–82.
9. Gerold G, Bruening J, Pietschmann T. Decoding protein networks during virus entry by quantitative proteomics. Virus Res. 2016;218:25–39.
10. Lum KK, Cristea IM. Proteomic approaches to uncovering virus-host protein interactions during the progression of viral infection. Expert Rev Proteomics. 2016;13(3):325–40.
11. Greco TM, Cristea IM. Proteomics tracing the footsteps of infectious disease. Mol Cell Proteomics. 2017;16(4):5–14.
12. Jean Beltran PM, Cook KC, Cristea IM. Exploring and exploiting proteome organization during viral infection. J Virol. 2017;91(18):00268–17.
13. Bailer S, Haas J. Connecting viral with cellular interactomes. Curr Opin Microbiol. 2009;12(4):453–9.
14. Spiropoulou CF, Kunz S, Rollin PE, Campbell KP, Oldstone MB. New world arenavirus clade c, but not clade a and b viruses, utilizes α-dystroglycan as its major receptor. J Virol. 2002;76(10):5140–6.
15. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, et al. The intact molecular interaction database in 2012. Nucleic Acids Res. 2012;40(D1):841–6.

16. Calderone A, Licata L, Cesareni G. Virusmentha: a new resource for virus-host protein interactions. Nucleic Acids Res. 2015;43(D1):588–92.
17. Chatr-Aryamontri A, Ceol A, Peluso D, Nardozza A, Panni S, Sacco F, Tinti M, Smolyar A, Castagnoli L, Vidal M, et al. Virusmint: a viral protein interaction database. Nucleic Acids Res. 2009;37(suppl-1):669–73.
18. Ammari MG, Gresham CR, McCarthy FM, Nanduri B. Hpidb 20: a curated database for host-pathogen interactions. Database. 2016;1:9.
19. Requião RD, Carneiro RL, Moreira MH, Ribeiro-Alves M, Rossetto S, Palhano FL, Domitrovic T. Viruses with different genome types adopt a similar strategy to pack nucleic acids based on positively charged protein domains. Sci Rep. 2020;10(1):1–12.
20. Rodrigo G, Daròs J-A, Elena SF. Virus-host interactome: putting the accent on how it changes. J Proteomics. 2017;156:1–4.
21. Gitlin L, Hagai T, LaBarbera A, Solovey M, Andino R. Rapid evolution of virus sequences in intrinsically disordered protein regions. PLoS Pathog. 2014;10(12):1004529.
22. Eid F-E, ElHefnawi M, Heath LS. Denovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics. 2016;32(8):1144–50.
23. Li Y, Ilie L. Predicting protein–protein interactions using sprint. In: Protein–protein interaction networks. Springer; 2020. p. 1–11.
24. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinform. 2017;18(1):1–8.
25. Li Y. Computational methods for predicting protein–protein interactions and binding sites. 2020.
26. Chen K-H, Wang T-F, Hu Y-J. Protein–protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. BMC Bioinform. 2019;20(1):1–17.
27. Sarkar D, Saha S. Machine-learning techniques for the prediction of protein–protein interactions. J Biosci. 2019;44(4):1–12.
28. Sudhakar P, Machiel, K, Vermeire S. Computational biology and machine learning approaches to study mechanistic microbiomehost interactions. 2020.
29. Mei S, Zhang K. In silico unravelling pathogen-host signaling cross-talks via pathogen mimicry and human protein–protein interaction networks. Comput Struct Biotechnol J. 2020;18:100–13.
30. Dick K, Samanfar B, Barnes B, Cober ER, Mimee B, Molnar SJ, Biggar KK, Golshani A, Dehne F, Green JR, et al. Pipe4: fast ppi predictor for comprehensive inter-and cross-species interactomes. Sci Rep. 2020;10(1):1–15.
31. Li BYS, Yeung LF, Yang G. Pathogen host interaction prediction via matrix factorization. In: 2014 IEEE international conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2014. p. 357–62.
