

Predicting the Interactions of Viral and Human Proteins

Fatma-Elzahraa S. Eid

Dissertation submitted to the Faculty of Engineering
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Lenwood S. Heath, Chair
Mahmoud ElHefnawi
Alexey Onufriev
Liqing Zhang
Bert Huang

April 19, 2017
Blacksburg, Virginia

Keywords: Protein-Protein Interaction, Virus, Machine Learning, Zika Virus
Copyright 2017, Fatma-Elzahraa S. Eid

Predicting the Interactions of Viral and Human Proteins

Fatma-Elzahraa S. Eid

Abstract

The world has proven unprepared for deadly viral outbreaks. Designing antiviral drugs and strategies requires a firm understanding of the interactions taken place between the proteins of the virus and human proteins. The current computational models for predicting these interactions consider only single viruses for which extensive prior knowledge is available. The two prediction frameworks in this dissertation, DeNovo and DeNovo-Human, make it possible for the first time to predict the interactions between any viral protein and human proteins. They further helped to answer critical questions about the Zika virus.

DeNovo utilizes concepts from virology, bioinformatics, and machine learning to make predictions for novel viruses possible. It pools protein-protein interactions (PPIs) from different viruses sharing the same host. It further introduces taxonomic partitioning to make the reported performance reflect the situation of predicting for a novel virus. DeNovo avoids the expected low accuracy of such a prediction by introducing a negative sampling scheme that is based on sequence similarity. DeNovo achieved accuracy up to 81% and 86% when predicting for a new viral species and a new viral family, respectively. This result is comparable to the best achieved previously in single virus-host and intra-species PPI prediction cases.

DeNovo predicts PPIs of a novel virus without requiring known PPIs for it, but with a limitation on the number of human proteins it can make predictions against. The second framework, DeNovo-Human, relaxes this limitation by forcing in-network prediction and random sampling while keeping the pooling technique of DeNovo. The accuracy and AUC are both promising ($> 85\%$, and $> 91\%$ respectively). DeNovo-Human facilitates predicting the virus-human PPI network.

To demonstrate how the two frameworks can enrich our knowledge about virus behavior, I use them to answer interesting questions about the Zika virus. The research questions examine how the Zika virus enters human cells, fights the innate immune system, and causes microcephaly. The answers obtained are well supported by recently published Zika virus studies.

This work received support from the VT-MENA program of Egypt, and ITIDA-Egypt.

Predicting the Interactions of Viral and Human Proteins

Fatma-Elzahraa S. Eid

General Audience Abstract

When a virus attacks a human body, it disturbs the host cells by interacting with their proteins. Identifying these interactions is key to fighting the virus. In this dissertation, I developed two computational tools to identify the interactions for any virus infecting the human. I further used these tools to answer interesting questions about the Zika virus behavior. The results are in agreement with recently published experimental studies about the virus.

Dedication

To all lives threatened by viruses,
THAT I MAY SERVE.

Acknowledgments

“That is from the favor of Allah upon us and upon the people.”, Yusuf 38, The Holy Quran.

First and foremost, I thank my generous Creator for making me thirsty to knowledge and exploring and for giving me the opportunity to direct my research for the benefit of human health; that I may serve. He has guided me through tough decisions and has sent me all those who supported and helped me during my Ph.D. journey.

To my supervisor, Lenwood Heath: because you have pushed me to become who I am today. Many Thanks! My Egyptian supervisor, Mahmoud ElHefnawi: because you opened the door of computational virology for me. Many Thanks! My committee members, Liqing Zhang, Alexey Onufriev, and Bert Huang: because your valuable comments directed my research to unvisited lands. Many Thanks!

I am grateful to my husband, Haitham Elmarakeby, who has been my best friend and supporter; he has made this exhausting journey much more pleasant. I am grateful to my son, Noureddeen, for paying his share in this journey while he does not understand why mommy is not playing with him or giving him more time. I am also grateful to my family members and my friends in Egypt who have supported me along the way.

A very special gratitude goes out to the Muslim community in Blacksburg, especially my friends who has been like my family, who helped me a lot taking care of my son, and who supported me through rough times. With a special mention to Doaa Altarawy for bearing my eccentric moments and ideas. Sally, Ayat, Eman, Nahla, Rehab, Asmaa, Amany, Nidaa, and Zulikha: you made a difference! Alida Palmisano: you definitely made a difference!

Special thanks go to Dr. Sedki Riad for his support and the launching of the VT-MENA program that funded my study. I am thankful also to the ITIDA program of Egypt and the Pratt Fellowship, to which I was nominated by my supervisor, for their funding.

The graduate school has facilitated my stay and let me feel at home. Thank you! What a comfortable place to work! The Computer Science Department at Virginia Tech has given me the opportunity to teach CS5046 that changed my life. Thank you! I am grateful to the students of CS5046 as well.

Thanks for all your support!

Contents

1	Introduction	1
1.1	Intra- and Inter-species Protein-Protein Interactions	1
1.2	Protein-Protein Interaction Prediction	2
1.3	Motivation	4
1.4	Objectives	5
1.5	Solution Approach	5
1.6	Contributions	7
2	Literature Review	9
2.1	Predicting Protein Interactions	9
2.2	Inter-species Prediction	11
2.3	Negative Example Selection	15
3	DeNovo	17
3.1	Introduction	17
3.1.1	Problem Definitions	17
3.1.2	Hypotheses	19
3.2	Methods	20
3.2.1	Data Set and Preparation	20
3.2.2	Dissimilarity-Based Negative Sampling	21
3.2.3	Data Partitioning	22
3.2.4	Learning Model	24

3.2.5	Supplementary Studies Data	26
3.2.6	Pilot and Generalization Studies	26
3.3	Results	29
3.4	Discussion	32
3.4.1	Negative Sampling	33
3.4.2	Comparison with Data Mining-based Sampling	35
3.4.3	Data Partitioning	36
3.4.4	Features	37
3.4.5	Prediction Approach	38
3.4.6	Generalization	39
3.4.7	Application Scenarios	40
3.5	Conclusion	40
4	DeNovo-Human	41
4.1	Introduction	41
4.2	Methods	43
4.2.1	Data set	43
4.2.2	Learning	44
4.2.3	Framework evaluation	45
4.2.4	Network prediction	46
4.3	Results	47
4.4	Discussion	50
4.4.1	System design	50
4.4.2	Comparison to related work	50
4.4.3	Comparison to PPI prediction in Human	53
4.4.4	DeNovo-Human advantage	53
4.4.5	Virus-human PPI network	55
4.4.6	Redundancy	55
4.4.7	Predicting for a query viral protein	56

4.5	Conclusion	56
5	Case Study: Zika Virus	57
5.1	Introduction	57
5.2	Research Questions	58
5.3	Data	59
5.4	DeNovo-Cascade	60
5.5	Results	62
5.6	Discussion	64
5.6.1	Viral perturbation to human signaling pathways	64
5.6.2	Research Question 1	67
5.6.3	Research Question 2	68
5.6.4	Research Question 3	70
5.6.5	Research Question 4	73
5.7	Conclusions	74
6	Conclusions and Future Work	75
6.1	Conclusions	75
6.2	Future Work	77
	Bibliography	79

List of Figures

1.1	PPI prediction general framework. A set of PPIs is fed into the framework, negative examples are generated, features are extracted, then positive and negative PPIs are split into training and testing. A classifier is trained on the training set and evaluated on the testing set. For the typical virus-host PPI prediction, the input data set consists of those PPIs among a specific virus and its host. In our solutions, the input set is a pool of PPIs between different viruses and their shared host.	3
1.2	Summary of the capabilities of the three models developed in this thesis for predicting virus-human PPI prediction. DeNovo can predict PPIs of a novel viral protein from its sequence but can only predict for a limited set of human proteins for each viral protein. DeNovo-Human can do the opposite; it can predict PPIs of a viral protein against all human proteins, but the viral protein cannot be novel, for DeNovo-Human requires few PPIs of that viral protein for training. DeNovo-Cascade combines the two models to achieve prediction of a novel virus against all human proteins.	6
3.1	DeNovo workflow. Virus-human PPIs are first cleaned, negative interactions are generated using the proposed dissimilarity-based method, and features are extracted from the protein sequences of the positive and negative interaction pairs. The feature vectors of the interactions are partitioned into multiple pairs of training and testing subsets based on a criterion Cr . A machine learning model is trained on the feature vectors from a training subset and tested on the corresponding testing subset in a cross validation fashion. . .	18
3.2	Taxonomy tree of the viruses with interactions in VirusMentha.	21
3.3	DISSIMILARITY-RANDOM-SAMPLING algorithm: selecting less likely interactions as negative PPI for virus-host PPI prediction applications.	23

3.4	An example of partial blindness partitioning. Each leaf node n_x in the taxonomic tree carries interactions of a virus v_x . Nodes at the same level belong to viruses of the same taxonomic rank (e.g. species or strain). Interactions of nodes sharing an immediate ancestor are grouped together into a single subset, denoted by a dashed contour.	24
3.5	Accuracy, sensitivity, and specificity measured at different sequence dissimilarity threshold values, testing with (A) partial blindness, and (B) full blindness grouping criteria. Dotted lines are for visualization purposes.	30
3.6	Confidence of prediction with (A) partial blindness grouping and (B) full blindness grouping, using the model optimal parameters (C^* , γ^* , and T^*). . .	30
3.7	ROC and precision-recall (PR) curves for testing DeNovo with the two partitioning criteria $Cr1$ and $Cr2$	32
3.8	Confidence of positive PPI prediction in studies ST4 (Left) and ST5 (Right). Sensitivity in both cases was about 94%.	33
3.9	Overview of DeNovo key ideas. [1] Instead of using a single virus PPIs with the host proteins (As in typical prediction of interactions between virus and host proteins), DeNovo utilizes PPIs from different viruses interacting with the host. This makes predictions possible for different viruses. [2] DeNovo does not divide the data set of PPIs randomly into training and testing. Instead, it groups the interactions of each virus (similar viruses or within a viral family) into a single partition. All partitions except one are then used for training and the excluded one for testing to imitate the situation of a novel virus. [3] Dissimilarity-based negative sampling basic idea. If two proteins are sequence-dissimilar, they are less likely to share interaction partners. Hence, the true partners of one may serve as the negative partners of the other.	34
4.1	DeNovo framework limitation on the testing library size. DeNovo is designed to discard specific human proteins from training (sampled for each viral protein) to reduce the noise in the negative random sampling process. As a result of this subsampling, for each query viral protein, a set of human proteins in the library (used to train DeNovo) is excluded as well. DeNovo performance cannot be extended to the excluded sets. The figure shows the distribution of the percentage of human proteins that can be tested for a set of 445 viral protein at each value of the dissimilarity threshold (used in the subsampling). Usually, DeNovo is operated in the range [0.4,0.7]. The theoretical optimal point is 0.8. Thus, the coverage power of DeNovo is mostly less than 20% of the human library size.	42

4.2	Area under ROC and PR curves for the three kernel methods compared for feature encoding. The three methods $M1$, $M2$, and $M3$ correspond to the three kernels K_1 , K_2 , and K_3 discussed in Section 4.2.2.	48
4.3	Performance for prediction over 37 single virus-human networks. Average, median, and maximum coverage are 351, 262, and 1569 human proteins. . . .	48
4.4	Estimated performance of DeNovo-Human with different number of PPIs given for a query viral protein. Accuracy, sensitivity, specificity, and AUPR are fitted to generate the tendency curves, with R^2 of 0.75, 0.78, 0.82, and 0.78, respectively. The balanced point occurs when a query viral protein has about 30 known PPIs with human proteins fed into DeNovo-Human. One can select different operating points according to their application tolerance and constraints.	54
5.1	Zika virus genome and proteins (Source: ViralZone: www.expasy.org/viralzone , SIB Swiss Institute of Bioinformatics) [47]	58
5.2	DeNovo Cascade. Prediction throughout this case study is prepared by cascading the two DeNovo models developed earlier. Positive and negative highly confident predictions from DeNovo (for a viral protein v_x with no known PPIs) are used as seed PPIs fed into DeNovo-Human.	60
5.3	Summary of Zika interference with IFN pathways.	65
5.4	How Zika virus causes microcephaly. ZIKV predicted interactions with the mTOR pathway show high interference potential, especially by ZIKV NS4A. The predicted suppression of mTORC1 signaling should lead to increased autophagy in fNSC, which in turn can lead to neurogenesis defects and consequently fetal microcephaly.	66
5.5	Dengue virus (DENV) potential interactions with the mTOR pathway demonstrate its interference capabilities with mTOR signaling and its potential to cause autophagy. The difference in ZIKV and DENV predicted PPIs suggests that either DENV cannot efficiently block the IGF to maintain upregulated autophagy or that DENV can induce autophagy but cannot reach the fetal cells at the first place to cause the downstream microcephaly.	67
5.6	Reasoning about how Zika virus might be causing microcephaly. Going backward from microcephaly, it is caused by disturbing neurogenesis by autophagy. Autophagy, in turn, is induced by suppressing mTOR. Thus, ZIKV is probably causing microcephaly by suppressing mTOR in fetal NSCs.	70
5.7	Relevant cascades in the mTOR pathway.	71

5.8	Simplified diagram of the key interference points of Zika virus and Dengue virus with the mTOR pathway. The blue and yellow hexagons represent proteins of Zika virus and Dengue virus, respectively.	72
-----	---	----

List of Tables

3.1	Summary of the pilot and generalization studies in Section 3.2.6.	27
3.2	Viral families resulted from partitioning the interactions according to the complete blindness criterion. The sample size for each family is the number of positive PPIs in the data set involving viruses from that family.	31
4.1	Comparison of the average performance over 40 sampling sets for the three top feature encoding kernel methods for the DeNovo-Human framework. The first method is the one I use in DeNovo-Human for reporting and final product.	47
4.2	Similarity Test 1. Comparison of the average performance when the viral proteins in the training set are at most Sim% similar to each other. This test examines the prediction performance when the model is trained on interactions from a viral protein v and other viral proteins similar to v and then tested to predict PPIs for v . The question is whether adding PPIs of viral proteins similar to v (along with PPIs of v) in training improves the prediction for v .	49
4.3	Similarity Test 2. Comparison of the average performance when the viral proteins in the training set are at least Sim% similar to the query viral protein v , but v does not have any of its PPIs in training. The question is whether having PPIs of viral proteins similar to v but none for v in training substitute for having the model trained on PPIs of v	49
5.1	Predicted interactions of Zika virus protein E with human cell entry receptors. The strict prediction reduced the search space (45 human cell receptor proteins previously reported to internalize some viruses) to 12 candidate receptors, while the average prediction reduced it to 21.	62
5.2	Predicted interactions of Zika virus proteins with the interferon circuit proteins.	63
5.3	Predicted interference of Zika virus and Dengue virus with the mTOR pathway, which is a key pathway whose interruption in the fetus may cause microcephaly.	64

Chapter 1

Introduction

1.1 Intra- and Inter-species Protein-Protein Interactions

Proteins do not work in isolation but interact with other proteins and molecules to perform functions including signal transduction, replication, and transportation inside and across the cell [10]. Protein-protein interaction (PPI) can be defined as the intentional functional and physical contact between non-generic interaction interfaces of two proteins in a biological context [26]. Temporal, spatial, and modification constraints control the occurrence of interactions between two proteins, leaving us far from a complete map of protein-protein interactions of any organism [84]. Such constraints include co-localization, co-expression or transportation, phosphorylation, regulatory mechanisms, and three-dimensional structure changes. Protein-protein interactions (PPIs) in the cell are typically part of signaling and metabolic pathways whose disturbance can cause diseases [54]. Connected pathways along with other interactions that are not mapped to known pathways form networks of protein-protein interactions. These networks tend to have central nodes whose knock-out can result in lethal phenotypes, node clusters, and other poorly connected nodes [10, 49].

There are two hypotheses on how two proteins may interact [13]. One hypothesis is that gradual conformational changes occur upon an initial interaction between the two proteins. The other hypothesis assumes that, prior to binding, a protein selects a high-energy conformation of its binding partner, among an ensemble of existing conformations, and binds to that conformation. Whichever of the two models governs the interaction mechanism, the physical interaction happens between domains or motifs of the two proteins. It is believed that a few residues, called hotspots, on the interaction interfaces are critical and that hotspot mutations can disable the interactions [88].

Several large-scale and small-scale experimental methods to identify protein-protein interac-

tions have been proposed and put into production over the last three decades [10]. Large-scale identification methods, such as yeast two-hybrid, are known to have high rates of false positives and false negatives, estimated at 30-60% and 40-80%, respectively [121]. Overlap of the identified interactions among the output of those methods and their reproducibility are poor [67, 115]. It is thus critical to validate the identified interactions and use quality control when considering the identified interactions for further studies [10, 67].

Protein-protein interactions also occur between proteins of different species, basically a pathogen and its host, to facilitate the pathogen attacking the host. In the virus-host case, a virus needs to hijack the host cell regulation for its benefit and to survive the host defenses [24]. Viruses disturb the signaling pathways of the cell by forcing their own proteins to interact with some host proteins or to disturb ongoing interactions in the host cell.

1.2 Protein-Protein Interaction Prediction

Identifying pairs of proteins that can interact can be done experimentally in wet labs, for example, using yeast two-hybrid screening, co-immunoprecipitation, or tandem affinity purification. However, each protein-protein interaction identification method has its own limitations. Additionally, some proteins cannot be prepared for some of these methods. Funding can also be a limiting factor when performing interaction identification on a large scale [54]. A computational tool that predicts whether two proteins are likely to interact can contribute to solving these limitations. Its top benefit is the immediate and inexpensive determination of results. Additionally, a computational method for protein-protein interaction determination may be able to avoid the wet lab assay limitations and operate on those proteins that cannot be isolated.

