Prediction of Novel Virus–Host Protein Protein Interactions From Sequences and Infectious Disease Phenotypes

Thesis by

Liuwei Wang

In Partial Fulfillment of the Requirements

For the Degree of

Masters of Science

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

November, 2020

EXAMINATION COMMITTEE PAGE

The thesis of Liuwei Wang is approved by the examination committee

Committee Chairperson: Prof. Jesper Tegner
Committee Co-Chair: Prof. Robert Hoehndorf
Committee Members: Prof. Hernando Ombao

ABSTRACT

Prediction of Novel Virus–Host Protein Protein Interactions From Sequences and Infectious Disease Phenotypes

Liuwei Wang

Infectious diseases from novel viruses have become a major public health concern. Rapid identification of virus–host interactions can reveal mechanistic insights into infectious diseases and shed light on potential treatments. Current computational prediction methods for novel viruses are based mainly on protein sequences. However, it is not clear to what extent other important features, such as the symptoms caused by the viruses, could contribute to a predictor. Disease phenotypes (i.e., signs and symptoms) are readily accessible from clinical diagnosis and we hypothesize that they may act as a potential proxy and an additional source of information for the underlying molecular interactions between the pathogens and hosts.

We developed DeepViral, a deep learning based method that predicts protein–protein interactions (PPI) between humans and viruses. Motivated by the potential utility of infectious disease phenotypes, we first embedded human proteins and viruses in a shared space using their associated phenotypes and functions, supported by formalized background knowledge from biomedical ontologies. By jointly learning from protein sequences and phenotype features, DeepViral significantly improves over existing sequence-based methods for intra- and inter-species PPI prediction. Lastly, we propose a novel experimental setup to realistically evaluate prediction methods for novel viruses.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Prof. Jesper Tegner, whose guidance and support for this project and my transition to the STAT program helped me grow as a more well-rounded student and a more independent researcher.

I also owe special thanks to Prof. Robert Hoehndorf, who supported me ever since I came to KAUST as a visiting student. He also conceived using phenotypes to improve prediction for infectious diseases, which is the core idea of DeepViral.

Sincere thanks also go to Prof. Hernando Ombao for agreeing to serve on my thesis committee and taking the time to evaluate my thesis.

My appreciation also goes to the members of Living Systems Lab and BORG. I benefited greatly from our interactions and discussions.

I would also like to thank my parents, for their love and support along the way.

Lastly, a special shout out to the essential service providers and the infectious disease researchers during the COVID-19 pandemic, whose sacrifice and dedication have been an inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Infectious diseases emerging unexpectedly from novel and reemerging pathogens have been a major enduring public health concern around the globe [2]. Pathogens disrupt host cell functions [3] and target immune pathways [4] through complex inter-species interactions of proteins [5], RNA [6] and DNA [7]. The study of pathogen–host interactions (PHI) can therefore provide insights into the molecular mechanisms underlying infectious diseases and guide the discoveries of novel therapeutics or provide a basis for the repurposing of available drugs. For example, a previous study of many PHIs showed that pathogens typically interact with the protein hubs (those with many interaction partners) and bottlenecks (those of central locations to important pathways) in human protein–protein interaction (PPI) networks [5]. However, due to cost and time constraints, experimentally validated pairs of interacting pathogen–host proteins are limited in number. Therefore, the computational prediction of PHIs is a useful complementary approach in suggesting candidate interaction partners from the human proteome.

Existing PHI prediction methods for novel viruses typically utilize protein sequence features of the interacting proteins [1, 8, 9, 10]. While protein functions have been shown to predict intra-species (e.g., human) PPIs [11, 12, 13] and such protein specific features exist for some extensively studied pathogens, such as *Mycobacterium tuberculosis* [14] and HIV [15], for most pathogens, these features are rare and expensive to obtain. As new virus species continue to be discovered [16], a method is needed to rapidly identify candidate interactions from information that can be ob-

tained quickly, such as the signs and symptoms of the host, which may be utilized as a proxy for the underlying molecular interactions between host and pathogen proteins.

The phenotypes elicited by pathogens, i.e., the signs and symptoms observed in a patient, may provide information about molecular mechanisms [17]. The information that phenotypes provide about molecular mechanisms is commonly exploited in computational studies of Mendelian disease mechanisms [18, 19], for example to suggest candidate genes [20, 21] or diagnose patients [22], but the information can also be used to identify drug targets [23] or gene functions [24]. We hypothesize that the host phenotypes elicited by an infection with a pathogen are, among others, the result of molecular interactions, and that knowledge of the phenotypes exhibited by the host can be used to suggest the protein perturbations from which these phenotypes arise.

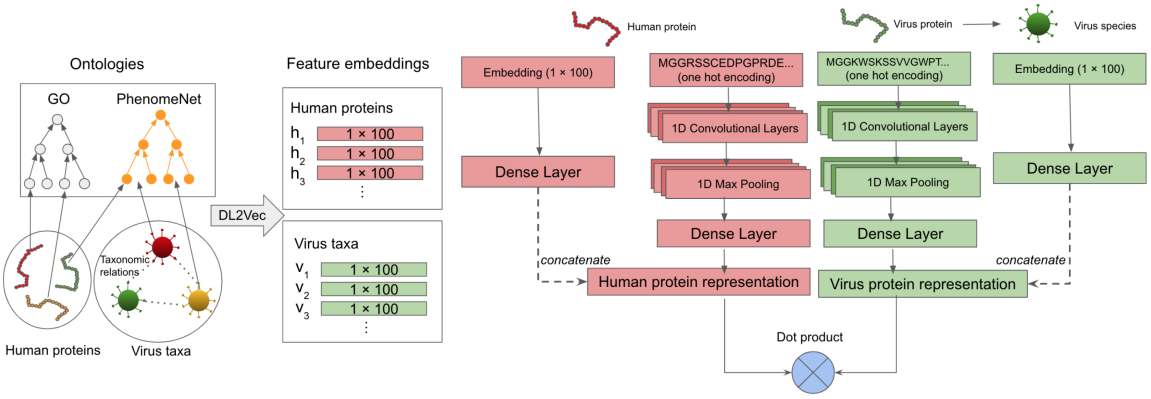
One major challenge of the novel PHI prediction problem is the lack of ground truth negative data. A recent method, DeNovo [1], adopted a "dissimilarity-based negative sampling": for each virus protein, the negatives are sampled from human proteins that do not have known interactions with other similar virus proteins (above a certain sequence similarity threshold). Another method based on protein sequences [8], samples negatives from only the set of host proteins that are less than 80% similar (in terms of sequence similarity) to the host proteins in the positive training data. However, the influence of sequence similarity on function is not uniform and while there is evidence for a number of general evolutionary rules, we are unable to determine cutoffs for any specific protein or function [25, 26]. By construction, these sampling schemes make the human proteins in the negative set different from the positive set; when used not only for training a model but also for evaluating its performance, this sampling scheme has the potential to over-estimate the actual performance for finding novel PHIs. In a more realistic evaluation for a novel virus species, a model would be evaluated on all the host proteins with which it could potentially interact, regardless of sequence similarity.

From these motivations, we developed a machine learning method, DeepViral, to predict potential interactions between viruses and all human proteins for which we can generate the relevant features. Firstly, the features of phenotypes, functions and taxonomic classifications are embedded in a shared space for human proteins and viruses. We then extended a sequence model by incorporating the phenotype features of viruses into the model. We show that the joint model trained on both the sequences and phenotypes can significantly outperform state-of-the-art methods and predict potential PHIs in a realistic experimental setup for novel viruses.

Chapter 2

Methods

DeepViral is a model that predicts potential protein interactions between viruses and human hosts from the protein sequences and feature embeddings of phenotypes, functions and taxonomies. To enable predictions based on such different features we embedded them in a shared representation space. We then combine these feature embeddings with a protein sequence model to predict potential PHIs of novel viruses. The workflow of DeepViral is illustrated in Figure 2.1.