32. Guven-Maiorov E, Tsai C-J, Ma B, Nussinov R. Interface-based structural prediction of novel host-pathogen interactions. In: Computational methods in protein evolution. Springer; 2019. p. 317–35.
33. Basit AH, Abbasi WA, Asif A, Gull S, Minhas FUAA. Training host-pathogen protein–protein interaction predictors. J Bioinform Comput Biol. 2018;16(04):1850014.
34. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16(12):1315–22.
35. Nouretdinov I, Gammerman A, Qi Y, Klein-Seetharaman J. Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method. In: Biocomputing. World Scientific; 2012. p. 311–22.
36. Nourani E, Khunjush F, Durmuş S. Computational prediction of virus-human protein–protein interactions using embedding kernelized heterogeneous data. Mol BioSyst. 2016;12(6):1976–86.
37. Mei S, Zhu H. A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. Sci Rep. 2015;5(1):1–13.
38. Cui G, Fang C, Han K. Prediction of protein–protein interactions between viruses and human by an SVM model. BMC Bioinform. 2012;13:1–10.
39. Kim B, Alguwaizani S, Zhou X, Huang D-S, Park B, Han K. An improved method for predicting interactions between virus and human proteins. J Bioinform Comput Biol. 2017;15(01):1650024.
40. Loaiza CD, Kaundal R. Predhpi: an integrated web server platform for the detection and visualization of host-pathogen interactions using sequence-based methods. Bioinformatics. 2020;37:622–4.
41. Zhou X, Park B, Choi D, Han K. A generalized approach to predicting protein–protein interactions between virus and host. BMC Genomics. 2018;19(6):69–77.
42. Ma Y, He T, Tan Y-T, et al. Seq-bel: sequence-based ensemble learning for predicting virus-human protein–protein interaction. IEEE/ACM Trans Comput Biol Bioinform. 2020;1:1.
43. Deng L, Zhao J, Zhang J. Predict the protein–protein interaction between virus and host through hybrid deep neural network. In: 2020 IEEE international conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2020. p. 11–16.
44. Dey L, Chakraborty S, Mukhopadhyay A. Machine learning techniques for sequence-based prediction of viral-host interactions between SARS-COV-2 and human proteins. Biomed J. 2020;43(5):438–50.
45. Yang X, Yang S, Li Q, Wuchty S, Zhang Z. Prediction of human-virus protein–protein interactions through a sequence embedding-based machine learning method. Comput Struct Biotechnol J. 2020;18:153–61.
46. Lanchantin J, Weingarten T, Sekhon A, Miller C, Qi Y. Transfer learning for predicting virus-host protein interactions for novel virus sequences. bioRxiv. 2021;2020-12.
47. Liu-Wei W, Kafkas S, Chen J, Dimonaco NJ, Tegnér J, Hoehndorf R. Deepviral: prediction of novel virus-host interactions from protein sequences and infectious disease phenotypes. Bioinformatics. 2021. https://doi.org/10.1093/bioinformatics/btab147.
48. Barman RK, Saha S, Das S. Prediction of interactions between viral and host proteins using supervised machine learning methods. PLoS ONE. 2014;9(11):112034.
49. Lasso G, Mayer SV, Winkelmann ER, Chu T, Elliot O, Patino-Galindo JA, Park K, Rabadan R, Honig B, Shapira SD. A structure-informed atlas of human-virus interactions. Cell. 2019;178(6):1526–41.
50. Liu D, Ma Y, Jiang X, He T. Predicting virus-host association by kernelized logistic matrix factorization and similarity network fusion. BMC Bioinform. 2019;20(16):1–10.

51. Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman JA, Braun J, Sun F, Ahlgren NA. A network-based integrated framework for predicting virus-prokaryote interactions. NAR Genomics Bioinform. 2020;2(2):044.
52. Biswas S. Principles of machine learning-guided protein engineering. PhD thesis; 2020.
53. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. String v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(D1):447–52.
54. Alonso-Lopez D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R, De Las Rivas J. Apid interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Res. 2016;44(W1):529–35.