Many protein-protein interaction prediction computational frameworks have been proposed since the first one in 1998 [22]. These methods can be divided into two main approaches: modeling protein-protein interaction concepts using domain knowledge and automatic learning to identify putative interacting pairs using statistical or machine learning. Most of the prediction methods to date belong to the learning approach, for modeling-based methods require a deep understanding of the interaction molecular mechanisms and have strong limitations that hinder their generalization [125]. In a typical PPI learning-based prediction framework, a set of unlikely interactions is generated from the available known interactions. A feature extraction procedure is then run over that set of positive and negative (unlikely) PPIs to generate feature vectors to represent each protein pair. The data set is then split into training and testing subsets. A machine learning model is trained on the training set and then evaluated on the testing set. See Figure 1.1.

Most of the frameworks in the field of PPI prediction are designed for interactions among proteins of the same species (intra-species prediction), mostly human and yeast. Interactions can occur between proteins of two organisms, mainly in a host-pathogen context. With the

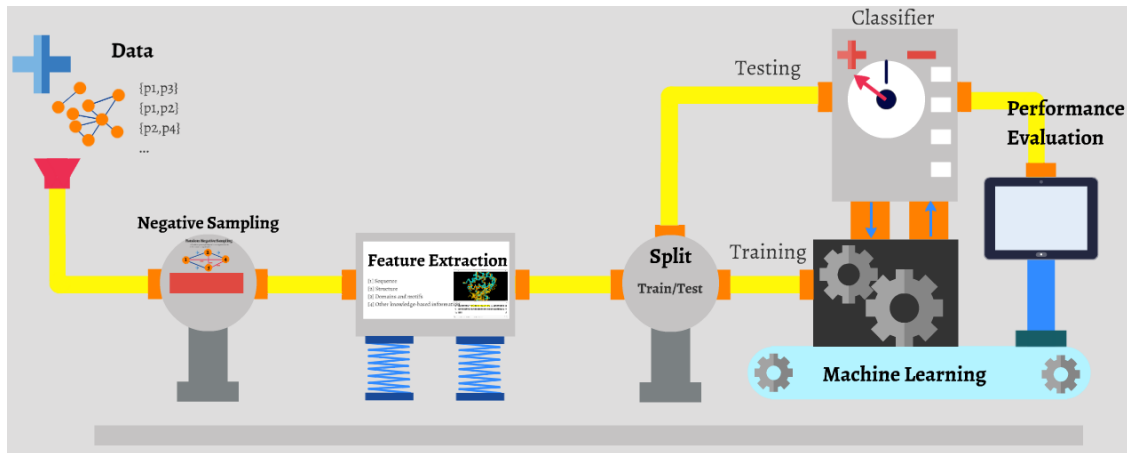


Figure 1.1: PPI prediction general framework. A set of PPIs is fed into the framework, negative examples are generated, features are extracted, then positive and negative PPIs are split into training and testing. A classifier is trained on the training set and evaluated on the testing set. For the typical virus-host PPI prediction, the input data set consists of those PPIs among a specific virus and its host. In our solutions, the input set is a pool of PPIs between different viruses and their shared host.

importance of determining interactions of the proteins in understanding the host-pathogen relations, several studies designed frameworks for predicting protein-protein interactions between two species (inter-species prediction) [85]. Experimentally reported host-pathogen interactions are few in number [15, 38, 60], increasing the need for computational methods to fill in this gap. Additionally, the experimental methods of identifying such interactions can be time-consuming, expensive, or life-threatening. It has long been thought that predicting interactions between proteins of different organisms is more difficult than for the single organism case, as it involves proteins from two systems different in nature.

Predicting the interactions between proteins of a virus and its host is an inter-species prediction problem. The prediction of HIV protein interactions with human proteins has dominated the field of inter-species PPI prediction, for the abundance of the reported interactions that can be used for training a model and the abundance of knowledge about HIV proteins [33, 37, 106]. However, some frameworks proposed solutions for the scarcity problem of the host-pathogen interactions by applying transfer and multi-task learning techniques, extending the prediction to HTLV (Human T-lymphotropic virus) [75]. A few studies applied the methods proposed for HIV to other viruses, but the data sets are much smaller than in the HIV case, making the learning, evaluation, and generalization efficiency questionable [21, 36].

All existing machine learning virus-host PPI prediction methods suffer four common limitations. First, they extend known interaction sets for a single virus only. Second, the learning models need relatively large data sets of known interactions. They are typically designed for the human immunodeficiency virus (HIV) proteins [33, 37, 113], for such large

interaction sets are unavailable for almost every other virus. Third, this centric single-virus approach wastes information that could otherwise be extracted from the large number of known interactions of host proteins with proteins of different viruses. Fourth, these methods use machine learning models as black boxes, with no biological interpretation of how they perform or what biological factors guide their learning.

1.3 Motivation

Knowledge of PPIs between a virus and its host can reveal much about how the virus infects the host, hijacks cell regulations, and proceeds in its life cycle, as well as how the host responds to the virus [82]. Such understanding supports the design of more effective therapeutics and antiviral strategies [37, 82]. Although numerous virus-host databases have been recently curated [15, 38, 60], they are incomplete and insufficient, demanding creative solutions to reveal more of these interactions [85]. Accurate computational prediction of virus-host PPIs stands as a promising solution to reduce the cost, time, and risk of wet lab experimentation [37, 85].

Having a computational tool to predict candidate interactions between a virus and its host should help to accelerate virus-host research. However, current approaches for predicting PPIs between a virus and its host require at least hundreds of known PPIs between them to learn from and make further prediction for putative PPIs. Except for HIV proteins, all viral proteins have few or no known interactions with their host proteins, hindering the application of the single virus-host prediction methods to those viruses [3]. This forms the major motivation of this thesis: Can we find a way to predict PPIs for other viruses with their hosts without the need for a large set of examples of their interactions? If we have a deadly outbreak like the recent outbreaks of Ebola virus [40] and Zika virus [20], should we wait for some years before being able to use the current virus-host PPI prediction frameworks? More valuable would be if we can predict the interactions between a virus and its host without any known interactions between them at all.

The second motivation, which is related to the one above, is the need to make a generalized tool that does not depend heavily on experimental data of proteins to learn from. Many PPI prediction models depend on protein features beyond sequence, like three-dimensional structure, localization, and knowledge of interaction domains. This experimental knowledge can be expensive or impossible to obtain for some viruses. Thus, it is essential to make a tool that depends on minimum protein knowledge to guarantee generalization ability.

The third major motivation of this work is to convince biologists to use this type of prediction framework. Looking at the common trend in using computational tools of PPI prediction by biologists, one can hardly find much biological research that builds upon similar computational prediction tools. One reason can be that biologists do not trust such tools over wet-lab work. Another reason can be that the available tools are not prepared or presented

as usable by the biologists. For example, almost all of the studies proposing the PPI prediction tools present the performance of the tool but without elaboration on how to use the tool for further research. These two issues might be resolved by showing a working example that demonstrates how to use a PPI prediction tool and how effective it can be.

1.4 Objectives

The four objectives of this thesis follow.

Objective 1. Building a framework that enables the prediction of protein-protein interactions between a virus and its host when the virus proteins do not have a large set of known interactions with that host (Chapter 4).

Objective 2. Developing a framework that predicts interactions between a novel virus with no known interaction and human proteins using easily obtainable features to ensure generalization (Chapter 3).

Objective 3. Designing a way to partition the known interactions employed in building the PPI prediction framework (and future frameworks dealing with virus-host) in a way that simulates predicting for a novel virus, taking in consideration the similarity between viral families (Chapter 3).

Objective 4. Demonstrating the value of the prediction frameworks designed to predict PPI between viruses and their hosts through a working example in which critical questions about a virus are answered (Chapter 5).

1.5 Solution Approach

In my first framework DeNovo (see [35] and Chapter 3), I address the generalized problem of predicting PPIs between a virus and its host without the need to know any of their interactions in advance. The key idea is to learn interaction features for the host proteins rather than for specific viral proteins. The model depends only on protein sequences to extract features and generate unlikely negative examples to ensure generalization. To simulate the situation when predicting for a new viral protein, I propose a taxonomic- and similarity-based partitioning scheme of the interactions used in training and evaluating the model. I further make three hypotheses on what attributes potentially guide the learning process and

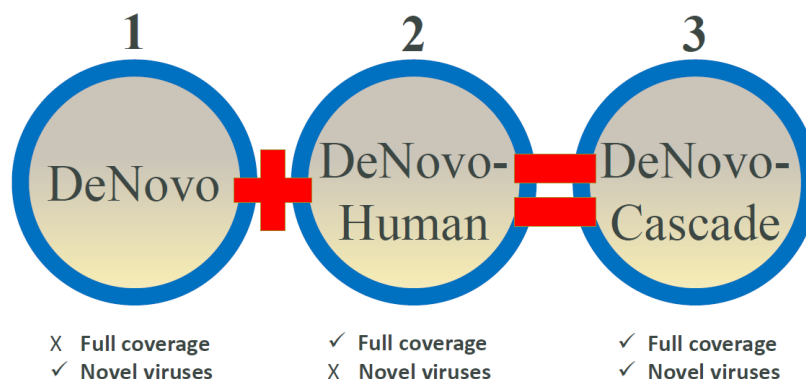


Figure 1.2: Summary of the capabilities of the three models developed in this thesis for predicting virus-human PPI prediction. DeNovo can predict PPIs of a novel viral protein from its sequence but can only predict for a limited set of human proteins for each viral protein. DeNovo-Human can do the opposite; it can predict PPIs of a viral protein against all human proteins, but the viral protein cannot be novel, for DeNovo-Human requires few PPIs of that viral protein for training. DeNovo-Cascade combines the two models to achieve prediction of a novel virus against all human proteins.

use the framework to test them to link biological and computational aspects of the prediction process.

DeNovo demonstrates the feasibility of predicting interactions of a viral protein with human proteins by learning from the available interactions of other viruses that also infect human, even with little or no sequence similarity to the viral proteins in training. The framework achieves accuracy comparable to the best reported for intra-species PPI prediction (using sequence information) despite the challenging nature of the generalized problem. I additionally demonstrate DeNovo generalization to other pathogens. DeNovo is limited to predicting the interactions of viruses infecting human because of the lack of data to validate the model for other hosts.

DeNovo provides prediction for novel viruses with no known interactions but cannot make the prediction against all the human proteins due to the way the unlikely interactions are sampled (Chapter 3). Particularly, PPI prediction is classified into three classes: C1, where both protein pairs of a testing sample have PPI examples used in training the prediction model; C2, where only one protein of the pair has examples; and C3, where neither of the testing protein pair has training PPIs examples [91]. Accuracy measures in C1 are much higher than those of C2 and C3. DeNovo belongs to C2 where testing viral proteins are novel. However, we managed to avoid the expected low accuracy of C2 in DeNovo by constraining the training sampling (at the cost of future testing limitation too). See Figure 1.2.

If we require a few PPIs of a testing viral protein to be known in advance and use them in training the prediction model, the problem that DeNovo addresses will be transformed from

a C2 class problem to a C1 problem (where accuracy is high without constraints). When random negative sampling is used instead of the dissimilarity sampling (which was proposed in DeNovo to obtain higher accuracy in C2), we can achieve broader coverage of human proteins to test while enjoying the high accuracy of C1. This approach forms the core of the DeNovo-Human framework proposed in Chapter 4.

This is an interesting time to demonstrate how the DeNovo models can be used to answer some critical questions about a virus behavior, especially with the recent viral outbreaks. I thus used the DeNovo models to predict the potential PPIs between Zika virus and all the human proteins known to interact with other viruses (Chapter 5). Then the human proteins of interest are examined to answer puzzling questions about Zika virus (ZIKV) behavior (Chapter 5): (1) how ZIKV enters the human cell; (2) how ZIKV fights interferon mechanism; (3) how ZIKV causes microcephaly; and (4) why a closely related virus to ZIKV does not cause microcephaly as ZIKV does. The answers obtained by applying the DeNovo models come in agreement with recently published experimental studies about Zika virus.

1.6 Contributions

The research in this thesis is exploratory in nature, leading to conceptual, computational, and biological contributions as summarized below.

- Changing the way the problem of predicting the interactions between viral and host proteins is addressed. First, predicting the interactions for a novel viral protein with a specific host is introduced (Chapter 3). Second, a data partitioning scheme is designed to split the known virus-human interactions into training and testing in a way that simulates the situation of predicting the interactions of a new virus (Chapter 3).
- Making it possible, for the first time, to predict PPIs between human and any virus infecting it, even if it is a novel or a less-studied virus. Two frameworks are proposed, DeNovo (Chapter 3) and DeNovo-Human (Chapter 4), to address these two aspects of the problem.
- Adopting molecular virology knowledge to find a solution for the computational challenge of predicting for a virus that has insufficient PPIs to train a learning model (Chapter 3 and 4).
- Demonstrating the capabilities and usefulness of the developed frameworks and how they can be used to answer critical questions about virus behavior only from viral protein sequences (Chapter 5). This should pave the way for biologists to use and put confidence in similar frameworks.
- Proposing a set of scenarios of how Zika virus causes fetal microcephaly while Dengue virus does not (Chapter 5). Additionally, I identify the human cell entry receptors

potentially responsible for internalizing the Zika virus and the association between Zika virus and Interferon- β pathways.

Chapter 2

Literature Review

2.1 Predicting Protein Interactions

Identifying proteins that can interact together can be done experimentally in wet labs, for example using yeast two-hybrid screening, co-immunoprecipitation, or tandem affinity purification. However, each protein-protein interaction identification method has its own limitations. Additionally, some proteins cannot be extracted and prepared for some of these methods. Funding can also be a limiting factor when performing interaction identification on a large scale [54].

A computational tool that predicts whether two proteins are likely to interact can contribute to addressing these limitations. Its top benefit would be the immediate and inexpensive determination of results. Additionally, a computational method for protein-protein interaction determination may be able to avoid the wet lab assay limitations and operate on those proteins that cannot be isolated.

To the best of my knowledge, the first model to predict protein interactions was proposed in 1998 [22]. In that study, the authors observed that conservation of gene neighborhood across multiple genomes suggests conservation of the interactions between those gene products. They further employed their observation to map known interactions between different organisms.

Several methods started to emerge afterward for predicting protein-protein interactions computationally. These methods can be divided into two main approaches: modeling and learning. In the modeling approaches, knowledge about how proteins interact are abstracted in computational models that can then be used to assess whether a protein pair matches that model and consequently can interact. The learning approaches utilize proteins that are reported to interact and learn computationally to discriminate between pairs of proteins that are likely or unlikely to be engaged in physical interactions. Most of the prediction

methods to date belong to the learning approach, for modeling-based methods require a deep understanding of the interaction process and have strong limitations that hinder their generalization [22].

Learning-based prediction employs statistics or machine learning approaches to perform the learning process based on some biological basis. Thus, PPI prediction methods can be classified based on the computational or biological paradigm they follow. From the biological point of view, learning-based PPI prediction methods utilize homology, domain-pairing, structure, gene neighboring, gene expression, and protein sequence [101]. From the computational point of view, the prediction framework may be built using classification, clustering, knowledge transfer, data mining, or association approaches [125].

Homology-based approaches transfer the knowledge of one particular domain to another based on homology relations. Basically, if protein A and protein B in some organism interact, their homologs in another organism are predicted to interact too. More sophisticated homology-based prediction considers the phylogenetic profiling of the two proteins [70, 94]. If two proteins that are reported to interact in one organism appear or disappear in a large number of organism genomes simultaneously, then they are likely to function together, and consequently likely to interact when they are present in an organism [94]. Another approach that employs the homology relation is to transfer the interaction likeliness between two proteins to other organisms if the two proteins co-evolve similarly, which can be measured by comparing their phylogenetic trees [112].

The gene fusion-based prediction is another form of homology-based prediction. This approach is based on the observation that some genes whose proteins interact together have homologs in another organism that are fused into a single multi-functional gene [70, 92]. Thus, if protein A and protein B in an organism exist in another organism as a fused protein, then they are predicted to interact [70, 71, 72]. This method is limited by the presence of fused proteins. Additionally, a fusion event of two proteins does not necessarily imply their interaction although it implies a functional relationship.

A similar approach to gene fusion knowledge transfer is the gene neighborhood one. If the neighborhoods are conserved across multiple genomes, then their product proteins are likely to have a functional relation [22]. That functional relation is assumed to be the ability to interact between the two gene products. This method fails when the interacting proteins are located distantly [81].

Domain-pairing is an association prediction approach. In domain-pairing, if protein A interacts with protein B through their domains d_1 and d_2 , then protein C and D that carry the domains d_1 and d_2 are predicted to interact [109]. This method is also extended to other interaction signatures such as the presence of interacting motifs [111]. If not enough information exists for pairs of interacting domains or motifs, pairing is determined by looking for domains or motifs that appear in interacting protein pairs more frequently than expected, usually using Bayesian learning methods [55].

Availability of some protein three-dimensional structures encourages the prediction of interactions based on structural information. If protein *A* and protein *B* interact and they are structurally similar to protein *C* and *D*, *C* and *D* are predicted to interact as well [101, 126]. Structural alignment is performed for two query proteins against the library of protein structures to determine if they are structurally homologous to an interacting protein pair and thus likely to engage in a physical interaction [2, 4, 89]. A more detailed method additionally considers the specific residues that contribute to interactions at the interaction interfaces, namely hot spots [53]. In that method, the structural alignment is constrained by the presence of these conserved residues.