(a) Generation of feature embeddings (b) Joint prediction model from embeddings and sequences

Figure 2.1: The workflow of DeepViral. (a) Generation of an embedding: the arrows of human proteins and virus taxa represent their annotations to the ontology classes. The dashed lines between viruses represent their taxonomic relations. The annotations, taxonomic relations and ontologies were fed into DL2Vec to generate feature embeddings of dimension 100 for each human protein and virus taxa. (b) Joint prediction model: latent representation learned from feature embeddings and protein sequences are concatenated into a joint representation, for human protein and virus protein respectively, on which a dot product is performed to predict interactions.

2.1 Data sources

Interactions between hosts and pathogens were obtained from the Host Pathogen Interaction Database (HPIDB; version 3) [27]. The phenotypes associated with pathogens were collected from the PathoPhenoDB [28], a database of manually curated and text-mined associations of pathogens, infectious diseases and phenotypes. We downloaded the PathoPhenoDB database version 1.2.1 (`http://patho.phenomebrowser.net/`).

The phenotypes associated with human genes were collected from the Human Phenotype Ontology (HPO) database [29], and the phenotypes associated with mouse genes and the orthologous gene mappings from mouse genes to human genes originated from the Mouse Genome Informatics (MGI) database [30]. The Entrez gene IDs in HPO and MGI were mapped to reviewed Uniprot protein IDs using the Uniprot Retrieve/ID mapping tool (`https://www.uniprot.org/uploadlists`) on March 6, 2020. The Gene Ontology annotations of human proteins (release date 2020-02-22) were downloaded from the Gene Ontology Consortium [31, 32]. Human PPI networks were downloaded from String [33] and filtered to only include the interactions with experimental evidence. The human protein sequences were obtained from the Swiss-Prot database [34].

To add background knowledge from biomedical ontologies of phenotypes and GO classes, we downloaded the cross-species PhenomeNET ontology [20, 35], from the AberOWL ontology repository [36] on September 13, 2018. We obtained the NCBI Taxonomy classification [37] as an ontology in OWL format (version 2018-07-27) from EMBL-EBI ontology repository (`https://www.ebi.ac.uk/ols/ontologies/ncbitaxon`).

2.2 Learning feature embeddings

To generate feature embeddings, we used DL2Vec [38], a recent method for learning features for entities (in our case, the human proteins and viruses) from their associations to ontological classes. DL2Vec first converted the ontologies and entity associations into a graph, with the classes and entities as the nodes and the associations and ontology axioms as the edges. Then a number of random walks were performed, starting from the entities over to the ontology graph and thereby generating a corpus of walks in the form of sentences capturing the graph neighborhoods and thereby the ontology axioms. After the construction of such sentences, a Word2vec skipgram model [39] was used to learn an embedding for each entity by learning from the corpus. Following the recommendations of the authors of DL2Vec, we fixed the number of walks to 100, the walk length to 30, the embedding dimension to 100 and the number of training epochs to 30. The embeddings were trained with the Word2Vec library in Julia (version 1.0.4). The resultant embedding was a vector representation of an entity capturing its co-occurrence relations with other entities within the walks generated by DL2Vec. As an example, the walks starting from a virus node explored its graph neighborhood, i.e., its associated phenotypes and its taxonomic relatives, and as an result, its feature embedding captured this information according to the co-occurrence patterns.

2.3 Supervised prediction models and parameter tuning

The neural network model of DeepViral consists of two components: a phenotype model based on the feature embeddings of viruses and human proteins and a sequence model based on the amino acid sequences of the human and viral proteins. The maximum input length of protein sequences is set to 1,000 amino acids and all shorter sequences are repeated up to the maximum length. The input protein sequences are

encoded as a one-hot encoding matrix of 22 rows that represents each amino acid type and the original sequence length (before being repeated), and 1,000 columns representing each position of the amino acid sequence.

To predict the likelihood of an interaction between a pair of proteins, we trained the network as a binary classifier, to minimize the binary cross-entropy loss defined below,

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_t \cdot log(y_p) + (1 - y_t) \cdot log(1 - y_p)$$

where N is the total number of predictions, y_t and y_p is the true label and predicted likelihood of y.

We implemented our model using the Keras library [40] and performed training on Nvidia Tesla V100 GPUs. The phenotype model consists of a fully connected layer with the feature embeddings as input. The sequence model is a convolutional neural network (CNN) with the sequences as input and consists of 1-dimensional convolution, max pooling and fully connected layers. We tuned the following hyper-parameters of the model: the sizes and numbers of the convolution filters, the size of the max pool and the number of neurons in the fully connected layers. We fixed these hyperparameters throughout all the experiments: 16 convolutional layers for each filter of 8, 16, ..., 64 in length, a pool size of 200 and 8 neurons for the dense layers. We also used dropouts [41] for the convolutional and dense layers with a rate of 0.5 and LeakyReLU as the activation function for the dense layer with an alpha set to 0.1.

Chapter 3

Results

3.1 Embedding features of viruses and human proteins from phenotypes, functions and taxonomies

We started with the biological hypothesis that phenotypes (i.e., symptoms) elicited by viruses in their hosts can act as a proxy for the underlying molecular mechanisms of the infection, and therefore may provide additional information to the prediction of potential PHIs for novel viruses.

To generate feature embeddings for human proteins and virus taxa, we applied a recent representation learning method DL2Vec [38], which learned feature embeddings for entities based on their annotations to ontology classes (see Section 2.2). DL2Vec takes two types of inputs: the associations of the entities with ontology classes (e.g., human proteins and their functions), and the ontologies themselves.

For representing virus taxa through the phenotypes they elicit in their hosts, we used the phenotype associations for viruses from PathoPhenoDB [28], a database of pathogen to host phenotype (signs and symptoms) associations. To increase the coverage of phenotypes beyond PathoPhenoDB, the taxonomic relations of the viruses were added from the NCBI Taxonomy [37]. By adding these taxonomic relations (as annotations of viruses to DL2Vec), we propagated the known phenotypes along the taxonomic hierarchies and learned a generalized embedding for viruses that do not have any phenotype annotations in PathoPhenoDB but have close relatives that do.

Similarly, for representing human proteins, we used the annotations of their as-

sociated phenotypes from the Human Phenotype Ontology (HPO) database [29], the phenotypes associated with their mouse orthologs from the Mouse Genome Informatics (MGI) database [30], and their protein functions from the Gene Ontology (GO) database [31, 32]. We propagated these annotations through the human PPI network, which has been shown to improve prediction for gene-disease associations [42].

To provide DL2Vec with structured background knowledge of human and mouse phenotypes as well as protein functions, we used the cross-species phenotype ontology PhenomeNET [20, 35], which is built upon and includes the Gene Ontology [31, 32]. These ontologies contain formalized biological background knowledge [43], which has the potential to significantly improve the performance of these features in machine learning and predictive analyses [44, 45].

3.2 A joint model for PPI prediction from sequences and phenotypes

DeepViral consists of a phenotype model trained on phenotypes caused by a viral infection and a sequence model trained on protein sequences, as shown in Figure 2.1 (b). The two models take a pair of virus and human proteins as input and predicts the probability score of their interaction. The inputs for a human protein are its feature embedding and its sequence, and the features for a viral protein are its sequence and the feature embedding of the virus species to which it belongs. The sequence model projects the protein sequence into a low dimension vector representation, which is concatenated with the vector projected from the embedding by the phenotype model to form a joint representation of the proteins. A dot product was performed over the two vector representations of the pair of proteins to compute their similarity, which was then used as input to a sigmoid activation function to compute their predicted probability of interaction. In an evaluation where the inputs were not symmetric, e.g., only using the feature embeddings of human proteins but not viruses (or vice

versa), an additional dense layer was added to project the longer representation to the same dimension as the other so that the dot product could be performed.