55. Consortium U. Uniprot: a hub for protein information. Nucleic Acids Res. 2015;43(D1):204–12.
56. Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E, Dana JM, De Las Rivas J, Dumousseau M, Galeota E, et al. Psicquic and psiscore: accessing and scoring molecular interactions. Nat Methods. 2011;8(7):528–9.
57. Martin S, Roe D, Faulon J-L. Predicting protein–protein interactions using signature products. Bioinformatics. 2005;21(2):218–26.
58. Mei S. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. PLoS ONE. 2013;8(11):79606.
59. Federhen S. The NCBI taxonomy database. Nucleic Acids Res. 2012;40(D1):136–43.
60. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Travé G, Gibson TJ. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci. 2008;13(6580):603.
61. Neduva V, Russell RB. Peptides mediating interaction networks: new leads at last. Curr Opin Biotechnol. 2006;17(5):465–71.
62. Le Q, Mikolov T. Distributed representations of sentences and documents. In: International conference on machine learning. PMLR; 2014. p. 1188–96.
63. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32:8026–37.
64. Welch BL. The generalization of 'student's' problem when several different population variances are involved. Biometrika. 1947;34(1–2):28–35.
65. Salzberg SL. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Discov. 1997;1(3):317–28.
66. Kafadar K. Handbook of parametric and nonparametric statistical procedures. Am Stat. 1997;51(4):374.
67. Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, Moest H, Omasits U, Gundry RL, Yoon C, et al. A mass spectrometric-derived cell surface protein atlas. PLoS ONE. 2015;10(4):0121314.
68. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
69. Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, Hartline E, et al. The gene ontology resource: enriching a gold mine. Nucleic Acids Res. 2021;49(D1):325–34.
70. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F. Cell entry mechanisms of SARS-COV-2. Proc Natl Acad Sci. 2020;117(21):11727–34.
71. Zhang Q, Xiang R, Huo S, Zhou Y, Jiang S, Wang Q, Yu F. Molecular mechanism of interaction between SARS-COV-2 and host cells and interventional therapy. Signal Transduct Target Ther. 2021;6(1):1–19.
72. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu N-H, Nitsche A, et al. SARS-COV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell. 2020;181(2):271–80.
73. Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, Somasundaran M, Sullivan JL, Luzuriaga K, Greenough TC, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature. 2003;426(6965):450–4.
74. Bao L, Deng W, Huang B, Gao H, Liu J, Ren L, Wei Q, Yu P, Xu Y, Qi F, et al. The pathogenicity of SARS-COV-2 in HACE2 transgenic mice. Nature. 2020;583(7818):830–3.
75. Winkler ES, Bailey AL, Kafai NM, Nair S, McCune BT, Yu J, Fox JM, Chen RE, Earnest JT, Keeler SP, et al. SARS-COV-2 infection of human ACE2-transgenic mice causes severe lung inflammation and impaired function. Nat Immunol. 2020;21(11):1327–35.
76. Wang N, Shi X, Jiang L, Zhang S, Wang D, Tong P, Guo D, Fu L, Cui Y, Liu X, et al. Structure of MERS-COV spike receptor-binding domain complexed with human receptor DPP4. Cell Res. 2013;23(8):986–93.
77. Vankadari N, Wilce JA. Emerging covid-19 coronavirus: glycan shield and structure prediction of spike glycoprotein and its interaction with human cd26. Emerg Microbes Infect. 2020;9(1):601–4.
78. Yeager CL, Ashmun RA, Williams RK, Cardellichio CB, Shapiro LH, Look AT, Holmes KV. Human aminopeptidase n is a receptor for human coronavirus 229e. Nature. 1992;357(6377):420–2.
79. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with alphafold. Nature. 2021;596:583–9.
80. Dong NT, Khosla M. A multitask transfer learning framework for novel virus-human protein interactions. bioRxiv. 2021. https://doi.org/10.1101/2021.03.25.437037.

Publisher's Note