Another association-based method is gene expression profiling. In this method, if protein *A* and protein *B* belong to genes that fall into a common expression profiling cluster while the genes of proteins *C* and *D* fall into another cluster, then *A* and *B* are more likely to interact than *A* and *C* or *A* and *D* [43].

From a computational perspective, in the classification-based paradigm, a set of known PPIs along with protein information are used to train a model to discriminate between likely and unlikely interactions. The protein information that is used to model the interactions can be protein sequence, structure, and localization. Different classification techniques are used, such as SVMs [44, 107], artificial neural networks, Naïve Bayes, K-nearest neighbors, decision trees, and random forests [125].

Amino acid sequence is the most fundamental and readily available information about a protein, thus used by several frameworks to predict protein-protein interactions [12, 44, 107]. For protein length variability and non-relativeness of amino acid location on the sequence, local descriptors and sequence signatures are commonly used in such methods [83, 107, 111].

While methods that utilize protein sequences work on local signals, other methods work on global signals, including network-based methods. Common topological features in PPI networks have been used to support or exclude edges between protein pairs in the network [31, 93, 105]. Those features include node degree, degree centrality, closeness centrality, betweenness centrality, and network motifs [125].

2.2 Inter-species Prediction

Most methods in the field of PPI prediction work on interactions among proteins of a single species (intra-species prediction), mostly human, yeast, or bacteria. Interactions can happen between proteins of two organisms, mainly in a host-pathogen context, including virus-host, bacterium-host, and parasite-plant. With the importance of determining interactions of the proteins in understanding the host-pathogen relations, several studies designed methods for predicting protein-protein interactions between two species (inter-species prediction) [85]. Experimentally reported host-pathogen interactions are sparse [15, 38, 60], thus increasing the need for computational models to fill in this gap. Additionally, the experimental methods

of identifying such interactions can be time-consuming, expensive, or life-threatening. It has long been thought that predicting interactions between proteins of different organisms is more difficult than for the single organism case, as it involves proteins from two systems different in nature [85].

Predicting the interactions between proteins of a virus and its host is an inter-species prediction problem. The prediction of interactions of HIV proteins with human proteins has dominated the field of inter-species PPI prediction, for the abundance of the reported interactions that can be used for training a model and the abundance of knowledge about HIV proteins [33, 37, 106]. However, one method proposes a solution to the scarcity problem of host-pathogen protein interactions by applying transfer and multi-task learning techniques, extending the prediction to HTLV (Human T-lymphotropic virus) [75]. A few studies applied the methods proposed for HIV to other viruses, but the data sets are much smaller than in the HIV case, making the learning, evaluation, and generalization efficiency questionable [21, 36].

There are four interleaving components to address for the case of predicting PPIs between viruses and their hosts. We need a PPI sample set to learn from, descriptive features to represent the PPIs in a way that allows generalization, a learning approach that deals with the difficulties imposed by this type of prediction, and a procedure to generate the negative PPI examples.

In predicting PPIs for a virus with its host, either there are a sufficient number of known PPIs, too few, or none. In the first case, usually, a model is trained on the available PPIs to predict additional ones [33, 113]. As most viruses do not have many known PPIs, these single-virus centered prediction approaches are not universally applicable. This is why most methods for predicting the interactions between viral and host proteins are built around HIV, for the abundance of its reported interactions [38, 85]. Recently, transfer learning has been applied to learn from PPIs between a pathogen and its host and make predictions for another pathogen through a transfer function that maps knowledge between the two domains [58, 75]. However, transfer function design can be a challenge to build between viral families where members share no sequence similarities and may require prior knowledge of the two domains.

Machine learning-based prediction of PPIs need features extracted from PPI examples to learn. Different feature extraction methods exist where they exploit different protein aspects to extract valuable discriminative information that may play a role in the interaction process. In virus-host prediction, not all features utilized with the general PPI prediction can be used, due to intrinsic limitations of viral proteins, the fact that the interactions are taking place between proteins of different organisms, or shortage of knowledge.

In sequence-based feature extraction methods, features are extracted from the amino acid sequences of the protein pairs. Those methods basically have emerged from the local descriptor in [107]. The 3-mer frequency of each protein is used after reducing the 20 amino acid vocabulary into 7 amino acid groups of interaction-related physiochemical properties.

This method and its variations are successfully used in single virus-host PPI prediction methods [21, 33]. In [33], different lengths of the descriptor were compared in HIV-human prediction. In [21], a different vocabulary reduction scheme was proposed for prediction of the interactions between proteins of human and HCV (hepatitis C virus) and HPV (human papillomaviruses). As obtaining protein sequences for training is not problematic, using sequence-based methods enables generalization of the prediction framework. However, the overall accuracy might not be as good as when more sophisticated protein features are used.

Using protein structures to describe protein interactions is more accurate than other features [85]. However, obtaining structural information hinders using this method for many organisms, including viruses, for the cost or the difficulty of the process [126]. The main concept in structure-based approaches is that similar interaction interfaces are repeatedly used in different proteins in binding. Thus, these methods use structures of protein complexes as models where putative protein pairs are compared to those models to determine their potential for interaction [1, 69]. The structure-based approach is applied only in HIV-human PPI prediction for the availability of HIV protein structures [31, 130].

Interaction interface co-occurrence (or pairing) is used as features in interface-based PPI prediction applications. A feature vector of the considered domains or motif is labeled with their presence or absence in each protein [55, 102, 110]. Such information requires a higher level of study for each protein individuality and are not readily available for most of the proteins. Still, some frameworks used this method with HIV-human PPI prediction, for HIV proteins are well studied, but it has not been used for any other virus. In those virus-host applications, domain-based approaches transfer known interactions from one protein to another based on sequence homology, structure similarity, or sharing of interaction interfaces [37, 106, 130].

In the homology-based feature extraction approach, knowledge about an interacting pair is transferred to their homologs based on the assumption of interaction conservation among homologs [85]. Although this approach is applied in host-pathogen PPI prediction [129], the sequence dissimilarities among viral families hamper the generalization of this knowledge-based approaches to viruses from different families. Additionally, viral fine mutation events are driven by the force of surviving the host defense system, and the fine differences among virus strains are linked to their functionality and interaction abilities. The homology-based knowledge transfer among viral relatives will overlook those mutations.

Other protein features are used in general PPI prediction applications like similarity of GO terms, gene-fusion approaches, and gene clustering [85]. However, their transfer from intra-species PPI prediction to the inter-species case is not applicable, for their concepts are single-organism centered. For example, gene fusion works on comparative genomics among several organisms to identify whether two query proteins appear as fused in some organisms. If a protein belongs to a host and another to a virus, what are the chances of finding them fused in a single organism? Although GO was used with HIV-human prediction [113], it was considered on the human protein side, not the viral one.

Most PPI prediction frameworks in inter-species context use machine learning. However, other approaches exist, like transfer learning, data mining, association, and statistical learning [85]. These approaches are discussed in Section 2.1; I here spotlight the ones used in inter-species prediction.

Different machine learning models are used, but SVMs have proven their excellence in inter-species PPI prediction [31]. Support vector machines are based on two concepts: the kernel and the margin maximization [9, 17, 19]. Kernels transform the feature vectors from the input space into a higher dimensional feature space where linear separability is possible. Additionally, kernels limit the curse of dimensionality by computing a pairwise distance measure between the data points. In that higher dimensional space, SVMs choose the decision hyperplane that maximizes the margin between the two classes.

The data mining approaches model true PPIs and measure a putative PPI similarity to them [86]. These approaches bypass the problem of generating negative PPI examples by modeling the true interactions. For example, conformal prediction is used in a recent study [86] where the conformance of a putative PPI is evaluated against the known PPIs and classified as positive if its prediction confidence is above a predefined threshold. In transfer learning, knowledge of PPIs is selectively transferred from a rich domain of PPIs to the target one for which prediction is needed using commonalities between the two domains [58, 75]. I anticipate that data mining and transfer learning approaches are not adequate for novel viral PPI prediction, mainly because virus families carry no sequence similarities among themselves sufficient to make data mining or transfer learning from available PPIs of other viruses possible.

In this thesis, I address the problem of predicting interactions between proteins of a host and a virus that have no known interactions to learn from, which has never been addressed before. Such a virus can be a newly discovered virus (novel) or a known virus with no known interactions with its host, which can be considered novel as well in terms of known PPIs. Except for HIV, all viruses have few or no known interactions with their hosts, hindering the application of the single virus-host prediction methods to those viruses. Thus, viruses with few known interactions can be considered novel from that perspective as well.

In Chapter 3, I propose DeNovo, a framework that addresses the problem of predicting the interactions between virus and host proteins when no interactions between them are known. DeNovo uses proteins sequences and a pool of interactions from different viruses with their host to predict putative interactions of a new viral protein. In Chapter 4, I propose DeNovo-Human, a framework that predicts interactions between a viral protein and human proteins when some of the viral protein interactions are known. Chapter 5 discusses a case study where I use DeNovo and DeNovo-Human to answer critical questions about Zika virus in a demonstration of their capabilities and application.

2.3 Negative Example Selection

A PPI prediction model must learn the difference between the positive and negative classes, i.e., between interacting and non-interacting pairs, to make a future prediction on new pairs. There exist some learning PPI prediction methods that operate on the positive PPIs only, but the majority of the prediction models function on a classification basis where positive and negative examples are needed to train the models. I call a pair of proteins that are not likely to interact a “negative PPI”.

To the best of my knowledge, except for the database Negaotome [11], there is no gold standard of protein pairs that are unlikely to engage in physical interactions. The Negatome database has a limited set of negative PPIs coming from different organisms. I examined the taxonomic lineage of the proteins in the largest set of negative PPIs in Negatome (the least constrained set). The set contains 6,532 negative pairs among 3,375 proteins, 1,241 of which are human proteins and 516 are mouse proteins. This data set might be of help to building PPI models for human proteins. However, it does not include any viral proteins.

As there have not been sufficient gold standard negative interactions to train interaction prediction models, most machine learning-based approaches generate sets of negative PPIs from the available positive PPI sets. I call the process of generating a pair of proteins that are not likely to interact from a set of positive PPIs “negative sampling”.

Negative PPI examples are typically generated randomly [7, 107]. Two random proteins from the positive data set are paired together if they do not appear in that set as interacting. Random negative sampling is expected to generate a considerable number of false negative examples. However, this random negative pairing method is widely accepted in intra-species PPI prediction, for it is believed to be unbiased, especially when used with some machine learning models (like support vector machines) that can tolerate high noise in the training data [7, 90]. The key idea behind accepting this method is that the negative PPI space is expected to be much larger than the positive space, making the expected ratio of false to true negative examples generated small.

Another PPI negative sampling method that received attention in the field is the exclusive co-localization negative sampling method [73, 75]. In this negative sampling method, protein A is paired with protein B to form a negative example if and only if protein B is not co-localized with protein A . The idea behind this method is two proteins that are not co-localized are prevented from participating in a physical interaction. As understood from [8], selecting the negative training examples using this method converts the problem to localization classification, which is much easier to accomplish from protein sequence than PPI prediction.

A third method uses reported positive interactions to generate negative interactions in a different way than sampling [77]. In that study, the set of known positive PPIs is classified using single-class SVMs into strong positive and weak positive. The weak positive PPIs are

considered negative PPIs. Then a 2-class SVM model is trained on the 1-class SVM positive and negative PPIs.

A recent method that works on top of sampling is the hierarchical clustering method [117]. In that method, the authors want to consider millions of negative PPIs in training without biasing the learning model. By clustering the negative examples using *k-means* in an iterative fashion, they created representative negative examples that are then used for training the models.

In reality, the number of protein pairs that may interact with a set of proteins is an extremely small fraction of all pair combinations. However, this exact ratio is unknown. In the PPI prediction application, the ratio of 1:1 is always used, for some machine learning models become biased when the ratio is unbalanced [90]. However, some methods take into consideration this natural unbalance of positive and negative examples. In [58, 75, 113], a ratio of 1:100 of positive to negative interactions was employed, while [33] reported the performance at different ratios. Those non-1:1 ratios are used to account for the expected unbalance, but the real ratios have never been estimated, to the best of my knowledge.

In Chapter 3, I propose a negative sampling method specific for virus proteins. That method employs observations and knowledge from molecular virology to reduce the false negatives when randomly sampling the negative examples from the set of unreported interactions. In Chapter 4, I use my developed framework, DeNovo-Human, to predict the network of interactions between many viral proteins and human proteins. Then I use that network to estimate the positive to negative ratio in the virus-human PPI network.

Chapter 3

DeNovo

3.1 Introduction

This study addresses the generalized problem of predicting interactions between a viral protein and a host protein without the need to know any of their interactions in advance. The key idea is to learn interaction features for the host proteins rather than for specific viral proteins. The model depends only on protein sequences to extract features and generate unbiased negative examples to ensure generalization. Data is partitioned in a way that simulates the discovery of a new virus to reasonably evaluate the performance. I further make three hypotheses on what attributes potentially guide the learning process and use my framework to test them to link biological and computational aspects of the prediction process. I name the proposed solution DeNovo as an analogy to *de novo* protein structure prediction, for it makes predictions without prior knowledge of any interactions between the two organisms to learn from; see Figure 3.1. For lack of data, I used interactions between virus and human in my experiments.

DeNovo demonstrates the feasibility of predicting interactions of a viral protein with human proteins by learning from the available interactions of other viruses that also infect human, even with little or no global sequence similarity to the viral proteins in training. The framework achieves accuracy comparable to the best reported for intra-species PPI prediction (using sequence information) despite the challenging nature of the generalized problem. I additionally demonstrate that DeNovo generalizes to other pathogens.

3.1.1 Problem Definitions

At the heart of DeNovo is the usage of known PPIs between a set of viruses and a shared host to predict PPIs of the host proteins with proteins of another virus that is not part of the available virus set. The novel virus PPI prediction problem can be formalized as:

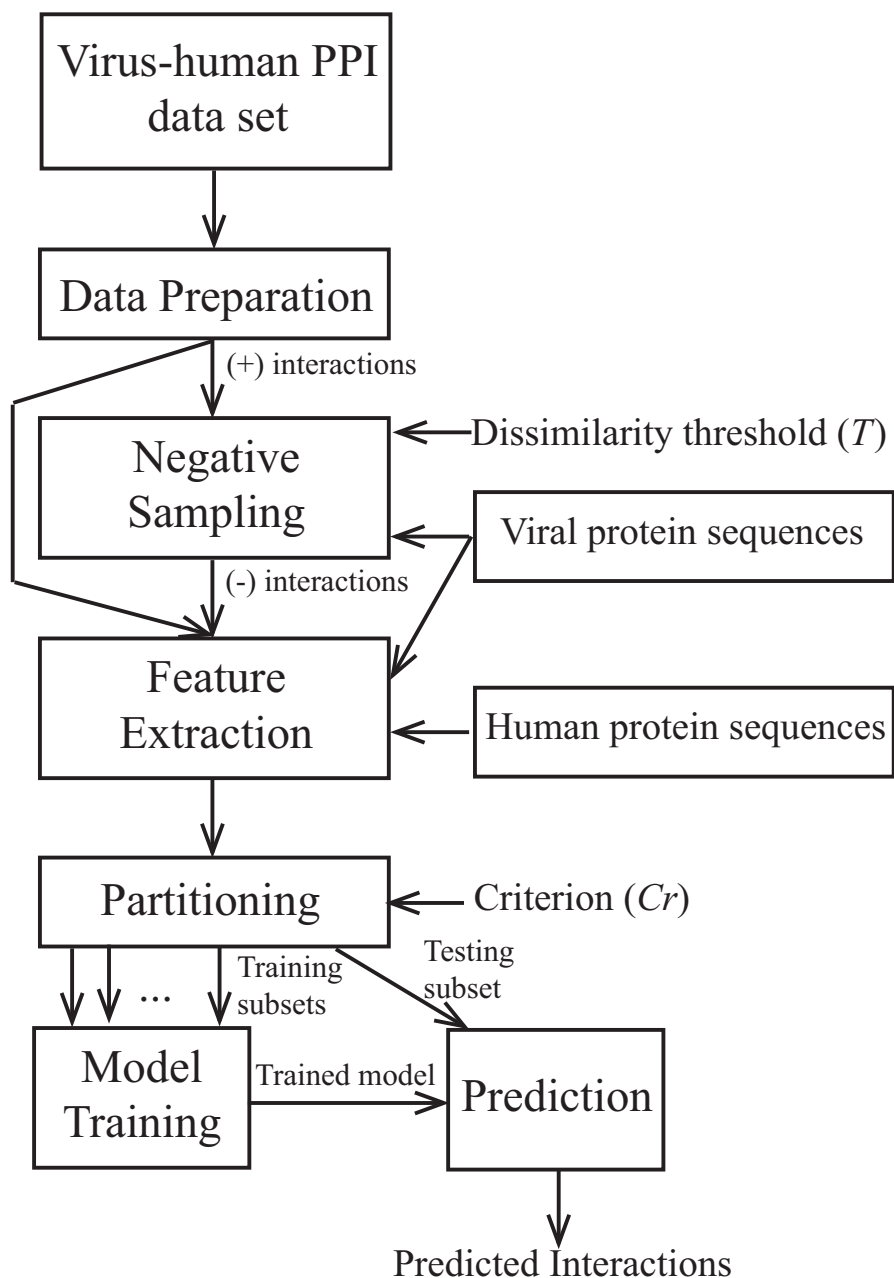


Figure 3.1: DeNovo workflow. Virus-human PPIs are first cleaned, negative interactions are generated using the proposed dissimilarity-based method, and features are extracted from the protein sequences of the positive and negative interaction pairs. The feature vectors of the interactions are partitioned into multiple pairs of training and testing subsets based on a criterion Cr . A machine learning model is trained on the feature vectors from a training subset and tested on the corresponding testing subset in a cross validation fashion.