Existing prediction methods for inter-species PPI (e.g., virus–human interactions) have rarely been compared with methods designed for intra-species (e.g. human) PPI prediction. To compare with the existing sequence-based methods for both intra- and inter-species PPI prediction, we evaluated DeepViral and RCNN [46], a recent method designed for intra-species prediction, on an existing dataset [1] that has been used to evaluate a number of PHI prediction methods [10, 9, 8]. The respective model performances and implementation details are shown in Table 3.1. DeepViral trained only on sequences achieves comparable performance with other sequence based methods, while the joint model is able to achieve the best performances in most metrics. However, the evaluation dataset suffers from several drawbacks: 1) negative sampling (to create a balanced dataset) was based on sequence dissimilarity; 2) the training and test sets are small relative to the current size of the PHI databases; 3) there are overlapping viruses (i.e., data leakage) at species level between the training and test sets, which makes it unsuitable for the problem of novel PHI prediction.

Method	SN (%)	SP (%)	ACC (%)	PPV (%)	NPV (%)	MCC	AUC	F1 (%)
DeNovo [1]	80.71	83.06	81.90	NA	NA	NA	NA	NA
VirusHostPPI [8]	80.00	88.94	84.47	87.86	81.64	0.692	0.897	NA
Doc2Vec + RF [10]	**90.33**	96.17	93.23	95.99	90.74	0.866	**0.981**	93.07
RCNN [46]	89.88	95.58	92.73	95.38	90.46	0.857	0.974	92.52
DeepViral (seq)	89.36	96.89	93.13	96.68	90.13	0.865	0.960	92.86
DeepViral (seq + human embedding)	88.43	96.22	92.32	95.94	89.23	0.849	0.955	92.02
DeepViral (seq + viral embedding)	88.29	97.24	92.76	96.97	89.26	0.859	0.967	92.42
DeepViral (joint)	90.27	**97.58**	**93.91**	**97.43**	**90.93**	**0.881**	0.976	**93.68**

Table 3.1: Comparison with the state-of-the-art methods on the datasets of [1] (the performances of first 3 methods are from the original papers respectively). RCNN and the variants of DeepViral are evaluated 5 times independently to compute the mean of the metrics: SN - sensitivity, SP - specificity, ACC - accuracy, PPV - positive predictive value (precision), NPV - negative predictive value, MCC - Matthews correlation coefficient, AUC - area under the ROC curve. DeepViral (seq) only utilizes the protein sequences and the joint model also includes both the human and virus embeddings as input. The bold numbers represent the best metric for a dataset.

3.3 Experimental setup, negative sampling and evaluation metrics for novel viruses

Motivated by the need for more representative datasets to evaluate methods for novel PHI prediction, we constructed a larger dataset from the curated virus–host interactions in HPIDB [27], a database of host–pathogen protein–protein interactions. We constructed our positive set by filtering HPIDB to include all virus–host interactions that 1) are provided with an MIscore, a confidence score for molecular interactions [47]; 2) are associated with an existing virus family in the NCBI taxonomy [37]; 3) are within 1,000 amino acids in length (for both human and viral proteins). The sequence length cut-off of 1,000 is chosen to include over 88.2% of the human proteins in Swiss-Prot and over 91.6% of the virus proteins in HPIDB. After filtering, the dataset includes 24,678 positive interactions and 1,066 viral proteins from 14 virus families and 292 virus taxa.

To realistically evaluate the prediction performance, we performed a leave-one-family-out (LOFO) cross validation: at each run, one virus family in our positive set was left out for testing, 20% of the remaining families for validation, and the rest 80% for training. The objective of the LOFO cross-validation is to evaluate the model under a scenario in which the novel virus emerges from a novel virus family - in our study, "novel" is defined as the situation in which we have no or very little knowledge about its protein interactions and the molecular functions of the viral proteins.

Instead of using "dissimilarity-based negative sampling" to construct a balanced dataset, we sampled our negatives from all the possible pairwise combinations of human and viral proteins, as long as the pair did not occur in the positive set. Essentially, we treated all "unknown" interactions as negatives. As the dataset was at this point unbalanced with more negatives than positives, we evaluated the model with the area under the receiver operating characteristic (ROC) curve [48]. A high ROCAUC indicates the ability of the model to prioritize the true positive interacting

proteins out of all the human proteins. We computed a ROCAUC for each virus family, and also for each viral protein and virus taxon in that family, for which we reported the mean across them, i.e. macro averages. Each model was evaluated 5 times independently to compute the 95% confidence interval of the ROCAUC, which is bounded by $mean \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$, where n is the sample size and σ is the standard deviation. Additionally, the mean ranks of the true positive proteins were provided as a more interpretable metric: for each viral protein, we ranked all of the 16,627 human proteins in Swiss-Prot (with a length limit of 1,000) as its potential interaction partner based on the prediction score and obtained the mean ranks of the true positives.

3.4 Phenotypes improve prediction for novel viruses

With the newly constructed dataset, we further evaluated the existing methods as well as the variants of DeepViral, under the scenario in which a novel virus (from a novel family) emerges and no previous knowledge (except about its protein sequences and the phenotypes elicited in its hosts) is known.

We compared DeepViral with two existing state-of-the-art methods based on protein sequences: Doc2Vec + RF [10], a recent method predicting for virus–human interactions; and RCNN [46], a recent deep learning based method for intra-species (e.g., human) PPI prediction. To adapt Doc2Vec + RF on our dataset, we used the pretrained Doc2Vec model provided by the authors and the same parameters for the random forest model for training. Similarly, for RCNN, we used the pre-trained embeddings for amino acids and the same model parameters for training. Since the stop criterion for Doc2Vec + RF was to have at most 2 samples at each leaf node, we did not use validation data and trained it with the entirety of the training data, while a validation set was used for both RCNN and DeepViral as described in the experimental setup.

The performance of each model is shown in Table 3.2. For models using only

sequence features, DeepViral and Doc2Vec + RF perform on a similar level across the metrics. As the current state-of-the-art method for intra-species PPI prediction, RCNN consistently yields the lowest performances. Adding human or virus embeddings individually shows a slight improvement in most metrics, compared to the sequence-only models, while the joint model with both embeddings achieved the best performances overall.

Method	Family-wise ROCAUC	Taxon-wise ROCAUC	Protein-wise ROCAUC	Mean rank
RCNN [46]	0.726 [0.717 - 0.734]	0.759 [0.750 - 0.768]	0.737 [0.731 - 0.743]	4669
Doc2Vec + RF [10]	0.764 [0.763 - 0.765]	0.768 [0.766 - 0.770]	0.751 [0.751 - 0.752]	3740
DeepViral (seq)	0.770 [0.763 - 0.777]	0.768 [0.759 - 0.777]	0.749 [0.742 - 0.756]	4064
DeepViral (seq + human embedding)	0.778 [0.766 - 0.790]	0.789 [0.776 - 0.801]	0.757 [0.742 - 0.771]	4245
DeepViral (seq + viral embedding)	0.788 [0.776 - 0.801]	0.782 [0.773 - 0.790]	0.757 [0.746 - 0.767]	3496
DeepViral (joint)	**0.813 [0.808 - 0.817]**	**0.829 [0.822 - 0.836]**	**0.800 [0.797 - 0.804]**	**3156**

Table 3.2: Comparison with the state-of-the-art methods on our dataset to evaluate the performances for novel viruses. The brackets after DeepViral indicate the features used for the model: seq – protein sequences, joint – both sequences and embeddings of human proteins and viruses. The square brackets behind ROCAUC scores indicate the 95% confidence interval.

Chapter 4

Discussion

4.1 Summary

We developed DeepViral, a machine learning method for predicting PHIs between viruses and human hosts. DeepViral is, based on our review of the literature, the first predictor using clinical phenotypes as an additional feature in PHI prediction and it has been seen to provide a significant improvement ($p < 0.05$; see confidence intervals in Table 3.2) over purely sequence based methods. Phenotype-based approaches have been successful in predicting disease-gene associations for Mendelian diseases [20] and intra-species PPIs [49], but have not yet been used for the prediction of (inter-species) PHIs in infectious diseases. Our model avoids the bottleneck of identifying the molecular functions of pathogen proteins by instead introducing a novel and – in the context of infectious diseases – rarely explored type of feature, the phenotypes elicited by pathogens in their hosts, as a "proxy" for the molecular mechanisms, which in turn eventually produce the observed clinical phenotypes.

4.2 Prediction with phenotypes

The focus of our method on utilizing features generated based on endo-phenotypes observed in humans and mice [50] has therefore the crucial advantage that we can identify host-pathogen interactions that may contribute to particular signs and symptoms. For example, our model consistently prioritizes the interaction between the proteins of Zika virus (`NCBITaxon:64320`) and DDX3X (`UniProt:O00571`) in humans. Infec-

tions with Zika virus have the potential to result in abnormal embryogenesis and, specifically, microcephaly [51]. Phenotypes associated with DDX3X in the mouse ortholog include abnormal embryogenesis, microcephaly, and abnormal neural tube closure [52]. DDX3X mutations in humans have been found to result in intellectual disability, specifically in females and affect individuals in a dose-dependent manner [53]. While DDX3X has previously been linked to the infectivity of the Zika virus [54], our model further suggests a role of DDX3X in the development of the embryogenesis phenotypes resulting from Zika virus infections.

4.3 Dataset construction for novel viruses

While we demonstrate quantitatively an improvement over existing methods on an existing dataset [1], we argue that the performances using this evaluation approach may have been over-estimated due to the negative sampling scheme based on sequence similarity that is used not only for training but also for evaluation of the model. Under a more realistic evaluation procedure that considers all host proteins as potential interaction partners for novel viruses, the achieved predictive performances are considerably lower. This calls for future efforts in the direction of PHI prediction of novel viruses, an issue today of increasing relevance to global public health. Accurate predictions of potential PHIs for novel pathogens with rapidly obtainable features would be an important development for understanding infectious disease mechanisms and the repurposing of existing drugs.

4.4 Application on the novel coronavirus

An example of such a novel virus is the novel coronavirus SARS-CoV-2, which as of 6th August 2020 reached more than 18 million infected cases and 707 thousand fatalities globally [55] in a timespan of 9 months. Based on a recently released dataset of 332 PHIs from 26 viral proteins of SARS-CoV-2 [56], we applied DeepViral by

treating it as a novel family (with no other Coronaviridae viruses in the dataset) and achieved a family-wise ROCAUC of 0.723 (0.699–0.747; 95% CI), which is within the observed variability in predicting for different virus families, as shown in Figure 4.1. This family-wise variability suggests that the learned features to predict for PHIs may have different generalization power across families, possibly a result of varying degrees of (dis)similarity between the virus families. Nonetheless, optimizing the predictive power for a single virus, e.g., SARS-CoV-2, requires a case-by-case experimental setup. Specifically in the case of SARS-CoV-2, one can potentially relax the leave-one-family-out evaluation, as we have prior knowledge about other species in its family, e.g., SARS-CoV and MERS-CoV, and their interactions with hosts and protein functions [57]. This is indeed a topic for further investigation.

4.5 Limitations and future work

There are several limitations that can be addressed by future efforts. One is the scarcity of training data for inter-species PPIs and this may be leveraged by transfer learning on the much larger intra-species PPI data available for humans and other model organisms. We also ignored other types of PHIs outside virus–human interactions in our current study, such as those of other hosts, e.g., plants and fishes, and other types of pathogens, e.g., bacteria and fungi. Additionally, predicting tissue-specific PHIs would also provide additional insights, as proteins of both human hosts [58] and viruses [59] often have tissue-specific expressions and functions.

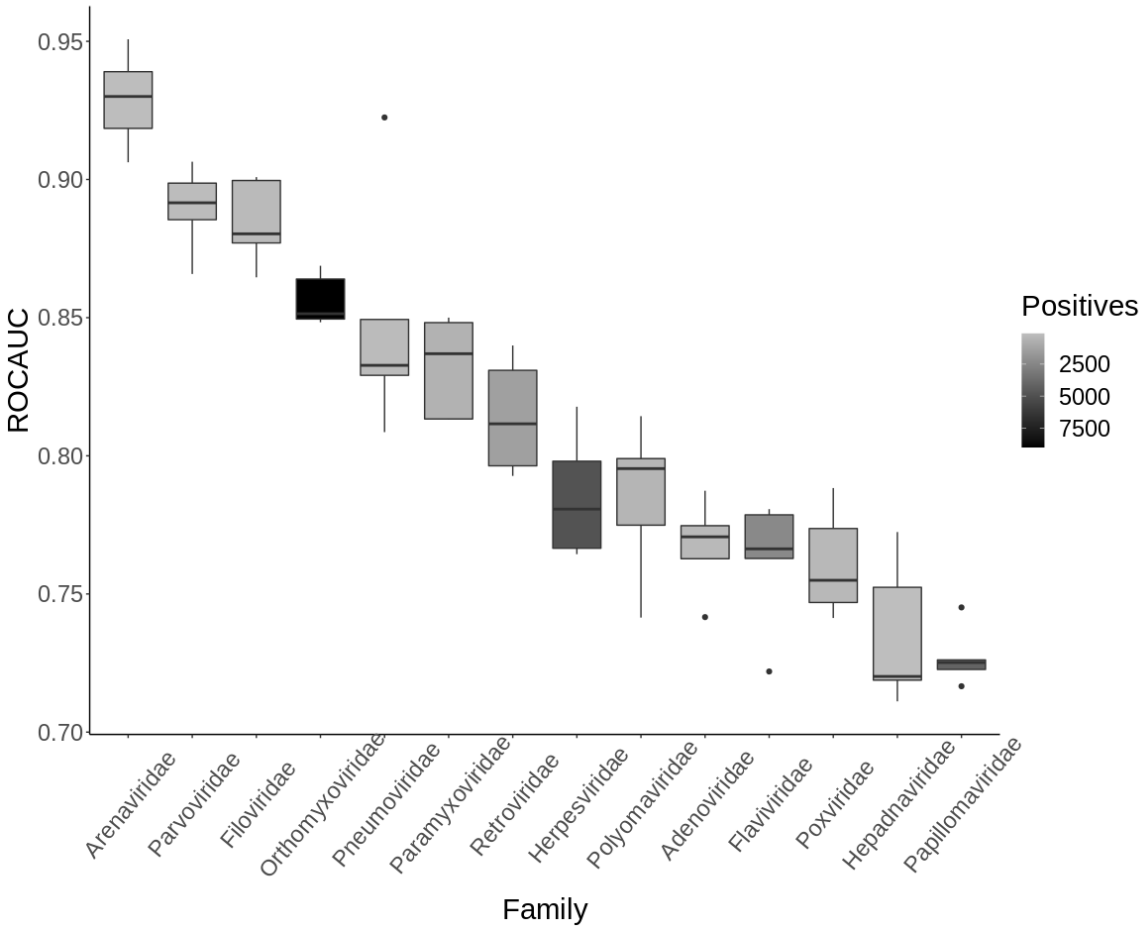


Figure 4.1: ROCAUC for each of the 14 virus families from the joint model, colored by the number of positives belonging to that family.

Chapter 5

Extensions of DeepViral

In this chapter, we briefly describe two extensions of the DeepViral model and the underlying HPI dataset, present the preliminary results, and discuss their implications for future work.

5.1 Transfer Learning from Human PPI data

5.1.1 Motivation

Transfer learning is a technique and a popular research topic in machine learning that focuses on adapting models trained for one task to other similar tasks, e.g., tuning neural networks trained for recognizing faces to hair styles. Previous work [8] has attempted to improve human–virus interaction prediction by adding training data from other host species such as plants, but shown mixed results for different virus species. As there exist much more intra-species protein–protein interaction data for humans and other model organisms, one natural hypothesis is that knowledge learned from predicting intra–species PPIs may be transferred to improve prediction for inter–species PPIs. To this end, we pre-trained DeepViral on human PPI data and performed fine-tuning on the same evaluation dataset we constructed in the previous sections. To the best of our knowledge, this work is the first attempt to leverage intra-species PPI data for inter-species prediction.