Given: a list of viruses $v_1, v_2, \dots, v_n, v_{n+1}$ infecting a host h , a set of proteins H for h , a set of proteins V for $v_1, v_2, \dots, v_n$, a set of positive interactions between members of H and V , and a set of proteins V_{n+1} for v_{n+1} .

Find: all likely interactions between members of H and V_{n+1} .

I call the process of generating negative examples from the positive ones “negative sampling”. In the context of PPI prediction, a negative interaction (example) resulting from this process is a pair of proteins unlikely to interact. For the special case of virus-host PPIs between multiple viruses and a shared host, the negative sampling problem can be defined as:

Given: a set of host proteins H , a set of viral proteins V , and a list of positive interactions between members of H and V .

Generate: a list of negative (unlikely) interactions between members of H and V .

Machine learning applications typically split the available data into training and testing sets. Here, I use “data partitioning” to refer to dividing the data set based on some criterion into multiple subsets, each of which can serve as a testing set while the remaining subsets are grouped into a corresponding training set. This problem can be defined in the context of virus-host PPI prediction as:

Given: a set of host proteins H , a set of viral proteins V , two sets of positive and negative interactions, PPI^+ and PPI^- , between members of H and V , and a criterion Cr on sets of interactions.

Find: a partitioning of PPI^+ and PPI^- into subsets, each of which satisfies Cr .

3.1.2 Hypotheses

Features Learning Hypothesis. I hypothesize that if I use a large number of interactions between different viruses and host proteins to train a classification model, it can learn to classify interactions with respect to the host proteins, rather than specific viral proteins. This hypothesis is based on the observation that viral proteins from the same family share host protein targets more often [74]. I anticipate that the classifier will extract features associated with the host proteins being interacting or not interacting. In such a case, PPI prediction would be possible for a foreign viral protein as the model will evaluate whether it is likely or unlikely to interact with each of the human proteins it was trained on.

Shared Interaction Partner Hypothesis. I hypothesize that two viral proteins with high sequence similarity can theoretically interact with a large number of similar host proteins. This hypothesis is based on two observations. First, viral protein interaction interfaces are likely to be different and more flexible than those of human proteins [114]. Second, their interactions with the host proteins are extensively mediated with viral short linear motifs (SLiMs), which give them the flexibility to interact with a larger number of host proteins

[24]. Using negative random sampling in predicting for the interactions between viral and host proteins will possibly lead to a large number of false negative examples.

Sequence Similarity Hypothesis. I hypothesize that strong sequence similarity between a tested viral protein and some viral proteins in the training set enhances the classification decision accuracy. Some interaction-related features may be conserved across the related viral proteins and consequently reflected in their sequences and the corresponding feature vectors. A novel virus not only has no known interactions, but it also may have no strong sequence similarities with any viral protein in the training set. I thus anticipate that the accuracy in the case of predicting PPIs of a novel virus can be lower than when there are some related viral proteins in the training set.

3.2 Methods

DeNovo divides the virus-human PPIs data set into training and testing pairs in a way that simulates a novel virus prediction situation. DeNovo learns only from the amino acid sequence of the interacting pairs. To overcome the low prediction accuracy when random negative examples are used in training, a sequence-based negative sampling method is proposed. Its performance is assessed using other data sets and compared with other methods to demonstrate generalization and robustness.

3.2.1 Data Set and Preparation

Virus-human PPIs were collected from VirusMentha (accessed June 2014) [15]. Each viral protein is given by its UniProtKB identifier and NCBI taxon identifiers of its corresponding virus and viral subfamily [122]. The data set contains 5,753 unique interactions between 2,357 human proteins and 453 viral proteins, covering 173 different virus species in 25 subfamilies. Sequences of the human and viral proteins were retrieved from UniProt [116]. See Figure 3.2.

The taxonomy lineage of each virus taxon identifier (associated with a viral protein) was retrieved to fix errors in subfamily taxon identifiers (assigned by VirusMentha to each viral protein) by consulting the ICTV virus taxonomy (2013 release) [56]; see the taxonomy tree of the viruses used in the study (Figure 3.2). Interactions were filtered down to 5,445 PPIs between 2340 human proteins and 445 viral proteins, covering 172 viral species in 28 subfamilies. A PPI was eliminated if one of its proteins has no corresponding reviewed sequence in UniProt [116], if the viral protein belongs to a virus whose taxon identifier is not annotated as hosted by human in UniProt, or if the viral protein belongs to a virus that lies on a branch (on the taxonomy tree) with limited representation in the data set (an arbitrary threshold of 10 PPIs was used).

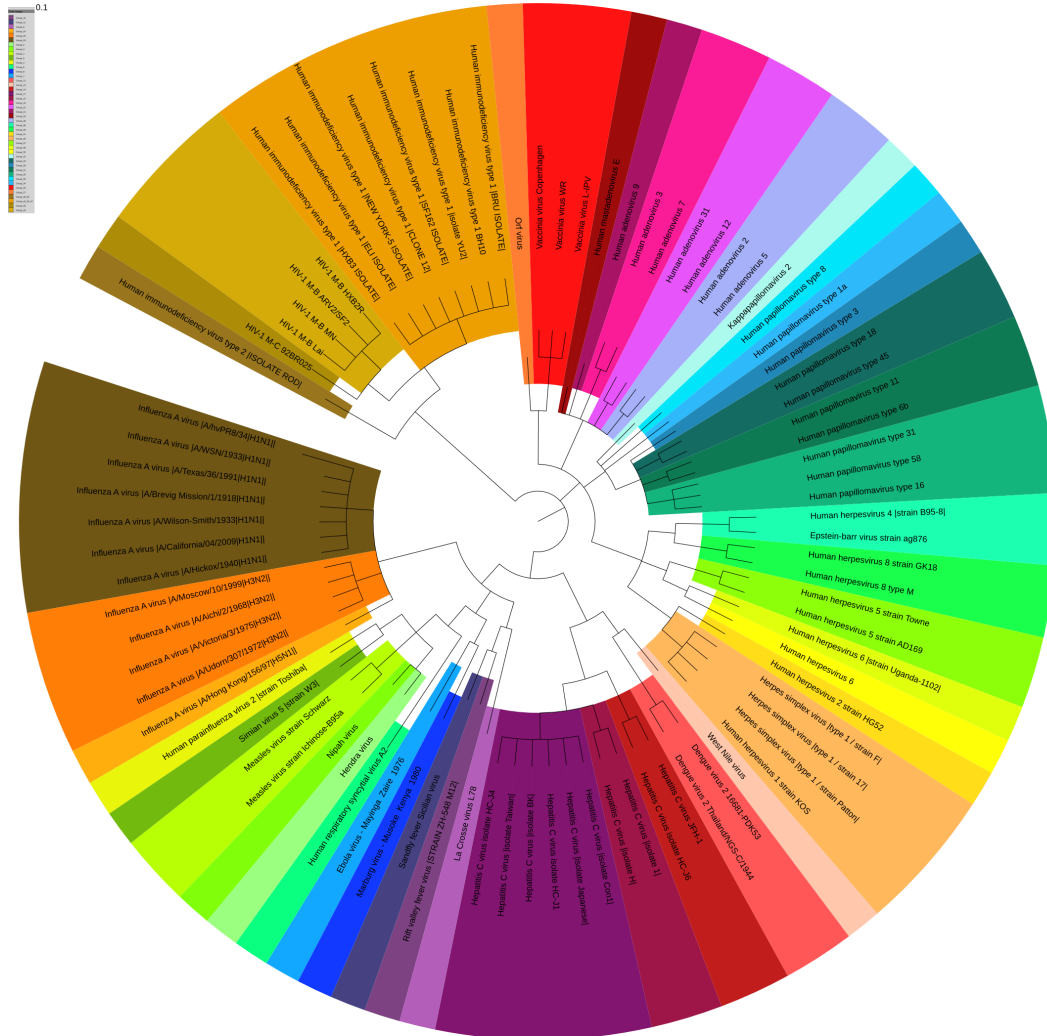


Figure 3.2: Taxonomy tree of the viruses with interactions in VirusMentha.

3.2.2 Dissimilarity-Based Negative Sampling

My negative sampling method aims at reducing the expected large number of false negative examples generated by random sampling. According to my hypotheses, if two viral proteins

are similar in sequence, a human protein that interacts with one of them cannot be paired with the other as a negative example. The DISSIMILARITY-RANDOM-SAMPLING algorithm below calculates all-versus-all global alignment bit-scores of the viral proteins. Bit-scores are then normalized, and their complements are used as the dissimilarity distances between the pairs of viral proteins. After excluding the unlikely negative examples based on a dissimilarity threshold T , random sampling is performed over the remaining negative interactions.

The DISSIMILARITY-RANDOM-SAMPLING algorithm, Figure 3.3, takes as input a viral protein x , a set of human proteins $H(x)$ interacting with x , and a list L of $(y, H(y))$ pairs for different viral proteins y and their corresponding human interacting partner sets $H(y)$. An input parameter T serves as a cutoff value for the sequence dissimilarity between the viral proteins when selecting the less likely interacting human partners. The algorithm returns a set of n human proteins $N(x)$ to be paired with x as negative interactions.

DISSIMILARITY-RANDOM-SAMPLING first calculates, for each viral protein in V , the *BitScore* of globally aligning that protein with the viral protein x . Global alignment (*GlobalAlign*) was performed using the Needleman-Wunsch algorithm with the BLOSUM30 matrix to capture distant similarities. As V may contain some outliers (viral proteins with small bit score, and hence large dissimilarity, to all other viral proteins), *RemoveOutliers* identifies these outliers and returns a reduced list of viral proteins V' . The bit scores are then normalized between 0 and 1 (*NormBitScore*) over all the bit scores of V' with x (*BitScore'*). $\min(\text{BitScore}')$ and $\max(\text{BitScore}')$ are the minimum and maximum values of *BitScore'*, respectively. One minus the normalized bit score for a viral protein y with x is used as a sequence dissimilarity distance between them (*Distance*(x, y)). If *Distance*(x, y) is close to zero, then y is highly similar in sequence to x compared to the other viral proteins in V .

If this dissimilarity distance is larger than the dissimilarity threshold T , then the human proteins interacting with y ($H(y)$) can be added to the set of unlikely interaction partners with x ($N(x)$). At $T = 0$, there is no dissimilarity enforced, and the dissimilarity-based method is equivalent to the non-constraint random sampling. Finally, n human proteins are randomly selected from $N(x)$ by *RandomPick*.

3.2.3 Data Partitioning

In the *Partial Blindness* scheme, *Cr1*, interactions associated with leaf nodes under the same immediate ancestor in the viral taxonomy tree are grouped together into a single subset. This partitioning makes interactions from very related viruses (and consequently related viral proteins) fall into the same testing subset, but still allows some more distantly related viral proteins in the same virus family to be utilized in training (as they fall in some other subsets), hence the name partial blindness. See Figure 3.4 for an example. The testing, in this case, will be equivalent to predicting for a novel virus of the same rank as the ancestor of the testing subset.

Algorithm 1 DISSIMILARITY-RANDOM-SAMPLING

Input: $x, n, T, V, H(V), L$ **Output:** $N(x)$

```

  // Selecting  $n$  negative interaction partners for viral protein
  //  $x$  from interaction partners of other viral proteins that
  // are more than  $T$  dissimilar in sequence to  $x$ .
  //  $V$  is the set of all viral proteins in the data set from which
  // negative samples are drawn.
  //  $L$  is a list of viral proteins in  $V$  and their corresponding human
  // interacting partners  $H(V)$ .
1:  $N(x) = \emptyset$ 
2: for  $y \in V$  do
3:    $BitScore(y) = GlobalAlign(x, y)$ 
4: end for
5:  $[V', BitScore'] = RemoveOutliers(V, BitScore)$ 
6: for  $y \in V'$  do
7:    $NormBitScore(y) = \frac{BitScore'(y) - \min(BitScore')}{\max(BitScore') - \min(BitScore')}$ 
8:    $Distance(x, y) = 1 - NormBitScore(y)$ 
9:   if  $Distance(x, y) > T$  then
10:     $N(x) = N(x) \cup H(y)$ 
11:   end if
12: end for
13:  $N(x) = N(x) \setminus H(x)$ 
14:  $N(x) = RandomPick(N(x), n)$ 
15: return  $N(x)$ 

```

Figure 3.3: DISSIMILARITY-RANDOM-SAMPLING algorithm: selecting less likely interactions as negative PPI for virus-host PPI prediction applications.

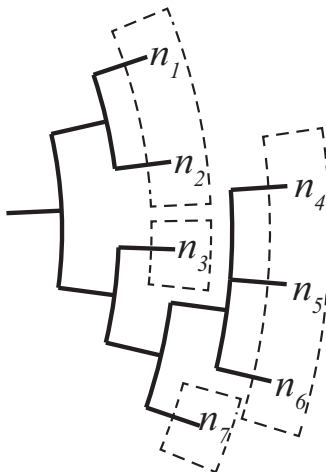


Figure 3.4: An example of partial blindness partitioning. Each leaf node n_x in the taxonomic tree carries interactions of a virus v_x . Nodes at the same level belong to viruses of the same taxonomic rank (e.g. species or strain). Interactions of nodes sharing an immediate ancestor are grouped together into a single subset, denoted by a dashed contour.

I want to examine how the framework performs in the case of a novel virus that not only has no known interactions, but also its proteins have no strong sequence similarities with any viral protein in the training set (Sequence Similarity Hypothesis, Section 3.1.2). To test this case, we need to make the viral proteins in the testing set S completely dissimilar in sequence to all viral proteins in the training set R . I thus developed the *Complete Blindness* scheme, *Cr2*, to describe this requirement. In this scheme, interactions of viral proteins within the same viral family are grouped together into a single subset. Members of a single viral family share some similarity in sequence among themselves and no similarities to the other viral families (according to the ICTV taxonomy).

3.2.4 Learning Model

I use a feature extraction scheme developed for intra-species PPI prediction and successfully adapted to the inter-species case [21, 33, 107]. It first clusters the 20 amino acids based on similarities of physiochemical properties known to drive most PPIs (dipoles and volumes of side chains) [107]. The amino acid clusters are $\{A, V, G\}$, $\{I, L, F, P\}$, $\{Y, M, T, S\}$, $\{H, N, Q, W\}$, $\{R, K\}$, $\{D, E\}$, and $\{C\}$. The amino acid sequence of each protein is mapped to the corresponding cluster numbers. For example, *ACDHNYA* is mapped to the vector (1, 7, 6, 4, 4, 3, 1), for *A* is in Cluster 1, *C* in Cluster 7, and so on. The frequency of each possible 3-mer is calculated in each mapped protein, generating a feature vector of length $7^3 = 343$ that is then normalized for each protein independently. If $F = (f_1, f_2, \dots, f_{343})$ is a

protein feature vector, then each normalized element f'_i is calculated as:

$$f'_i = \frac{f_i - \min(F)}{\max(F) - \min(F)} \quad (3.1)$$

The two normalized vectors of an interacting (or negative) pair are concatenated into a single feature vector representing the interaction, with the viral protein vector coming first.

For each of the (R, S) training-testing pairs, the values of each feature (f'_i) are independently standardized across all feature vectors in the training set. The mean and variance for each feature (as computed to standardize the training set) are used to standardize the values of the corresponding feature in the testing set. At this stage, each subset (R or S) should be represented by a set of feature vectors for the positive and negative interaction along with their labels.

I used the radial basis function kernel $K(u, v) = \exp(-\gamma * |u - v|^2)$ and the SVM implementation from LIBSVM version 3.18 [17]. To pick the optimal values for the model parameters C (error penalty) and γ (the kernel parameter), a grid search over C and γ was performed with nested 5-fold cross-validation. Five exponentially-spaced values for each parameter were used; C in the range $[10^{-3}, 10^1]$ and γ in the range $[10^{-4}, 10^0]$, as suggested by best practice [17]. The search was conducted at each value of the dissimilarity threshold T in the range $[0, 1]$ with a 0.1 step by training and testing SVM models on each training-testing pair for each (C, γ, T) combination. The resultant optimal parameter values are $C^* = 10$ and $\gamma^* = 10^{-3}$. This pair of values for C^* and γ^* are already known to operate the LibSVM implementation of SVMs optimally, thus, the grid search does not have an effect on the hyperparameter value choice at the end.

The data set was divided into 49 and ten subsets according to Criteria $Cr1$ and $Cr2$, respectively. The i^{th} subset was used for testing at a time against a model trained on the interactions from all the remaining subsets, forming the i^{th} (R, S) training-testing pair. A negative to positive examples ratio of 1 : 1 is maintained throughout the study, as it is the common practice in PPI prediction studies [91]. Rounds of training and testing were performed over the different pairs with different dissimilarity thresholds T . The performance of the system was averaged over the entire data set for each of the partitioning schemes. Thus, no data are lost in testing while the performance evaluation is consistent with the goal of the study.

The performance was evaluated using standard accuracy, sensitivity, and specificity for each (R, S) pair [85]. I also report the support vector ratio for each model and prediction confidence. To quantify the overall performance across all the different (R, S) pairs, averages of the above measures were calculated (for each parameter value set), weighted by the sample size of the testing unit S in each pair.

3.2.5 Supplementary Studies Data

I used VirusMentha and other data sets to perform the pilot and generalization studies in Section 3.2.6. For ST1, a new data set of virus-human PPIs was collected from HPIDB (accessed September 21st, 2015). The data were originally generated in a separate study [34]. I used PPIs from three bacteria with human: *Bacillus anthracis*, *Yersinia pestis*, and *Francisella tularensis*; call them B1, B2, and B3 respectively. 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 3,057 PPIs, B2 has 402, and B3 has 1,346. Approximately, each bacteria shares half of its human interaction partners with at least one of the other two bacteria: B1, B2, and B3 interact uniquely with 426, 1197, and 861 human proteins, and share 572, 911, and 887 partners with the other two bacteria, respectively.

The host in ST2 is *Arabidopsis thaliana*. In the training set, the interactions are collected from the PPIRA database of *A. thaliana* and the bacterium *Ralstonia solanacearum* (Accessed September 26, 2015). There are 3,074 PPIs between 119 of the bacteria proteins and 1,442 of the host proteins. The testing data set consists of PPIs between the same host and other pathogens including some bacteria different from *R. solanacearum*. A set of 543 PPIs were collected from HPIDB (accessed September 24, 2015) of 149 unique pathogen proteins. The testing set contains 127 PPIs that belong to bacteria, 384 to fungi, and 30 to viruses.

All VirusMentha PPIs used in the main study of DeNovo are used in ST3 for training. For testing, 8,857 bacteria-human PPIs from HPIDB were employed. A set of 2,765 interactions in this testing set has human partners that appear in VirusMentha. For ST4 and ST5, all of VirusMentha PPIs serve as positive training examples, and the negative examples are generated with T^* (Section 3.4.1). The testing set was obtained from an external data set, HPIDB, with 1,612 PPIs of foreign human proteins for ST4, and 1,830 PPIs of the known human proteins for ST5. The viral proteins in both cases are foreign to the training set.

For study ST6, I downloaded the SLiMs known in viruses from the ELM database, accessed Oct. 5, 2015. Using protein UniProt IDs, I intersected viral proteins carrying these ELMs with the viral proteins in VirusMentha (the data set used in the main study).

3.2.6 Pilot and Generalization Studies

I conducted a set of pilot studies to assess DeNovo generalization and robustness. Each study is denoted STx, where x is the study number, for ease of reference. See Table 3.1.

The first study, ST1, examines how well DeNovo generalizes to predict bacteria-human PPIs for a novel bacterium. It also compares DeNovo performance to a recent transfer learning approach [58]. In ST1, features were extracted from the positive examples of the three bacteria-human PPI sets (Section 3.2.5). My dissimilarity-based negative sampling scheme was used to generate the negative examples at T^* with a positive to negative ratio of 1:10

No.	Goal	Summary
ST1	Testing DeNovo generalization to bacteria.	Using DeNovo with partitioning criterion $Cr2$ over the interactions of the proteins of three bacteria with human proteins.
ST2	Testing DeNovo generalization to different pathogens.	Training on bacteria-host PPIs and testing on PPIs of other pathogens.
ST3	Testing DeNovo generalization to different pathogens.	Training on virus-human PPIs and testing on bacteria-human PPIs.
ST4	Prediction scenario for foreign viral and human proteins.	Testing viral and human protein pairs not in training.
ST5	Prediction scenario for foreign viral proteins.	Testing viral proteins not in training, paired with human proteins in training.
ST6	Assessing using SLiM-based features with DeNovo.	Masking known SLiMs in viral proteins before extracting the features.
ST7	Testing 1-class SVMs positive PPIs modeling.	Evaluating sensitivity with $Cr1$ and $Cr2$.
ST8	Testing 1-class SVMs negative sampling.	Dividing the data set using 1-class SVMs into positive and negative PPIs, then using DeNovo.

Table 3.1: Summary of the pilot and generalization studies in Section 3.2.6.

for each bacterial protein. An SVM model in each training-testing round was trained and tested with C^* and γ^* . I followed the same testing scheme as in the family partitioning, where each of these different bacteria was set aside for testing while the interactions from the other two bacteria were used for training.

ST2 trained DeNovo on interactions of *A. thaliana* proteins with proteins of a specific bacterium and tested the model on PPIs from other pathogens with *A. thaliana*. Features were extracted from the positive interactions of the training set. All-versus-all sequence dissimilarity distances were calculated for the bacterial proteins to generate the negative examples with a 1:1 ratio of positives to negatives. An SVM model was trained on these examples. Features were extracted for the testing set as normal. The dissimilarity distances were measured for each protein of the pathogens with the bacterial proteins in training to generate the negative examples for testing.

DeNovo was further trained on all VirusMentha virus-human PPIs and tested on bacteria-human PPIs in ST3. Features of the VirusMentha PPIs were extracted, and negative examples generated at T^* using a ratio of 1:1 of positives to negatives. An SVM model was trained on all of the PPIs of VirusMentha at once with C^* and γ^* and tested on the bacteria-human PPIs positive examples after their features were extracted. No negative examples were created, for our concern here is how much of the true interactions can be captured from different

pathogen interactions.

DeNovo was tested on virus-human PPIs for foreign viral and human proteins (not in the training set) in ST4, and on PPIs of foreign viral proteins with known human proteins in ST5. ST4 and ST5 work as application scenarios on how to use DeNovo to estimate the probability of a list of putative PPIs of a novel virus (ST4) and to generate a list of interacting and non-interacting human partners for that virus (ST5). In ST5, the dissimilarity distance between the novel viral proteins and those of VirusMentha were calculated to generate negative examples for the testing set with T^* and a 1:1 ratio. The trained model was tested on the positive and negative interactions.

Assessing how adding short linear motif (SLiM) features may improve DeNovo accuracy, in ST6, I masked the SLiMs known in the viral proteins, grouped the PPIs of these proteins into a testing data set, trained DeNovo on the remaining PPIs, and compared the prediction accuracy when the SLiMs were masked and not masked. I divided VirusMentha into two sets of interactions: the ELM-set that contains PPIs of viral proteins that have some of the ELMs, and the non-ELM set that contains the remaining PPIs. The non-ELM set works as a training set. I then generated the negative examples for the training set and the ELM-set using the dissimilarity method of DeNovo with a 1:1 ratio of positive to negatives.

A masked set was generated from the ELM-set, in ST6, by masking the known SLiMs in each viral protein so that these SLiMs are not considered in feature extraction. Features were extracted from the training, masked testing, and non-masked testing sets. An SVM model with $C^* = 10$ and $\gamma = 10^{-3}$ was trained on the training set. The trained model was tested against the masked and non-masked testing sets. The confidence of prediction in both cases was retrieved and then used to calculate standard error.

To assess how data mining techniques that learn from positive PPIs alone can be affected by the viral dissimilarity property among families, I conducted ST7. I used a one-class classifier to model the true PPIs in the two grouping criteria and test the accuracy of prediction as in the main study. A one-class SVM model was trained on the positive interactions of the viral groups in Criteria *Cr1* and *Cr2* in a leave-one-out fashion. The trained model in each case was then used to classify the positive examples in the remaining group into positive and negatives. The accuracy was measured as how many of the examples were classified as positives.

ST8 is a three-fold study where I compared my negative sampling method to one recently proposed [77], which uses one-class SVMs to separate true PPIs into positives and negatives. I used that method on the ten viral family sets, and further tested the models on the negative samples of these families (generated by the negative sampling method I proposed). I additionally compared the result of splitting the PPIs into positive and negatives against their original confidence score.

In the first part of ST8, I used the family grouping (Criterion *Cr2* in the main study) to assess how exceptional accuracy this negative sampling method can reach. In each one of the

ten families of PPI sets, I train an SVM model on the PPIs of that family and then used the same model to give a score for each PPI. The 0.5 threshold was used as a balancing point. A PPI with a higher score is considered a positive example and one below is considered a negative example. I used the leave-one-out testing strategy on the families, by training a two-class SVM model on the examples from nine families at a time and test on the examples from the remaining family. The accuracy measures were averaged over the ten testing sets.

In the second part of ST8, I tested the negative examples that my dissimilarity-based method generates for each family on the models generated above. Each negative set from a family was tested on the model trained on the other families as usual.

In the third part of ST8, I retrieved the scores of confidence in positive PPIs from the VirusMentha data set for the total of 5,445 PPIs used in the main study. Only 301 PPIs have scores above 0.5. I used a one-class SVM to learn the structure or model of this data set. I then used the same model to generate scores of how likely each PPI example from the training set is to belong to the major component of similar interactions as modeled by the one-class SVM. I used the threshold of 0.5 again for comparison with the real score in the data set.

3.3 Results

The partial and full simulation of predicting PPIs of a novel virus (Criteria *Cr1* and *Cr2*) give 49 and ten viral subsets respectively. See the taxonomy tree in Figure 3.2 with the 49 subsets color-coded. The ten family names are listed in Table 3.2. Influenza A virus PPIs are grouped into three subsets (H1N1, H3N2, and H5N1), solely representing *Orthomyxovirus* in family grouping. Available PPIs from Hepatitis C virus (HCV) genotypes fall into three subsets in criterion *Cr1*, which contribute along with two other viruses into the viral family group of *Flaviviridae*. PPIs of *Herpesviridae* fall into seven subsets in criterion *Cr1*, each of which roughly corresponds to one of the human herpesvirus (HHV) types. Similarly, human viral types of *Adenoviridae* and *Papillomaviridae* fall into five and seven subsets, respectively.

When tested on the partial and full simulation of novel virus prediction (partitioning criteria *Cr1* and *Cr2* at $T = 0$), the random sampling accuracy achieved 60% and 48%, respectively. With the same features and SVM models, random sampling in intra-species and single virus-host PPI prediction achieved a maximum of 86% and 81.6%, respectively [21, 107]. Testing with the 49 (R, S) pairs resulting from the partially blind partitioning, the accuracy raised to 76.4% at $T^* = 0.8$, and eventually increased to 86% with $T = 0.9$. Accuracy, sensitivity, and specificity were increased when T^* was used with 16%, 16.8%, and 17.6%, respectively, over those achieved with random sampling. See Figures 3.5 and 3.7.

For ST1, in average, prediction accuracy was about 97%. Sensitivity, specificity, accuracy were 94%, 97.2%, and 96.42% for B1; 94.8%, 98.3%, and 97.47% for B2; and 94.9%, 98.3%,

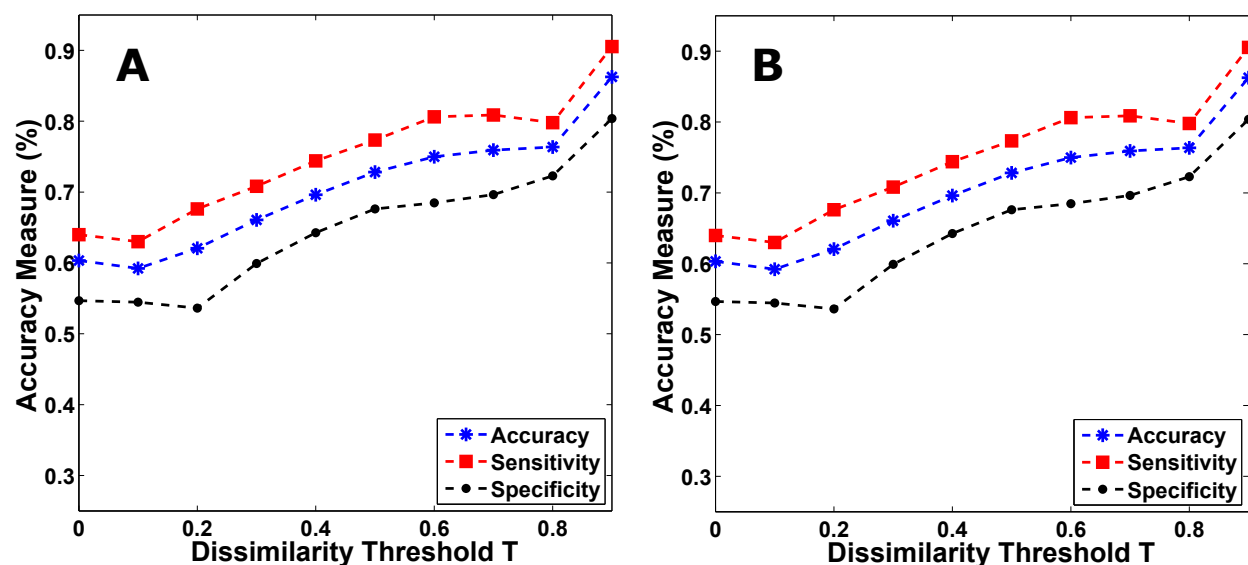


Figure 3.5: Accuracy, sensitivity, and specificity measured at different sequence dissimilarity threshold values, testing with (A) partial blindness, and (B) full blindness grouping criteria. Dotted lines are for visualization purposes.

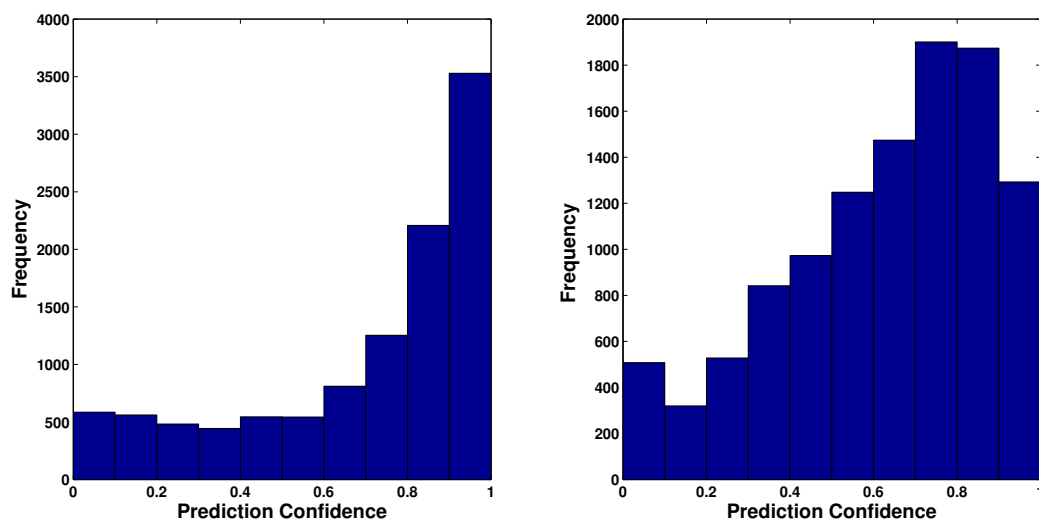


Figure 3.6: Confidence of prediction with (A) partial blindness grouping and (B) full blindness grouping, using the model optimal parameters (C^* , γ^* , and T^*).

and 97.32% for B3. In ST2, sensitivity achieved was 84.6% while specificity was 65%. In ST3, DeNovo could recognize 67.7% of the interactions with known human proteins and 69.9% of the interactions with foreign human proteins. The overall sensitivity was 69.2%. Recognition of the positive interactions in both studies ST4 and ST5 was 94%. The specificity in study

No.	Name	Sample size
1	<i>Paramyxoviridae</i>	762
2	<i>Filoviridae</i>	114
3	<i>Bunyaviridae</i>	159
4	<i>Flaviviridae</i>	291
5	<i>Adenoviridae</i>	88
6	<i>Orthomyxoviridae</i>	664
7	<i>Chordopoxviridae</i>	237
8	<i>Papillomaviridae</i>	245
9	<i>Herpesviridae</i>	1459
10	<i>Retroviridae</i>	2307

Table 3.2: Viral families resulted from partitioning the interactions according to the complete blindness criterion. The sample size for each family is the number of positive PPIs in the data set involving viruses from that family.

ST5 was 81%. Figure 3.8 shows the distribution of the confidence of prediction in both studies.

In ST6, the SLiM set consists of 219 instances in 316 unique viral protein identifiers. Virus-Mentha originally has 445 viral proteins. The two sets intersect in 39 proteins in 66 ELM instances. There were a total of 2,015 positive PPIs and 1,940 negative PPIs for the ELM-set, and 3,430 positives and 3,219 negatives in the non-ELM set. Accuracy, sensitivity, and specificity for the non-Masked set were 81.90%, 80.71%, and 83.06%, respectively; and for the masked set were 82.59%, 81.65%, and 83.53%, respectively. Standard error was 0.82%. ST7 assesses how mining the true interactions alone may affect the prediction for novel viruses. With a confidence threshold of 0.5, the accuracy of prediction in the partitioning criterion *Cr1* was 38% and for criterion *Cr2* was 25%.

Grouping true PPIs from the different ten families into positive and negative examples of each family independently in ST8, and then testing using two-class SVMs resulted in exceptionally high accuracy. The weighted average accuracy was 91.6%, the weighted average sensitivity was 88.6% and the weighted average specificity was 94.5%. The support vector ratio was 19% in average. Using the same models to test DeNovo negative examples in each case, there was no single example correctly classified. ST8 also compares the original confidence scores to the assigned scores by one-class SVM models. 100 out of 301 examples originally of a score above 0.5 were correctly identified, and 2,722 examples of a score below 0.5 were assigned score above 0.5. If we name the class of examples above 0.5 as positive, the results correspond to 33% sensitivity, 53% specificity, and 52% total accuracy. The support vector ratio in this study is 50%, reflecting the ease of discrimination between the two clusters.