5.1.2 Methods

Transfer learning with deep learning models typically consists of two steps: 1) a pre-training step where the model is trained for a more general task; and 2) a fine-tuning step where the model is further tuned on data for the specific task. In our study, the pre-training task is the prediction of human intra-species PPI and the fine-tuning task is the prediction of virus–human interspecies PPI. As a first attempt, we focus on extending the sequenced-based DeepViral model, since the interacting sequences likely carry motifs or protein domains that are shared between both intra- and inter-species PPIs.

To extend DeepViral on human data where two interacting proteins are both from the same species, it is necessary to adjust the original DeepViral model to reflect the communitative property – the model should extract the same features from both proteins and the input ordering of the sequence pair should not change the output prediction. Thus, on contrast to DeepViral where we use different parameters of the convolutional neural networks to learn from human and viral protein sequences, the transfer learning model uses a parallel model where model parameters are shared on both sides. The model architecture is illustrated below in Figure 5.1.

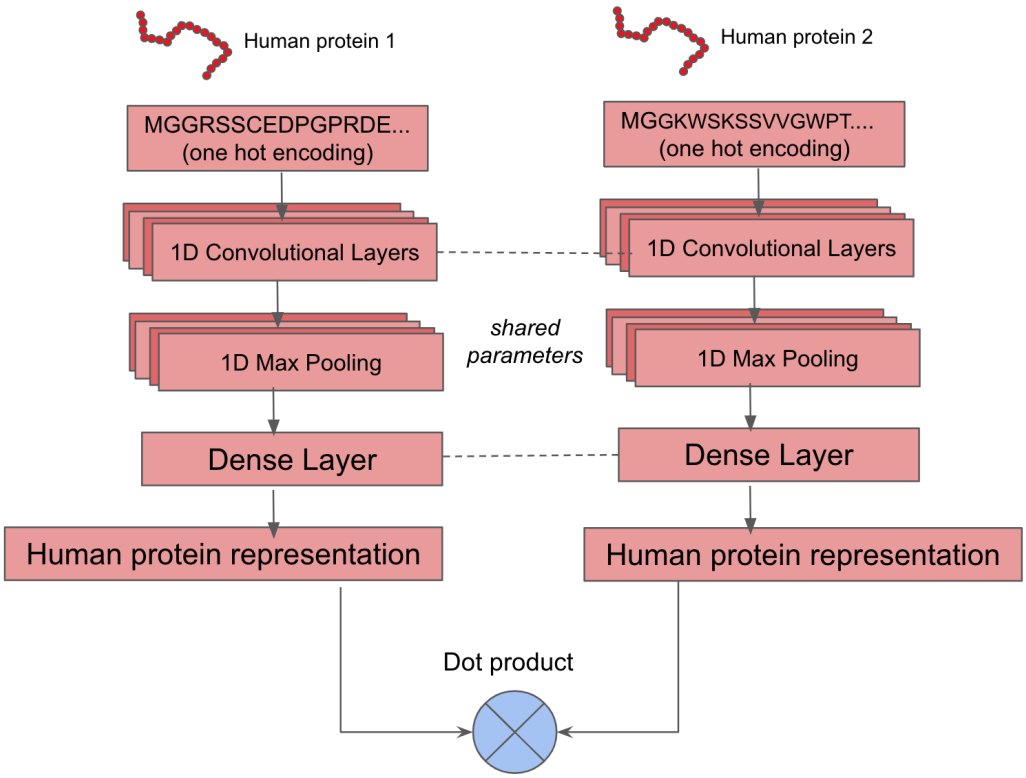


Figure 5.1: The model architecture of the pre-training model based on DeepViral for interacting human proteins. The same model is used for the fine-tuning task except that the second protein is replaced with the virus protein.

We pre-trained the model with the physical human PPI data from the STRING database [33] and saved the best performing model with a left-out testing set. The pre-trained model was then used for fine-tuning on our constructed dataset for novel virus prediction, followed by the same evaluation procedure in previous sections.

5.1.3 Results

We ran the transfer learning model, denoted as TransferViral, both with and without pre-training to test if there is an improvement after adding human data. The results are summarized in Table 5.1. With pre-training, the model gained improvements in a few metrics, which aligns with our hypothesis. However, the overall predictive performance is much lower compared to the original DeepViral model, the main difference

being that the parameters are not shared between human and viral proteins.

Method	Family-wise ROCAUC	Taxon-wise ROCAUC	Protein-wise ROCAUC	Mean rank
TransferViral (not pre-trained)	0.559 [0.506 - 0.613]	0.552 [0.501 - 0.602]	0.605 [0.572 - 0.638]	7753
TransferViral (pre-trained)	**0.628 [0.618 - 0.639]**	**0.665 [0.657 - 0.674]**	**0.606 [0.591 - 0.620]**	**7229**

Table 5.1: Comparison of the transfer learning model with and without pretraining.

5.1.4 Discussion

From the preliminary results, even though the pre-training with human PPI data improved over the model without pre-training, all the metrics are significantly lower than the original DeepViral described in previous chapters. The decline in performance may be due to the difference in the underlying interaction mechanisms between intra-human and virus–human PPIs, and thus the shared parameters for learning features of both proteins cannot extract the meaningful features required for virus–human protein pairs. One way to test this is to only used the pre-trained parameters for the human protein part of the original model, while keeping the virus side untouched. Another possibility for the performance drop is simply the size difference between the two models – due to the shared parameters, the new model has only half the number of parameters compared to before. To fully investigate and understand the potential of transfer learning in host–pathogen interaction prediction, it is necessary to test the model under different experimental settings and parameters, and we hope to explore this in the future.

5.2 Identifying proteins that elicit virus phenotypes

5.2.1 Motivation

In the previous sections, we used virus–human protein–protein interaction data as training data and infectious disease phenotypes as features for machine learning models. In this section, we investigate the possibility of directly mining statistically significant associations between infectious disease phenotypes and human proteins from our datasets.

Previously, biochemists have looked at whether they can find significant relationships between drugs side effects and human proteins based on a association network of drug–target and drug–side effects relationships [60]. Here, we follow a similar approach as [60] to identify potential human proteins that cause certain phenotypes from the association network of virus–human proteins and virus–phenotype associations, as illustrated in Figure 5.2. The identification of protein–phenotype associations can potentially suggest drug targets that inhibit certain proteins to alleviate their associated disease symptoms.

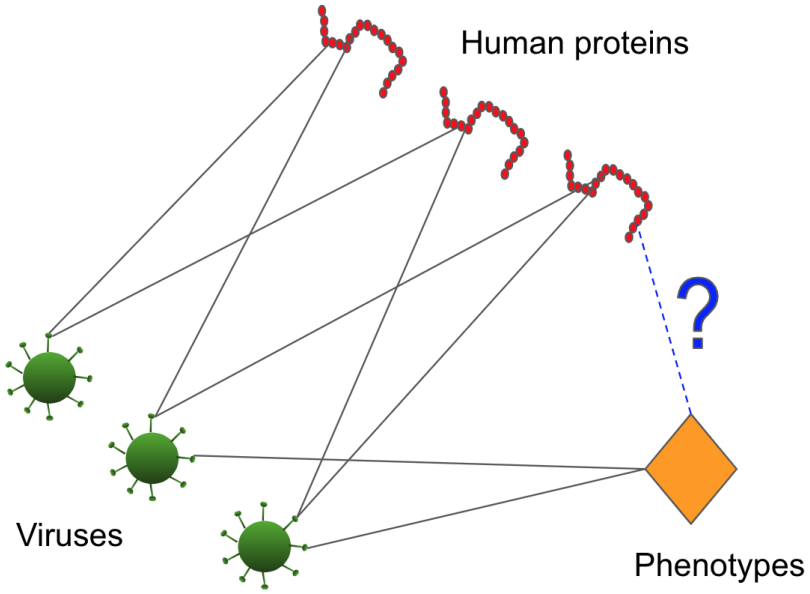


Figure 5.2: The association network based on the HPI dataset and the virus phenotype data. The dashed line indicates the edge that we want to identify.