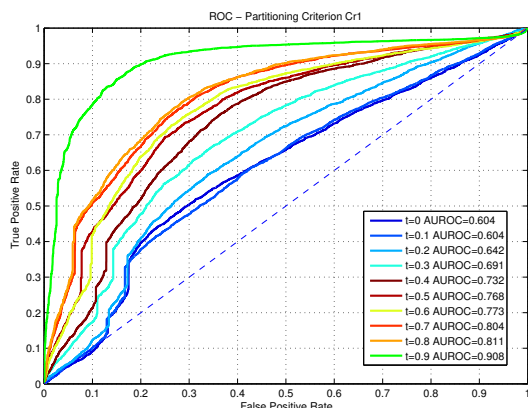
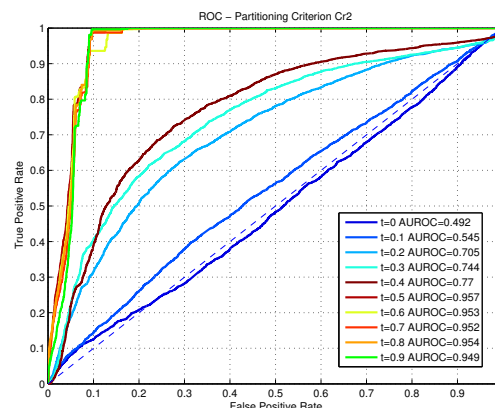
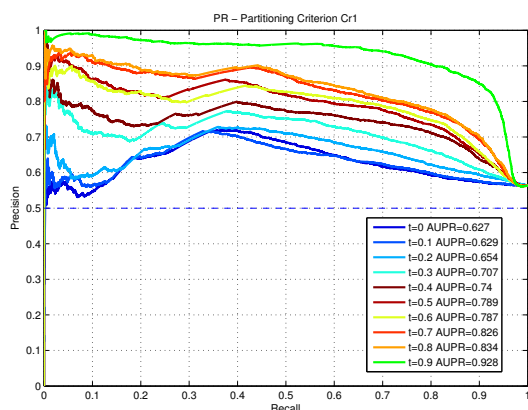
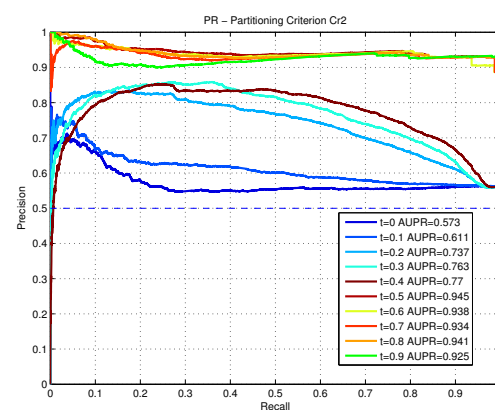
(a) ROC for $Cr1$ (b) ROC for $Cr2$ (c) PR for $Cr1$ (d) PR for $Cr2$

Figure 3.7: ROC and precision-recall (PR) curves for testing DeNovo with the two partitioning criteria $Cr1$ and $Cr2$.

3.4 Discussion

DeNovo makes it possible to predict interactions of proteins from a novel virus with human proteins accurately by using PPIs from different viruses, sequence-based features, and a biology-inspired negative subsampling method. The results demonstrate the generalization and robustness of DeNovo. See Figure 3.9 for an overview of the main ideas of DeNovo that cooperatively have solved the problem of predicting for novel viruses.

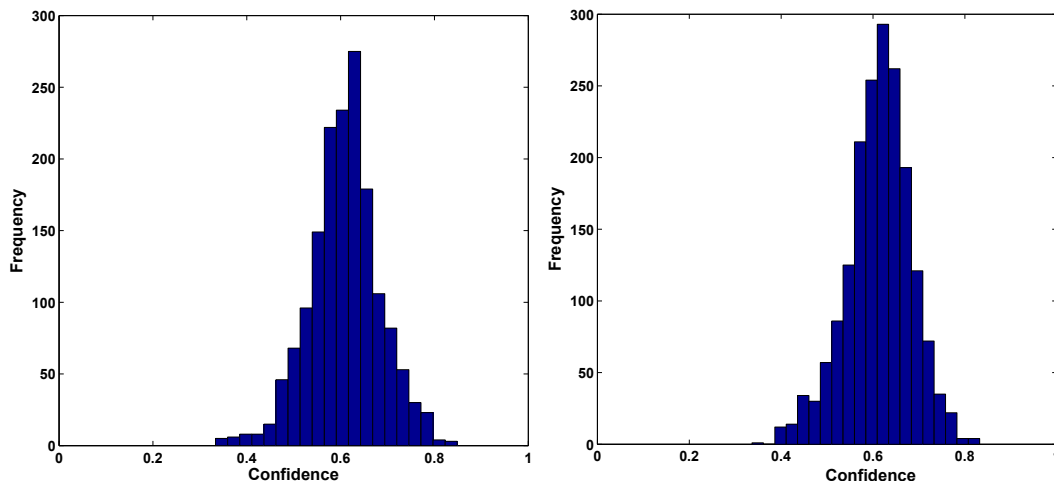


Figure 3.8: Confidence of positive PPI prediction in studies ST4 (Left) and ST5 (Right). Sensitivity in both cases was about 94%.

3.4.1 Negative Sampling

In the typical random negative sampling method, for a viral protein x , the set of human proteins that serve as partners in the negative intersections with x are picked randomly from the set of all human proteins (in the data set) that x does not interact with. Random sampling is prone to produce false negative examples when used to generate negative PPIs, but it is widely accepted in intra-species PPI prediction, for it is proven to be unbiased [7].

I expect random sampling to produce much more incorrect negative examples than expected in the intra-species PPI case (according to Hypotheses 2 and 3 in Section 3.1.2), misleading the learning process and lowering the prediction sensitivity. The resultant significant difference in accuracy between the inter-species and intra-species prediction with random sampling (Section 3.3) demonstrates how difficult predicting for novel viruses is and the high noise introduced by random sampling in the inter-species case.

The proposed negative sampling method outperforms the typical random sampling ($T = 0$) in both cases of partial and full simulation of novel virus prediction; see Figure 3.5. Testing with the 49 (R, S) pairs resulting from the partially blind partitioning (Criterion $Cr1$), the accuracy raised to 76.4% at $T^* = 0.8$, and eventually increased to 86% with $T = 0.9$. Accuracy, sensitivity, and specificity were increased when T^* was used with 16%, 16.8%, and 17.6%, respectively, over those achieved with random sampling. The accuracy gain is about 60-fold of the standard error (0.27%). Additionally, a 12% drop was observed in support vectors ratio memorized by the trained model, which reflects less over-fitting and an easier learning task.

The accuracy improvement is maintained even when the sequences of the novel virus proteins are completely dissimilar to all other viral proteins used in training the model (Criterion

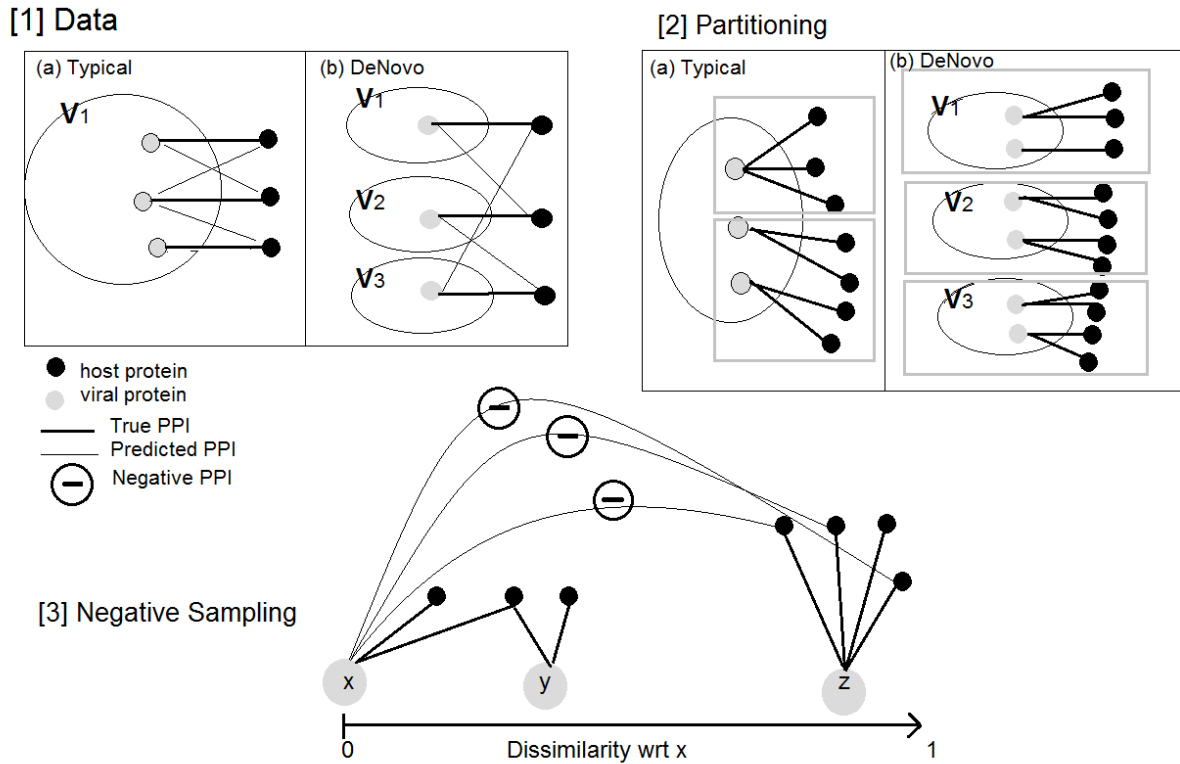


Figure 3.9: Overview of DeNovo key ideas. [1] Instead of using a single virus PPIs with the host proteins (As in typical prediction of interactions between virus and host proteins), DeNovo utilizes PPIs from different viruses interacting with the host. This makes predictions possible for different viruses. [2] DeNovo does not divide the data set of PPIs randomly into training and testing. Instead, it groups the interactions of each virus (similar viruses or within a viral family) into a single partition. All partitions except one are then used for training and the excluded one for testing to imitate the situation of a novel virus. [3] Dissimilarity-based negative sampling basic idea. If two proteins are sequence-dissimilar, they are less likely to share interaction partners. Hence, the true partners of one may serve as the negative partners of the other.

$Cr2$); see Figure 3.5. An increase in accuracy of 23.12% was achieved, which corresponds to a 96-fold standard error improvement over random sampling, supporting my Features Learning Hypothesis (Section 3.1.2).

Comparative high accuracies were also achieved when the model was trained with parameter values other than the optimal ones (C^* and γ^*), reflecting the stability of this method. Additionally, the monotonic increase in accuracy is maintained with the increase in T , regardless of the partitioning criterion or the data set used for training; see Figure 3.5.

The results from this negative sampling method also support my Shared Interaction Partner

Hypothesis (Section 3.1.2), for when more dissimilarity was enforced in pairing the negative example partners (by increasing T), the accuracy and the sensitivity improved. Such improvement suggests that less noisy negative examples were generated, facilitating the classification learning process.

My negative sampling method picks the least likely interactions as negative examples to ease the classifier task for better discrimination between the true and unlikely interactions. Meanwhile, the dissimilarity distance constraint maintains the unbiased characteristic of the typical random sampling. First, it does not exclude human proteins similar to those interacting with the viral protein in question, x , but it excludes those interacting with viral proteins similar in sequence to x . Second, the dissimilarity distances are normalized independently for each viral protein.

Dissimilarity threshold T is used to pick less likely interactions as negative examples. We need to define an optimal value for the dissimilarity threshold (T^*) that best serve this purpose. According to the twilight zone concept of protein sequence alignments, when the sequence identity score between a pair of proteins falls below 20%, structure similarity between them is minimal [104]. Considering the central dogma of genomics (sequence determines structure determines function), by selecting T^* at 0.8 (which corresponds to 20% sequence similarity), we increase the probability that the two proteins have no similar interaction interfaces, and consequently, they are less likely to share an interaction partner. Thus, if viral proteins x and y interact with human proteins h_x and h_y respectively, and the dissimilarity distance between x and y is less than 0.8, then x is less likely to interact with h_y , and thus the pair (x, h_y) can serve as a negative (unlikely) interaction.

When generalizing this constrained sampling method to human-human PPI prediction, no improvement was observed. I trained a model over human-human PPIs with the negative examples for both training and testing are drawn using the proposed constrained sampling method. I also tested the model when the testing negative set is drawn randomly from the set of unreported PPIs without constraints. This robustness of performance aligns with the design purpose of the constrained method, for it is built to respect the special interaction nature of viral proteins (acting like hubs, interacting through short linear motifs, and sharing many host protein target especially within the same viral family).

3.4.2 Comparison with Data Mining-based Sampling

Data mining approaches are used in PPI prediction to bypass the problem of generating negative examples needed for classic machine learning techniques. A recent negative sampling method clusters the true PPIs into two subsets: a large cluster of highly similar PPIs and a smaller distant cluster [77]. The study hypothesizes that most of the PPIs reported should be true ones and will cluster together, and the fewer PPIs falsely reported as positives will cluster away from the true ones. A model is then trained and tested on subsets of the two clusters. When I replicated the study for the case of viral families scenario, exceptionally high

accuracy was achieved (91.6%). However, when I examined the members of each cluster (in ST8), I found that (at least in the case of viral proteins) the two clusters do not necessarily correspond to real positive and negative PPI classes, suggesting that this method is not suitable for viral-host PPIs.

In ST8, the low support vector ratio (in the models trained on examples generated by the one-class SVMs on viral family groups) indicates the easiness of the classification task. This observation does not align with the fact that virus families have almost no sequence similarities among themselves, which makes the prediction task complicated. The reported increase in accuracy corresponds to about 80-fold of standard error in the investigated case of family prediction.

From another perspective, in ST8, testing the generated models on the negative examples, my dissimilarity-based negative sampling method resulted in a flat zero specificity. Taking into consideration that my negative examples at dissimilarity threshold of $T = 0.8$ are highly likely to be true negatives as reasoned in the main study, the flat zero indicates the clustering performed using the one-class classifier does not align with the true positive and negative PPI classes.

From the bias in the scoring examination in ST8, it is evident that scoring using a one-class SVM model does not correspond to positive and negative classes, and that the features used by the classifier to model the highly similar positive examples do not align with discriminating features that can separate true PPIs from negative ones. This further demonstrates the bias in this one-class negative sampling method presented in [77] when applied to virus-host PPIs.

These three results from ST8 support my intuition of the one-class SVM negative sampling bias (in the case of PPIs from multiple virus-host pairs). My understanding of this problem is that the non-linear complicated problem of discriminating between positive and negative PPIs was transferred to a linearly separable problem using a clustering-like technique. There was no control on the features that validates the two linearly separable classes as indeed corresponding to positive and negative PPIs. The two-class classifier was then used to train and test on these two classes; this is like clustering some data and then using a classifier to assess how well the two clusters are separated, thus exceptional high accuracy is expected. But these clusters do not necessarily correspond to the two PPI classes we are interested in, as demonstrated by my results above.

3.4.3 Data Partitioning

The data partitioning schemes I proposed address the need to divide the data set into training and testing subsets such that the testing subset contains interactions of viruses taxonomically far from those in the training set in a way that imitates predicting for a novel virus (or for a known virus with no or few known interactions). Additionally, they address dividing the entire data set into multiple subsets, each of which can serve as a testing subset, while

the remaining subsets are grouped into a corresponding training subset (Section 3.1.1). A random split cannot satisfy these requirements.

When a novel virus emerges, it is more likely to be of a low taxonomy rank (an isolate or strain) than to be of higher rank (a species or genus), because a new virus typically evolves from known ones. However, we cannot specify a fixed taxonomy rank (species for example) as a partitioning criterion because the branches of the viral taxonomy tree significantly vary in length (Figure 3.2). Some subsets would carry interactions of viral proteins nearly identical or entirely different from the other viral proteins in the remaining subsets, making performance evaluation inconsistent.

The accuracy difference between the two partitioning schemes is explained by and supports my Sequence Similarity Hypothesis (Section 3.1.2). In the partial blindness case, PPIs from related viral proteins (to the ones in testing) contribute to the training set, making the learning task easier. Testing on all interactions from a single family means that the model is not trained on any viral proteins with sequence similarity to the ones in testing. Thus, the model learns most of the discriminating features with respect to human proteins and other viral proteins and may miss some features that distinguish the interactions of the foreign viral proteins. This result confirms the generalization characteristic of DeNovo and further supports my learning hypothesis (Features Learning Hypothesis, Section 3.1.2).

Prediction in the case of partial blindness is more accurate and more confident than with full blindness partitioning; see Figure 3.6. However, partitioning does not affect and is not affected by the negative sampling at $T^* = 0.8$. At this threshold (in both schemes), negative human partners are picked from the partners of the other viral proteins that are placed in the midnight zone with respect to the viral protein in question. The purpose of partitioning is to simulate the situation of a novel virus, not to aid the negative sampling.

3.4.4 Features

The main reason I utilize sequence-based features is to maintain generalization of DeNovo to any other PPI domain and to maintain prediction ability to any novel virus once its proteins are sequenced. Advanced features (structure, gene ontology, or interaction domains) require prior knowledge that can be difficult and time-consuming to obtain [44, 107]. Sequence dissimilarity among viral families hampers using homology-based and structure-based features in novel virus-host PPI prediction. Amino acid sequence information has been proven to be sufficient for PPI prediction (I also demonstrate this in Section 3.4.5) [44].

Although domain-based prediction is applied (on a limited scale) in the pathogen-host area [130], application for novel virus prediction needs prior knowledge of virus domains and related information [85]. As such knowledge is not available for many viruses and is difficult to obtain, one solution is to predict viral domains computationally for new viral proteins and apply domain-based methods on them. However, viral proteins are different from human

proteins in that they are mostly loosely packed, enriched in disordered regions [114], or of unknown and unmapped structure [95], making prediction of viral domains using methods designed for human proteins questionable.

Similarly, domain-based methods that use short linear motifs (SLiMs) in predicting viral PPIs face the same objections: the generalization problem for lack of data and difficulty of SLiM prediction (for SLiMs are short and degenerate in nature). Some studies assign viral proteins whose sequences are predicted or known to have SLiMs similar to ones known in human as interacting candidates with human proteins carrying the corresponding counter domains [37, 106]. These methods are limited by the 239 known human SLiM classes and the 219 known viral SLiM instances [28].

I assessed how sequence motif information may help to improve the accuracy of DeNovo prediction by masking the viral SLiMs in testing PPIs (in ST6, Section 3.2.6). No significant change in accuracy was observed when the known viral SLiMs were masked, suggesting that the SVM model captures other interaction associated features that are sufficient to make the discrimination. Another possibility is that the available number of SLiMs is insufficient to make a difference.

3.4.5 Prediction Approach

I reasoned that my negative sampling method is robust and not biased (Section 3.4.1). I thus found that the supervised machine learning approaches (which require addressing the challenge of negative sampling) are more suitable for the domain of predicting PPI of a novel virus with its host than data mining and transfer learning approaches, as discussed below.

In data mining-based learning, a putative PPI is classified as likely or unlikely based on its similarity to the known PPIs, without the need of negative examples. My main concern is that dissimilarities among viral families will turn a model representing true PPIs of some viral families to be insufficient to describe interactions of foreign viruses. I tested this hypothesis in ST7 (Section 3.2.6). With a similarity threshold of 0.5, accuracies of 38% and 25% were obtained for partitioning schemes 1 and 2, respectively. These low accuracies suggest that modeling true PPIs of some viruses is insufficient to capture interaction-discriminating features for other viruses.

My concern of using transfer learning in virus-host PPI prediction for novel viruses is the need of prior knowledge to design a transfer function specific for each novel virus. I compared the performance of DeNovo, which uses sequence-based features, with transfer learning that uses more advanced features [58] on PPIs from 3 bacterial pathogens of different taxonomic orders (ST1 in Section 3.2.6). With near optimal accuracy of 97% and low support vector ratio, DeNovo demonstrates that its use of SVMs is sufficient (and probably a better fit) to capture discriminative features that otherwise need to be explicitly modeled in transfer learning. Another study uses homology knowledge transfer of gene ontology (GO) features

in HIV proteins [75]. Such advanced features hamper generalization to most viruses where GO data can be time-consuming and a challenge to obtain.

3.4.6 Generalization

Prediction for novel bacteria using DeNovo trained only on PPIs of the other two bacteria with human results in near optimal accuracy. Approximately half of the human proteins in testing appeared in training. This result supports my intuition that predicting for bacteria is an easier task than for viruses, for viral families have the constraint of dissimilarity in sequence.

DeNovo generalizes well. I tested my model on five data sets assessing how it behaves in different pathogen domains, across pathogens with a shared host, and on PPIs whose protein partners are foreign to the model (Section 3.2.6). These tests demonstrate how to apply DeNovo and its generalization ability and robustness. However, I did not examine generalization to non-human host PPIs with viral proteins due to lack of data.

I used DeNovo on three PPI sets from different bacteria with human (in ST1) in a cross validation fashion. Each set was tested once against a model trained on the other two, similar to the Criterion 2 scenario. DeNovo achieved average accuracy, sensitivity, specificity, and support vector ratio of 97%, 94.5%, 97.5%, and 28.8%, respectively. This near-optimal result demonstrates the prediction ability of DeNovo. It also shows how predicting for bacteria is easier than for novel viruses, for bacteria do not have the dissimilarity constraint among families.

To further assess the difference between viral and bacterial proteins in PPI prediction, DeNovo was trained on all VirusMentha PPIs (between human and viruses) and tested on bacteria-human PPIs (in ST3). The model could recognize up to 69% of the tested PPIs. DeNovo was further trained on PPIs from a plant host with a bacterial pathogen and tested on proteins of different pathogens (viruses, fungi, and bacteria) that target the same host (in ST2). Sensitivity reached 84.6%, while specificity was 65%, reflecting some differences of these pathogen proteins and their interaction interfaces.

When DeNovo was tested on PPIs from viral proteins foreign to VirusMentha, it correctly classified 94% of them (in ST5). When both human and viral proteins were foreign (in ST4), the prediction accuracy was approximately the same. This result supports my Features Learning Hypothesis (Section 3.1.2) that DeNovo can learn with respect to host proteins rather than specific viral proteins.

ST4 and ST5 show that the model is sufficient to recognize the positive PPIs even if the human proteins are foreign. This observation suggests that the same interaction discriminating features are maintained in the foreign human proteins and across the foreign viral proteins, unlike the case in ST3 and ST2 where the recognition was relatively low when the prediction was for PPIs of different pathogens than those used in training the models.

3.4.7 Application Scenarios

ST4 and ST5 serve as testing scenarios. Suppose a novel virus emerges, and we want to give a list of potential interacting and non-interacting human proteins, or we have a list of putative interactions of that novel virus with human proteins, and we want to assess how probable they are. Here I test both cases. In ST4, foreign viral proteins are tested against foreign human proteins to imitate the second situation. In ST5, only the viral proteins are foreign to imitate the first situation.

In real scenarios, where a list of putative virus-human PPIs is given to estimate their probabilities, confidence scores will help the researchers make informative decisions. Figure 3.8 shows the distribution of the confidence of prediction in both studies.

If a researcher's interest is to generate a list of likely and unlikely interactions for a set of novel viral proteins, she should (1) pair each viral protein with all human proteins in the training set, (2) extract their features vectors, and (3) use the trained model to assess the probability (confidence) of each interaction. The researcher then can select the interactions above (or below) a certain confidence threshold.

3.5 Conclusion

I introduced the problem of predicting virus-human PPIs for a novel virus, for which we do not have any known interactions. My proposed solution, DeNovo, solved this problem efficiently by (1) exploiting the known PPIs between a large number of viruses and their shared host (human) to learn from; (2) introducing a method to reduce the expected noise in generating the negative PPI examples from the true ones; and (3) partitioning the data set in a way that simulates the situation of a novel virus, hence makes performance evaluation realistic. The model achieved significantly high accuracy while only using protein sequences to learn and predict, making the prediction possible for virtually any virus infecting human. I demonstrated that DeNovo generalizes well and is robust and biologically sound.

Chapter 4

DeNovo-Human

4.1 Introduction

Interactions between viral and human proteins define viral tactics in invading human cells and hijacking their regulation [24]. Identifying the interactions taking place between the virus and its host proteins facilitates greater understanding of the virus life cycle and consequently helps in antiviral drug design [25]. Over the past decade, progress has been made in predicting the interactions between viral proteins of a specific virus and human proteins using machine learning and different viral protein features. However, these solutions are single-virus centered, require a large set of known protein-protein interactions (PPIs) for the virus, and usually require advanced knowledge of features beyond protein sequence (such as protein structure), making them inapplicable for all viruses other than HIV. Except for my first framework, DeNovo, no previous approach makes it possible to predict PPIs between proteins of novel viruses and human [35] (Chapter 3). However, DeNovo suffers from a subsampling problem, which limits the size of the human protein library that a virus can be tested against. For each viral protein, there is a set of human proteins for which DeNovo cannot make predictions, with estimated average coverage of less than 20% of the library size; see Figure 4.1.

The problem DeNovo-Human addresses can be defined as follows: Given a set of human proteins, a set of viral proteins, and a set of PPIs between them, including some PPIs for a viral protein v , predict which of the human proteins can interact with v . This problem can be viewed as predicting missing edges between viral and human nodes in an existing PPI network. My first framework, DeNovo, on the other hand, addresses the problem of extending an existing network by adding new viral protein nodes and predicting their edges (interactions) with the human protein nodes in the network.

The two problems addressed by DeNovo and DeNovo-Human can be understood in the light of a new PPI prediction classification scheme [91]. In that scheme, PPI prediction is classified

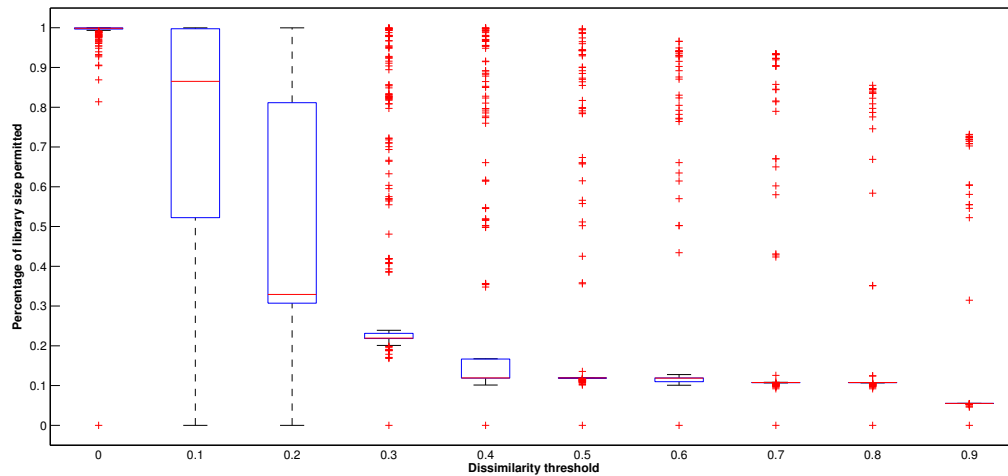


Figure 4.1: **DeNovo framework limitation on the testing library size.** DeNovo is designed to discard specific human proteins from training (sampled for each viral protein) to reduce the noise in the negative random sampling process. As a result of this subsampling, for each query viral protein, a set of human proteins in the library (used to train DeNovo) is excluded as well. DeNovo performance cannot be extended to the excluded sets. The figure shows the distribution of the percentage of human proteins that can be tested for a set of 445 viral protein at each value of the dissimilarity threshold (used in the subsampling). Usually, DeNovo is operated in the range $[0.4, 0.7]$. The theoretical optimal point is 0.8. Thus, the coverage power of DeNovo is mostly less than 20% of the human library size.

into three classes. In the first class, C1, the two proteins in a tested pair have PPI examples used in training the prediction model. C1 is equivalent to having the two proteins of existing nodes in the PPI training network. In the second class, C2, only one protein of the pair has training examples, while the other does not belong to the PPI training network. In the third class, C3, neither of the tested proteins in a pair has training PPIs examples, which means both of them are foreign to the training PPI network. Accuracy measures in C1 are demonstrated to be much higher than those of C2 and C3 because the model has seen training examples of the two proteins in a testing PPI pair [91].

Relating these PPI prediction classes to DeNovo, it is apparent that DeNovo predicts for C2, where the testing viral proteins are novel but the human proteins belong to the training PPI network. However, I managed to avoid the C2 expected low accuracy in DeNovo by constraining the training sampling, but at the cost of the testing library size (Figure 4.1). I found that, if we demand a few PPIs of a testing viral protein be known in advance and use them in training the model, the problem of DeNovo will be transformed from C2 to C1 (where accuracy is much higher). If, additionally, random negative sampling is used instead of dissimilarity sampling (which I proposed to achieve higher accuracy in the C2 case of

DeNovo), we can enjoy both the broader coverage of the full human protein testing library and the high accuracy of C1 prediction. These transformations led to the DeNovo-Human problem, where I seek to find the missed edges connecting a viral protein node to human protein nodes in the network.

DeNovo-Human is a framework for predicting the interactions between viral and human proteins for the case when the viral proteins have some known interactions. DeNovo-Human exploits the shared human proteins for different viral interaction partners to learn for human proteins rather than for a specific viral protein, thus overcoming the data shortage limitation (the required large PPIs set for single-virus prediction). The few known interactions for each viral proteins help increase the precision (and accuracy) of the prediction up to 85% and the AUPR up to 92%, a 10% performance increase over human-human PPI prediction mode, which is thought of as an easier problem [91]. Thus, we can now predict PPIs for virtually any virus once some interactions of its proteins with human proteins are available.

DeNovo-Human is designed by applying the pooling hypothesis and employing some viral hijacking strategies. Instead of training a classifier to recognize how a specific viral protein interacts, training it to recognize how a set of human proteins interact will allow recognition of their PPIs with other viral proteins. This pooling of PPIs from different viruses provides a larger human testing library and reduces the number of required training PPIs for each query viral protein. I leveraged two common viral hijacking strategies to aid the learning process: viruses frequently interact with common human targets, and their interactions are frequently mediated by short linear motif interfaces.

DeNovo-Human is trained using only the amino acid sequence of the viral and human proteins of the known PPI data set. For testing, the trained model uses only the sequence and a few PPIs of the query viral protein to make predictions against a large human protein library. Viruses tend to target a subset of human proteins more frequently, for they are associated with cell processes viruses need to hijack [32]. As the human proteins in the library are those reported to previously interact with some viral proteins, prediction of a viral protein PPIs against this library can provide a near-complete coverage of the human proteome.

4.2 Methods

4.2.1 Data set

I collected 24,077 unique PPIs between 5,738 human proteins and 1,245 viral proteins in 381 different taxonomic groups (mostly species) from HPIDB [3, 59]. The protein sequences were retrieved directly from UniProt. Negative PPIs were selected according to the typical random negative PPI sampling technique [108], with positive to negative ratio of 1:1, as is the convention in PPI prediction studies [91]. Random negative sampling is believed to be the least biased for PPI prediction applications [8, 90]. The PPI set (positive and negative)

was randomly divided into training and testing to simulate the effect of predicting in the C1 mode.

4.2.2 Learning

SVMs use kernel functions to transform the features from the input space into a higher dimensional space where linear separation is possible. Additionally, kernels measure the similarity between data points. A kernel must be a dot product of the feature vectors of the two data points between which the similarity is measured. For example, if the feature vectors are X_1 and X_2 , then the linear kernel is $K_l(X_1, X_2) = X_1 \cdot X_2$ and the Gaussian kernel is $K_g(X_1, X_2) = \exp(-\gamma \|X_1 - X_2\|^2)$.

Kernels are designed to compare two data points. In the case of protein-protein interactions, we need a kernel that considers the relation between the four proteins in the two data points (two interactions, positive or negative) [7]. Here I compare three of these pairwise kernels proposed to measure the similarity between pairs of proteins. I denote them K_1 , K_2 , and K_3 . I selected these kernels as they are reported to perform the best in human-human and yeast-yeast PPI prediction applications [91].

K_1 and K_2 use a signature kernel to measure the similarity between two proteins; they differ in which proteins they compare to measure the similarity between two interactions (or two protein pairs) [62, 73, 119]. The signature they use, $s(A)$, is an array of the frequency of all the 3-mer combinations of the twenty amino acids of the sequence A . However, the direction of the sequence is not considered. For example, an amino acid sequence “LVMTTM” will have “MTT” counted twice, one for each direction. The signature kernel between two proteins A and B is the dot product between their signatures $s(A)$ and $s(B)$. See Equation 4.1.

$$K(A, B) = s(A) \cdot s(B) \quad (4.1)$$

K_1 compares (or measures the similarity between) the counterpart proteins only. If we have two interactions (or pairs of proteins) (p_1, p_2) and (p_3, p_4) , K_1 compares p_1 to p_3 and p_2 to p_4 and corrects for the variability of protein lengths. See Equation 4.2. K_2 performs the comparison between the counterparts and also accounts for the similarity between the non-counterparts (in the previous example p_1 and p_4 and not counterparts). See Equation 4.3. Both equations are written in terms of human proteins h_1 and h_2 and viral proteins v_1 and v_2 to make the comparison order clearer and directly linked to the case of comparing interactions between human and viral proteins.

$$K_1((h_1, v_1), (h_2, v_2)) = \frac{K(h_1, h_2)}{\sqrt{K(h_1, h_1)K(h_2, h_2)}} \cdot \frac{K(v_1, v_2)}{\sqrt{K(v_1, v_1)K(v_2, v_2)}} \quad (4.2)$$

$$K_2((h_1, v_1), (h_2, v_2)) = (K(h_1, h_2) - K(h_1, v_2) - K(v_1, h_2) + K(v_1, v_2))^2 \quad (4.3)$$

K_3 considers the similarity between the counterpart proteins only when measuring the similarity between two interactions (or pairs of proteins) [108]. See Equation 4.4. Like K_1 and K_2 , the signature used is the 3-mer frequency of the amino acid sequence, but unlike them, the 3-mers are only counted in the forward direction of the sequence. Although the two signatures conceptually differ, there is no noticeable difference in the result when either of them is used with K_3 . Thus, I wrote the equation of K_3 (Equation 4.4) in terms of the signature kernel used with K_1 and K_2 (Equation 4.1).

$$K_3((h_1, v_1), (h_2, v_2)) = K(h_1, h_2) + K(v_1, v_2) \quad (4.4)$$

4.2.3 Framework evaluation

The performance was evaluated in a cross-validation fashion over 40 rounds. For each round, the data set was randomly split into training and testing with a ratio of 9 : 1. The positive to negative ratio used is 1 : 1, as is the convention in PPI prediction applications. I used both the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR).

To evaluate the effect of pooling PPIs from different viruses on the case of predicting for a single virus, I run the training and testing rounds in DeNovo-Human on each viral taxonomic group (mostly species) as a single virus and report the performance over 40 different rounds for each group. To make the experiment unbiased, taxonomic groups are excluded if they have fewer than three viral protein in the data set (for generating negative examples is not possible in this case), if they have few human proteins (< 10 , an arbitrary threshold), or if they have few positive examples (< 100 , an arbitrary threshold).