5.2.2 Methods

To identify over-represented pairs of human proteins and virus phenotypes, we test the null hypothesis that a certain protein is independent from a certain phenotype based on the association data. We construct a 2 by 2 contingency table by counting, for every pair of human proteins and virus phenotypes, the number of viruses that both interact with the protein and cause the phenotype, that only interact with the protein, that only cause the phenotype and that do neither, as illustrated in Table 5.2.

	Interacts with protein A	Does not interact with protein A	Row total
With Phenotype P	a	b	a+b
Without Phenotype P	c	d	c+d
Column total	a+c	b+d	a+b+c+d=n

Table 5.2: A contingency table to test for significance between a pair of protein and phenotype. n is the total number of viruses in the analysis.

To test for statistical significance, we use Fisher's exact test, which computes the probability of the contingency table if the null hypothesis is true, i.e., the phenotype and the protein are independent of each other. The probability of a contingency

table based on Fisher's exact test is the same as the probability mass function of a hypergeometric distribution, given as,

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

where a, b, c, d, and n are defined the same way as in Table 5.2. We use the Scipy library and the Statsmodels library in Python, respectively, to perform the Fisher's exact test and subsequently correcting for multiple comparisons with false discovery rate.

Following [60], to ensure sufficient statistical power, we filter the dataset to exclude the human proteins and the phenotypes that are associated with less than five viruses.

5.2.3 Results

After filtering out the pairs that do not associate with enough viruses, we obtained a association network of 1,921 viruses, 1,643 human proteins, 2,240 phenotypes, and in total 230,213 pairs for testing significance. After performing the tests and correcting for false discovery rate, we reached a total of 3,523 statistically significant associations of human proteins and virus phenotypes ($\alpha = 0.05$). A few examples of such associations are listed in Table 5.3.

Protein Name	Protein Description	Phenotype	p-adjusted
PAF1	RNA polymerase II-associated factor 1 homolog	Brain inflammation	0.001
UBR4	E3 ubiquitin-protein ligase UBR4	Seborrheic keratosis	0.001
PTPN1	Tyrosine-protein phosphatase non-receptor type 1	Squamous Papilloma	0.002

Table 5.3: A list of statistically significant associations of proteins and phenotypes.

5.2.4 Discussion

After obtaining the statistically significant associations between human proteins and phenotypes, the next challenge is to validate them biologically, e.g., *in vivo* gene

knockout studies with mice [60]. Moreover, the number of highly significant associations is large and they are likely to contain false positives. This can be potentially alleviated by: 1) designing a more robust statistical test; 2) implementing a more stringent filtering to exclude phenotypes and human proteins that are less likely to be true positives due to biological relevance. For example, phenotypes related to "anoperineal fistula" are frequently occurring in the significant associations, but they could be of little biological interests. To fully understand the power of such a testing procedure with our datasets, it is necessary to work with biologists in the future to examine the biological relevance of the results.

REFERENCES

[1] F.-E. Eid, M. ElHefnawi, and L. S. Heath, "Denovo: virus-host sequence-based protein–protein interaction prediction," *Bioinformatics*, vol. 32, no. 8, pp. 1144–1150, 2016.

[2] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak, "Global trends in emerging infectious diseases," *Nature*, vol. 451, no. 7181, pp. 990–993, 2008.

[3] B. B. Finlay and P. Cossart, "Exploitation of mammalian host cell functions by bacterial pathogens," *Science*, vol. 276, no. 5313, pp. 718–725, 1997. [Online]. Available: http://science.sciencemag.org/content/276/5313/718

[4] M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali, and B. W. Sobral, "The human-bacterial pathogen protein interaction networks of bacillus anthracis, francisella tularensis, and yersinia pestis," *PLOS ONE*, vol. 5, no. 8, pp. 1–12, 08 2010. [Online]. Available: https://doi.org/10.1371/journal.pone.0012089

[5] M. D. Dyer, T. M. Murali, and B. W. Sobral, "The landscape of human proteins interacting with viruses and other pathogens," *PLOS Pathogens*, vol. 4, no. 2, pp. 1–14, 02 2008. [Online]. Available: https://doi.org/10.1371/journal.ppat.0040032

[6] T. Fajardo, Jr., P.-Y. Sung, and P. Roy, "Disruption of specific rna-rna interactions in a double-stranded rna virus inhibits genome packaging and virus infectivity," *PLOS Pathogens*, vol. 11, no. 12, pp. 1–22, 12 2015. [Online]. Available: https://doi.org/10.1371/journal.ppat.1005321

[7] M. D. Weitzman, C. T. Carson, R. A. Schwartz, and C. E. Lilley, "Interactions of viruses with the cellular dna repair machinery," *DNA Repair*, vol. 3, no. 8, pp. 1165 – 1173, 2004, bRIDGE OVER BROKEN ENDS - The Cellular Response to DNA Breaks in Health and Disease. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568786404000771

[8] X. Zhou, B. Park, D. Choi, and K. Han, "A generalized approach to predicting protein-protein interactions between virus and host," *BMC genomics*, vol. 19, no. 6, p. 568, 2018.

[9] S. Alguwaizani, B. Park, X. Zhou, D.-S. Huang, and K. Han, "Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids," *Journal of healthcare engineering*, 2018.

[10] X. Yang, S. Yang, Q. Li, S. Wuchty, and Z. Zhang, "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method," *Computational and structural biotechnology journal*, vol. 18, pp. 153–161, 2020.

[11] P. H. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: assessment with biological features and issues," *Briefings in Bioinformatics*, vol. 13, no. 5, pp. 569–585, 12 2011. [Online]. Available: https://doi.org/10.1093/bib/bbr066

[12] S. Jain and G. D. Bader, "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC bioinformatics*, vol. 11, no. 1, p. 562, 2010.

[13] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLoS computational biology*, vol. 5, no. 7, 2009.

[14] T. Huo, W. Liu, Y. Guo, C. Yang, J. Lin, and Z. Rao, "Prediction of host - pathogen protein interactions between mycobacterium tuberculosis and homo sapiens using sequence motifs," *BMC Bioinformatics*, vol. 16, no. 1, p. 100, Mar 2015. [Online]. Available: https://doi.org/10.1186/s12859-015-0535-y

[15] A. Mukhopadhyay, S. Ray, and U. Maulik, "Incorporating the type and direction information in predicting novel regulatory interactions between hiv-1 and human proteins using a biclustering approach," *BMC Bioinformatics*, vol. 15, no. 1, p. 26, Jan 2014. [Online]. Available: https://doi.org/10.1186/1471-2105-15-26

[16] M. Woolhouse, F. Scott, Z. Hudson, R. Howey, and M. Chase-Topping, "Human viruses: discovery and emergence," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1604, p. 2864, 2012.

[17] G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf, "The anatomy of phenotype ontologies: principles, properties and applications," *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 1008–1021, September 2018. [Online]. Available: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx035

[18] A. Oellrich, N. Collier, T. Groza, D. Rebholz-Schuhmann, N. Shah, O. Bodenreider, M. R. Boland, I. Georgiev, H. Liu, K. Livingston, A. Luna, A.-M. Mallon, P. Manda, P. N. Robinson, G. Rustici, M. Simon, L. Wang, R. Winnenburg, and M. Dumontier, "The digital revolution in phenotyping," *Briefings in Bioinformatics*, vol. 17, no. 5, pp. 819–830, 2016. [Online]. Available: http://dx.doi.org/10.1093/bib/bbv083

[19] P. N. Schofield, R. Hoehndorf, and G. V. Gkoutos, "Mouse genetic and phenotypic resources for human genetics," *Human Mutation*, March 2012. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/humu.22077/abstract

[20] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "Phenomenet: a whole-phenome approach to disease gene discovery," *Nucleic acids research*, vol. 39, no. 18, pp. e119–e119, 2011.

[21] T. F. Meehan, N. Conte, D. B. West, J. O. Jacobsen, J. Mason, J. Warren, C.-K. K. Chen, I. Tudose, M. Relac, P. Matthews, N. Karp, L. Santos, T. Fiegel, N. Ring, H. Westerberg, S. Greenaway, D. Sneddon, H. Morgan, G. F. Codner, M. E. Stewart, J. Brown, N. Horner, International Mouse Phenotyping Consortium, M. Haendel, N. Washington, C. J. Mungall, C. L. Reynolds, J. Gallegos, V. Gailus-Durner, T. Sorg, G. Pavlovic, L. R. Bower, M. Moore, I. Morse, X. Gao, G. P. Tocchini-Valentini, Y. Obata, S. Y. Y. Cho, J. K. K. Seong, J. Seavitt, A. L. Beaudet, M. E. Dickinson, Y. Herault, W. Wurst, M. H. H. de Angelis, K. C. K. Lloyd, A. M. Flenniken, L. M. J. Nutter, S. Newbigging, C. McKerlie, M. J. Justice, S. A. Murray, K. L. Svenson, R. E. Braun, J. K. White, A. Bradley, P. Flicek, S. Wells, W. C. Skarnes, D. J. Adams, H. Parkinson, A.-M. M. Mallon, S. D. M. Brown, and D. Smedley, "Disease model discovery from 3,328 gene knockouts by the international mouse phenotyping consortium." *Nature genetics*, vol. 49, no. 8, pp. 1231–1238, Aug. 2017.

[22] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson, "Clinical diagnostics in human genetics with semantic similarity searches in ontologies," *The American Journal of Human Genetics*, vol. 85, no. 4, pp. 457 – 464, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002929709003991

[23] R. Hoehndorf, T. Hiebert, N. W. Hardy, P. N. Schofield, G. V. Gkoutos, and M. Dumontier, "Mouse model phenotypes provide information about human drug targets," *Bioinformatics*, October 2013. [Online].

Available: http://bioinformatics.oxfordjournals.org/content/early/2013/10/23/bioinformatics.btt613.abstract

[24] R. Hoehndorf, N. W. Hardy, D. Osumi-Sutherland, S. Tweedie, P. N. Schofield, and G. V. Gkoutos, "Systematic analysis of experimental phenotype data reveals gene functions," *PLoS ONE*, vol. 8, no. 4, p. e60847, 04 2013. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0060847

[25] J. C. Whisstock and A. M. Lesk, "Prediction of protein function from protein sequence and structure," *Quarterly reviews of biophysics*, vol. 36, no. 3, p. 307, 2003.

[26] C. P. Ponting, "Issues in predicting protein function from sequence," *Briefings in bioinformatics*, vol. 2, no. 1, pp. 19–29, 2001.

[27] M. G. Ammari, C. R. Gresham, F. M. McCarthy, and B. Nanduri, "Hpidb 2.0: a curated database for host-pathogen interactions," *Database*, vol. 2016, p. baw103, 2016. [Online]. Available: http://dx.doi.org/10.1093/database/baw103

[28] S. Kafkas, M. Abdelhakim, Y. Hashish, M. Kulmanov, M. Abdellatif, P. N. Schofield, and R. Hoehndorf, "Pathophenodb: linking human pathogens to their disease phenotypes in support of infectious disease research," *bioRxiv*, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/12/09/489971

[29] S. Köhler, L. Carmody, N. Vasilevsky, J. O. B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry, D. Osumi-Sutherland, V. Cipriani, J. P. Balhoff, T. Conlin, H. Blau, G. Baynam, R. Palmer, D. Gratian, H. Dawkins, M. Segal, A. C. Jansen, A. Muaz, W. H. Chang, J. Bergerson, S. J. F. Laulederkind, Z. Yüksel, S. Beltran, A. F. Freeman, P. I. Sergouniotis, D. Durkin, A. L. Storm, M. Hanauer, M. Brudno, S. M. Bello, M. Sincan, K. Rageth, M. T. Wheeler, R. Oegema, H. Lourghi, M. G. Della Rocca, R. Thompson, F. Castellanos, J. Priest, C. Cunningham-Rundles, A. Hegde, R. C. Lovering, C. Hajek, A. Olry, L. Notarangelo, M. Similuk, X. A. Zhang, D. Gómez-Andrés, H. Lochmüller, H. Dollfus, S. Rosenzweig, S. Marwaha, A. Rath, K. Sullivan, C. Smith, J. D. Milner, D. Leroux, C. F. Boerkoel, A. Klion, M. C. Carter, T. Groza, D. Smedley, M. A. Haendel, C. Mungall, and P. N. Robinson, "Expansion of the human phenotype ontology (hpo) knowledge base and resources," *Nucleic Acids Research*, p. gky1105, 2018. [Online]. Available: http://dx.doi.org/10.1093/nar/gky1105

[30] C. L. Smith, J. A. Blake, J. A. Kadin, J. E. Richardson, C. J. Bult, and the Mouse Genome Database Group, "Mouse genome database (mgd)-2018: knowledgebase for the laboratory mouse," *Nucleic Acids Research*, vol. 46, no. D1, pp. D836–D842, 2018. [Online]. Available: http://dx.doi.org/10.1093/nar/gkx1006

[31] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[32] The Gene Ontology Consortium, "Expansion of the gene ontology knowledgebase and resources," *Nucleic Acids Research*, vol. 45, no. D1, pp. D331–D338, 2017. [Online]. Available: http://dx.doi.org/10.1093/nar/gkw1108

[33] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork *et al.*, "String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research*, vol. 47, no. D1, pp. D607–D613, 2019.

[34] U. Consortium, "Uniprot: a worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.

[35] M. Á. Rodríguez-García, G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf, "Integrating phenotype ontologies with phenomenet," *Journal of biomedical semantics*, vol. 8, no. 1, p. 58, 2017.

[36] R. Hoehndorf, L. Slater, P. N. Schofield, and G. V. Gkoutos, "Aber-owl: a framework for ontology-based data access in biology," *BMC bioinformatics*, vol. 16, no. 1, p. 26, 2015.

[37] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 37, no. suppl_1, pp. D5–D15, 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkn741

[38] J. Chen, A. Althagafi, and R. Hoehndorf, "Predicting candidate genes from phenotypes, functions, and anatomical site of expression," Mar. 2020. [Online]. Available: https://doi.org/10.1101/2020.03.30.015594

[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[40] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[42] M. Alshahrani and R. Hoehndorf, "Semantic disease gene embeddings (smudge): phenotype-based disease gene prioritization without phenotypes," *Bioinformatics*, vol. 34, no. 17, pp. i901–i907, 2018.

[43] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective," *Briefings in Bioinformatics*, March 2015. [Online]. Available: http://bib.oxfordjournals.org/content/early/2015/04/10/bib.bbv011.abstract

[44] F. Z. Smaili, X. Gao, and R. Hoehndorf, "Formal axioms in biomedical ontologies improve analysis and interpretation of associated data," *Bioinformatics*, 12 2019, btz920. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz920

[45] M. Kulmanov, F. Z. Smaili, X. Gao, and R. Hoehndorf, "Machine learning with biomedical ontologies," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/05/08/2020.05.07.082164

[46] M. Chen, C. J.-T. Ju, G. Zhou, X. Chen, T. Zhang, K.-W. Chang, C. Zaniolo, and W. Wang, "Multifaceted protein–protein interaction prediction based on siamese residual rcnn," *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, 2019.

[47] J. M. Villaveces, R. C. Jimenez, P. Porras, N. del Toro, M. Duesbury, M. Dumousseau, S. Orchard, H. Choi, P. Ping, N. Zong *et al.*, "Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study," *Database*, vol. 2015, 2015.

[48] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn Lett*, vol. 27, no. 8, pp. 861 – 874, 2006.

[49] M. Alshahrani, M. A. Khan, O. Maddouri, A. R. Kinjo, N. Queralt-Rosinach, and R. Hoehndorf, "Neuro-symbolic representation learning on biological knowledge graphs," *Bioinformatics*, vol. 33, no. 17, pp. 2723–2730, 04 2017. [Online]. Available: https://doi.org/10.1093/bioinformatics/btx275

[50] P. N. Schofield, R. Hoehndorf, C. L. Smith, J. T. Eppig, and G. V. Gkoutos, "25 - the informatics of developmental phenotypes," in *Kaufman's Atlas of Mouse Development Supplement*, R. B. B. R. D. Morriss-Kay, Ed. Boston: Academic Press, January 2016, pp. 307 – 318. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128000434000257

[51] A. Wang, S. Thurmond, L. Islas, K. Hui, and R. Hai, "Zika virus genome biology and molecular pathogenesis," *Emerging Microbes & Infections*, vol. 6, no. 3, p. e13, 2017.

[52] C.-Y. Chen, C.-H. Chan, C.-M. Chen, Y.-S. Tsai, T.-Y. Tsai, Y.-H. Wu Lee, and L.-R. You, "Targeted inactivation of murine ddx3x: essential roles of ddx3x in placentation and embryogenesis," *Human Molecular Genetics*, vol. 25, no. 14, pp. 2905–2922, 2016. [Online]. Available: http://dx.doi.org/10.1093/hmg/ddw143

[53] L. S. Blok, E. Madsen, J. Juusola, C. Gilissen, D. Baralle, M. R. Reijnders, H. Venselaar, C. Helsmoortel, M. T. Cho, A. Hoischen, L. E. Vissers, T. S. Koemans, W. Wissink-Lindhout, E. E. Eichler, C. Romano, H. V. Esch, C. Stumpel, M. Vreeburg, E. Smeets, K. Oberndorff, B. W. van Bon, M. Shaw, J. Gecz, E. Haan, M. Bienek, C. Jensen, B. L. Loeys, A. V. Dijck, A. M. Innes, H. Racher, S. Vermeer, N. D. Donato, A. Rump, K. Tatton-Brown, M. J. Parker, A. Henderson, S. A. Lynch, A. Fryer, A. Ross, P. Vasudevan, U. Kini, R. Newbury-Ecob, K. Chandler, A. Male, S. Dijkstra, J. Schieving, J. Giltay, K. L. van Gassen, J. Schuurs-Hoeijmakers, P. L. Tan, I. Pediaditakis, S. A. Haas, K. Retterer, P. Reed, K. G. Monaghan, E. Haverfield, M. Natowicz, A. Myers, M. C. Kruer, Q. Stein, K. A. Strauss, K. W. Brigatti, K. Keating, B. K. Burton, K. H. Kim, J. Charrow, J. Norman, A. Foster-Barber, A. D. Kline, A. Kimball, E. Zackai, M. Harr, J. Fox, J. McLaughlin, K. Lindstrom, K. M. Haude, K. van Roozendaal, H. Brunner, W. K. Chung, R. F. Kooy, R. Pfundt, V. Kalscheuer, S. G. Mehta, N. Katsanis, and T. Kleefstra, "Mutations in ddx3x are a common cause of unexplained intellectual disability with gender-specific effects on wnt signaling," *The American Journal of Human Genetics*, vol. 97, no. 2, pp. 343 – 352, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002929715002803

[54] P. Doñate-Macián, J. Jungfleisch, G. Pérez-Vilaró, F. Rubio-Moscardo, A. Perálvarez-Marín, J. Díez, and M. A. Valverde, "The trpv4 channel links calcium influx to ddx3x activity and viral infectivity," *Nature Communications*, vol. 9, p. 2307, 2018.

[55] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet Infectious Diseases*, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1473309920301201

[56] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, M. J. O'Meara, J. Z. Guo, D. L. Swaney, T. A. Tummino, R. Huettenhain, R. M. Kaake, A. L. Richards, B. Tutuncuoglu, H. Foussard, J. Batra, K. Haas, M. Modak, M. Kim, P. Haas, B. J. Polacco, H. Braberg, J. M. Fabius, M. Eckhardt, M. Soucheray, M. J. Bennett, M. Cakir, M. J. McGregor, Q. Li, Z. Z. C. Naing, Y. Zhou, S. Peng, I. T. Kirby, J. E. Melnyk, J. S. Chorba, K. Lou, S. A. Dai, W. Shen, Y. Shi, Z. Zhang, I. Barrio-Hernandez, D. Memon, C. Hernandez-Armenta, C. J. Mathy, T. Perica, K. B. Pilla, S. J. Ganesan, D. J. Saltzberg, R. Ramachandran, X. Liu, S. B. Rosenthal, L. Calviello, S. Venkataramanan, J. Liboy-Lugo, Y. Lin, S. A. Wankowicz, M. Bohn, P. P. Sharp, R. Trenker, J. M. Young, D. A. Cavero, J. Hiatt, T. L. Roth, U. Rathore, A. Subramanian, J. Noack, M. Hubert, F. Roesch, T. Vallet, B. Meyer, K. M. White, L. Miorin, O. S. Rosenberg, K. A. Verba, D. Agard, M. Ott, M. Emerman, D. Ruggero, A. García-Sastre, N. Jura, M. von Zastrow, J. Taunton, A. Ashworth, O. Schwartz, M. Vignuzzi, C. d'Enfert, S. Mukherjee, M. Jacobson, H. S. Malik, D. G. Fujimori, T. Ideker, C. S. Craik, S. Floor, J. S. Fraser, J. Gross, A. Sali, T. Kortemme, P. Beltrao, K. Shokat, B. K. Shoichet, and N. J. Krogan, "A sars-cov-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/03/27/2020.03.22.002386

[57] V. Thiel, K. A. Ivanov, A. Putics, T. Hertzig, B. Schelle, S. Bayer, B. Weißbrich, E. J. Snijder, H. Rabenau, H. W. Doerr *et al.*, "Mechanisms and enzymes involved in sars coronavirus genome expression," *Journal of General Virology*, vol. 84, no. 9, pp. 2305–2315, 2003.

[58] L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund *et al.*, "Analysis of the human tissue-specific expression by genome-wide integration of tran-

scriptomics and antibody-based proteomics," *Molecular & Cellular Proteomics*, vol. 13, no. 2, pp. 397–406, 2014.

[59] K. W. Jarosinski, S. Arndt, B. B. Kaufer, and N. Osterrieder, "Fluorescently tagged pul47 of marek's disease virus reveals differential tissue expression of the tegument protein in vivo," *Journal of Virology*, vol. 86, no. 5, pp. 2428–2436, 2012. [Online]. Available: https://jvi.asm.org/content/86/5/2428

[60] M. Kuhn, M. Al Banchaabouchi, M. Campillos, L. J. Jensen, C. Gross, A.-C. Gavin, and P. Bork, "Systematic identification of proteins that elicit drug side effects," *Molecular systems biology*, vol. 9, no. 1, p. 663, 2013.

APPENDICES

A Implementation details for comparison on the [1] dataset

The dataset contains contains 5,020 positives and 4,734 negatives in the training set, and 425 positives and 425 negatives in the testing set. Since no validation set was used previously, we constructed a validation set by randomly sampling 10% of the training set, which was used for choosing the best epoch for RCNN and DeepViral. We truncated all longer sequences than 2,000 amino acids to only the first 2,000 for RCNN, due to the maximum sequence length limit of the model (similarly, first 1,000 amino acids for DeepViral). RCNN and DeepViral (seq) were implemented and evaluated for the entirety of the test set. For DeepViral variants with feature embeddings, a limited number of protein pairs, i.e. 2% of the test set, do not have relevant features available (some proteins are obsolete due to database updates) and thus are excluded from the test set. We evaluated both RCNN and the variants of DeepViral for 5 times and report the mean of the metrics.

B Papers Submitted and Under Preparation

• Wang Liu-Wei, Şenay Kafkas, Jun Chen, Nicholas Dimonaco, Jesper Tegnér, Robert Hoehndorf, "DeepViral: infectious disease phenotypes improve prediction of novel virus–host interactions", *Under review at Bioinformatics*.