I also want to compare the prediction performance for the virus-human and human-human cases. I run 40 rounds of training and testing, using the same learning model as in DeNovo, on a data set of 13,944 human-human PPIs among 4,303 human proteins. The PPIs in this data set are obtained from high-quality reproducible yeast two-hybrid assays using all versus all examination [103]. I chose this yeast two-hybrid data set, rather than collecting reported human-human interactions from the available databases, for three reasons. First, the Y2H data set was built by performing high-quality and reproducible all versus all assays, thus, the negative examples are all known and less noisy than what can be randomly sampled for reported human-human PPIs. Second, the Y2H interaction basis is physical, regardless of localization or any other interaction factors. Third, the interactions being physical in the assays makes the positive PPI set less noisy, for it does not contain reported PPIs of other interaction types (association and indirect activation for example). The Y2H data set was processed the same way as the virus-human data set, except that I tested it twice. In one

test setting, I considered the PPIs as taking one direction (the same as virus-human testing), and in the other, I considered the bi-directional nature of the intra-species PPIs by adding the PPIs with the reverse direction to the data set before processing.

Additionally, I want to quantify how DeNovo-Human performs with a varying number of PPIs the query viral protein may have in the training set. Within the 40 rounds of training and testing (performed earlier for evaluating the model), each viral protein that has n training positive examples has its testing outputs added to the n^{th} stack. Each stack performance is evaluated independently to give an estimate of how the framework will perform on average in case of a new viral protein that has n known PPIs with human proteins.

To estimate the expected accuracy when testing some random viral protein against a specific human protein in the library, I performed a quantitative experiment. For each human protein, I collected all its occurrence in the test sets over the 40 cross-validation rounds along with the true labels of the PPI and their predicted estimates. I then calculated the performance measures for each human protein independently.

4.2.4 Network prediction

I used DeNovo-Human to perform an all versus all prediction for the human-virus PPI network. The study data set was used to train DeNovo-Human, along with a generated negative PPIs with a ratio of 1:1 positive to negative. Then all versus all pairs of the viral and human proteins were tested to predict the missing edges in the current network. The test was repeated 40 times to reduce the noise of the negative random sampling. The predicted score was calculated for each edge. The network adjacency matrix was determined by the predicted edge scores on three bases: reporting an edge positive if its average score (over the cross-validation rounds) is above the mid-point threshold (averaging), if the edge is predicted positive for the majority of the experiments (majority voting), and if the edge is predicted positive in all of the experiments (all ones, complete agreement, or strict prediction).

I used the virus-human PPI network predicted by DeNovo-Human to estimate the positive to negative ratio in the virus-human network. As this is a two species network where we are not interested in intra-species edges, the ratio ($Ratio_{vh}$) was estimated as for how many edges in average connecting a single virus node to the human proteins in a network of n human proteins. If the set of viral protein nodes is V , and $d(v)$ is the node degree of viral protein v , then we can estimate $Ratio_{vh}$ as the average positive to negative connectivity of V ; see Equation 4.5.

$$Ratio_{vh} = \frac{1}{V} \sum_{p=1}^V \frac{d(p)}{n - d(p)} \quad (4.5)$$

I randomly selected 1000 different n values and sampled the predicted network 100 times

Method	SVr	Accuracy	Sensitivity	Specificity	AUROC	AUPR
M1	74.9 ± 0.2	85.2 ± 0.46	85.2 ± 0.67	85.1 ± 0.7	91.8 ± 0.4	91.1 ± 0.51
M2	58.0 ± 0.3	85.3 ± 0.56	87.1 ± 0.64	83.4 ± 0.79	91.2 ± 0.43	89.2 ± 0.69
M3	59.2 ± 0.30	85.3 ± 0.55	86.8 ± 0.74	83.7 ± 0.8	92.0 ± 0.41	91.2 ± 0.57

Table 4.1: Comparison of the average performance over 40 sampling sets for the three top feature encoding kernel methods for the DeNovo-Human framework. The first method is the one I use in DeNovo-Human for reporting and final product.

at each value to estimate the ratio $Ratio_{vh}$. I also used the same approach with human-human PPI network collected from String DB in December 2016 to estimate the positive to negative ratio for human PPI networks, $Ratio_{hh}$, with the number of nodes (n) in the subnetworks sampled 4000 times. The same equation above can be used to estimate $Ratio_{hh}$ directly, with replacing the number of viral proteins V with the number of human proteins in the human-human PPI network (n). I do not need to correct for the inter-species case, as self-interactions are considered in this estimate; see Equation 4.6.

$$Ratio_{hh} = \frac{1}{n} \sum_{p=1}^n \frac{d(p)}{n - d(p)} \quad (4.6)$$

4.3 Results

I evaluated how well DeNovo-Human can predict protein interactions of a query viral protein. On average, DeNovo-Human predicted PPIs of different viral proteins (with human protein) with an average area under the curves AUROC and AUPR of 91.8 and 91.5, respectively (Table 4.1 and Figure 4.2). Compared to the only other solution available for the same problem, my first framework DeNovo, the area under curves are almost the same, but with the testing coverage mostly limited to 20% of the human protein library (Figure 4.1).

The obtained result out-performs the prediction for the single virus-human PPI case. The single virus prediction achieved an average AUC of 84.1 and an average coverage of 351 human proteins (Figure 4.3). The results from DeNovo-Human pooling are also close to when all the sophisticated knowledge about HIV is used for training [99].

The framework allows predicting for a viral protein with as few known PPIs as a single one. DeNovo-Human predicts for a query viral protein by training the model with the query protein PPIs added to the data set. The framework then predicts the possibility of having interactions between the query viral protein and the library of human proteins (in the training data). To enable informative decisions on the prediction results, I estimated the expected accuracy given a different number of known PPIs for a query viral protein using

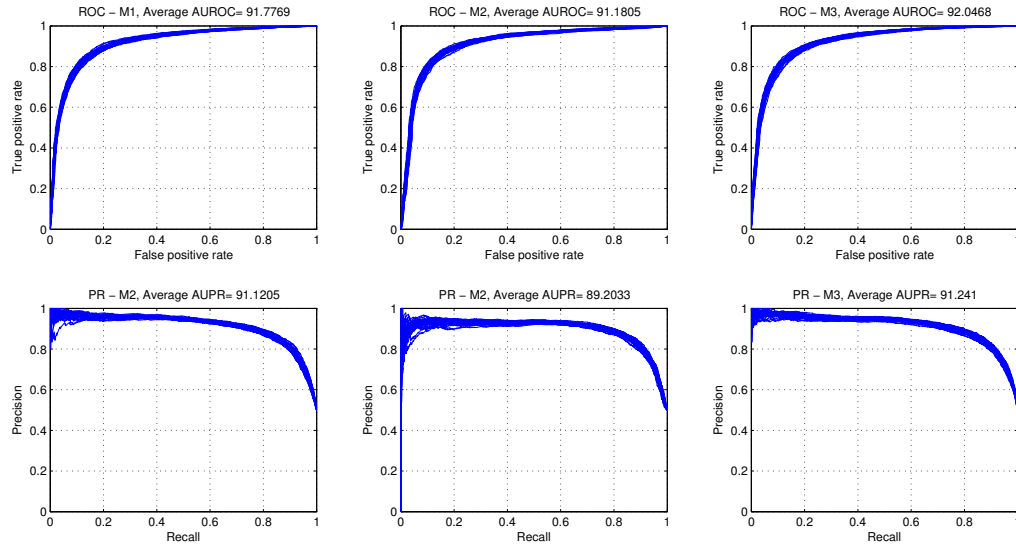


Figure 4.2: Area under ROC and PR curves for the three kernel methods compared for feature encoding. The three methods $M1$, $M2$, and $M3$ correspond to the three kernels K_1 , K_2 , and K_3 discussed in Section 4.2.2.

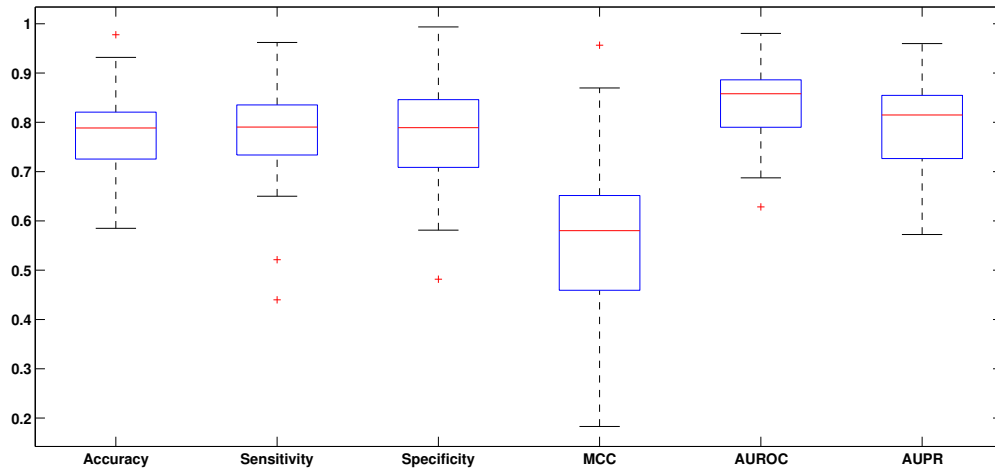


Figure 4.3: Performance for prediction over 37 single virus-human networks. Average, median, and maximum coverage are 351, 262, and 1569 human proteins.

37 single virus-human networks for larger comparison quantification; see Figure 4.3. I also estimated the expected accuracy when testing against each human protein in the library.

I investigated whether the redundancy in viral proteins that are similar in sequence in train-

Sim%	Accuracy	Sensitivity	Specificity	AUROC	AUPR
40	84.10	79.66	88.54	90.21	89.36
50	84.24	80.42	88.06	90.35	89.29
60	84.05	80.44	87.65	90.14	88.87
70	84.64	81.79	87.48	90.38	89.08
80	84.74	82.06	87.43	90.69	89.45
90	85.59	83.00	88.18	91.36	90.71
100	85.2	85.2	85.1	91.8	91.1

Table 4.2: Similarity Test 1. Comparison of the average performance when the viral proteins in the training set are **at most** Sim% similar to each other. This test examines the prediction performance when the model is trained on interactions from a viral protein v and other viral proteins similar to v and then tested to predict PPIs for v . The question is whether adding PPIs of viral proteins similar to v (along with PPIs of v) in training improves the prediction for v .

ing enhances the performance by performing two similarity tests. In the first test, I reduced the training set into proteins that share similarity less than a similarity threshold. I considered similarity thresholds of 40, 50, 60, 70, 80, and 90% (Table 4.2). The goal of this test is to quantify the effect of having PPIs in the training set from viral proteins similar to a query protein in addition to its own training PPIs. In the second test, I investigated whether having PPIs in training from proteins similar to the query viral protein, while none of that viral protein PPIs exist in training, helps to improve the prediction for that viral protein. The test was performed by limiting the training set to PPIs from proteins that are similar to the query viral proteins (Table 4.3).

Sim%	Accuracy	Sensitivity	Specificity	AUROC	AUPR
40	56.70	29.64	83.92	61.31	60.73
50	56.33	34.29	78.37	59.60	59.45
60	57.04	35.81	78.27	60.39	60.17
70	56.57	33.78	79.35	60.73	60.31
80	55.87	31.34	80.40	60.67	60.04
90	55.85	32.12	79.58	60.27	59.76
100	85.2	85.2	85.1	91.8	91.1

Table 4.3: Similarity Test 2. Comparison of the average performance when the viral proteins in the training set are at least Sim% similar to the query viral protein v , but v does not have any of its PPIs in training. The question is whether having PPIs of viral proteins similar to v but none for v in training substitute for having the model trained on PPIs of v .

4.4 Discussion

4.4.1 System design

Computational PPI prediction started with intra-species (single organism) prediction, as for human or yeast. When transferred to inter-species cases (between two organisms), prediction models have continued to be between a single virus and its host. The pooling used in DeNovo-Human intends to extend the prediction to multiple viruses taking advantage of their shared host. Pooling allows the model to be trained on PPIs pairing different viruses with human. The goal is to draw the classifier attention to how human proteins interact with viral proteins so that in future testing the classifier can predict more of their interactions.

Viruses commonly use some techniques to target human proteins, making the pooling strategy possible and fruitful. One common technique is the employment of linear interfaces (short sequence stretches called SLiMs, or Short Linear Motifs) due to the compactness of most viral genomes [24]. The short length of these interfaces facilitates capturing interaction features locally, without the need to encode complex relations to capture global or distant features. Another characteristic is the shared human protein targeted by many viral proteins. Different viruses tend to commonly target human proteins of specific functions or members in specific pathways to achieve their survival goal [24, 32]. Pooling the PPIs of proteins of different viruses with human proteins along with the shared interfaces make it easier to computationally encode how human proteins interact with viral proteins. By learning human protein specific interaction features, DeNovo-Human can predict for new viral proteins.

4.4.2 Comparison to related work

Computational prediction studies for predicting interactions between viral and human proteins are all built for single virus cases, except for my model DeNovo [35] (Chapter 3). Almost all of these studies focus on HIV protein interactions with human proteins, due to the wealth of knowledge about the virus and its proteins. Here I review some of the outstanding frameworks used for the prediction in the HIV-human case and other single viruses and compare them to DeNovo-Human.

To the best of my knowledge, the best-performing methods reported so far for predicting interactions between HIV and human proteins in terms of accuracy are those of the Tasthan and Oznur group [99, 113]. Their supervised learning framework utilizes an extensive set of advanced protein features, including “co-occurrence of functional motifs and their interaction domains and protein classes, gene ontology annotations, post-translational modifications, tissue distributions and gene expression profiles, topological properties of the human protein in the interaction network and the similarity of HIV-1 proteins to human proteins known binding partners” [113]. A random forest classifier over these features achieved an average accuracy of 91.5% (± 1.2) and mean average precision (MAP) score of 23% (± 2.17), which

means that of all pairs predicted as interacting, 23% are true positive. Their follow-up semi-supervised multi-task framework achieved marginal improvement in accuracy but with a more sophisticated learning framework [99]. It is not clear in their paper what is the average accuracy the new model achieved, but the maximum accuracy reported throughout the study is 91.9%.

DeNovo-Human achieves an average accuracy of 85.2%; if compared to the 91.5% achieved in the most sophisticated HIV model discussed above [113], one can recognize the significant achievement DeNovo-Human represents. First, my model only uses protein sequence to reach that accuracy, while the framework in [113] uses an extensive set of advanced protein knowledge that is expensive and hard to obtain for a virus. The feature complexity in their framework makes it inapplicable for viruses other than HIV, while my framework can be used with any virus with few known interactions and no further prior knowledge. Second, less-studied viruses, which previously had no chance of having their PPIs predicted, now can have predicted PPIs with only 6.5% less accurate results than the most sophisticated model made for the most well-studied virus using extensive prior knowledge. Third, the framework in [113] makes it possible to predict for HIV-1 proteins against a set of 1,406 human proteins, while my framework allows testing a query viral protein against a library of 5,738 human proteins. Similarly, comparing to the 91.9% of their semi-supervised model in [99] demonstrates the value of my framework for predicting for the less studied viruses.

There is a set of other prediction frameworks built for HIV-human PPIs, but their accuracy measures are either not clear or not directly comparable to mine. However, even if they happen to perform as well as the framework in [113], still my model has the advantage of delivering accurate performance using viral protein sequence with no sophisticated knowledge. Among those frameworks is that in [29], where structural similarity of HIV-1 proteins to human proteins is employed. Although the framework presented in [33] would have been a good comparison to mine (for it uses sequence frequencies with other domain and network features), the reported performance is not directly comparable. Please refer to the review in [5] for more models using sophisticated knowledge to predict HIV-human PPIs with less or comparable performance as those discussed above.

There is also a set of single virus-human prediction frameworks (for HIV or other viruses) that I do not consider for comparison, for I have major concerns about bias in their design that are likely to produce an unrealistic performance. The transfer learning framework in [75] has the problem of generating negative examples using exclusive co-localization negative sampling. In this negative sampling method, protein *A* is paired with protein *B* to form a negative example if and only if protein *B* is not co-localized with protein *A*. As understood from [8], selecting the negative training examples using this method converts the problem to localization classification, which is much easier to predict from protein sequences than PPI prediction. Adding that the framework in [75] is already using GO terms as features, an artificially high accuracy is expected. Similarly, the model on HTLV virus PPIs [77] has a negative sampling problem, yet it is different from the one in [75]. In [77], the set of known positive PPIs for HTLV proteins and human proteins are classified using a single class SVM

model into strong positive and weak positive. The weak positive set is then considered as a negative set. A 2-class SVM model is trained on the positive and negative PPIs generated by the 1-class SVM model. My concern is that this method of splitting the positive PPIs into positive and negative is biased towards the maximum marginal split regardless of what the two sets represent, whether strong and weak positives or some other knowledge embedded in the participating proteins (like co-localization or amino acid distributions, for example), especially that the features used are GO terms.

Another single-virus framework (other than HIV) that uses prior knowledge is the one proposed in [6], developed for predicting the interactions of human proteins with hepatitis B virus (HBV) and hepatitis E virus (HEV) proteins separately. The framework integrates various features, including protein sequence, disordered regions, amino acid composition, domain-domain association, and the PPI network topology. The study compared 44 descriptors out of these features. The best achieved cross-validation accuracy the study reports from a combination of the best performing descriptor and the best performing machine learning model is 72.41%. If compared to the above 91% accuracy achieved in [99, 113], the difference in accuracy can be due to the relative difference in the data set size and the difference in the prior knowledge amount and type [6]. A recent single-virus framework in [76] is proposed to work on small data sets of PPIs and tested on Epstein-Barr virus (EBV). The framework transfers GO knowledge from proteins homologous to those of the viral protein to make a better prediction than using only those GO terms assigned to the viral proteins. The model achieved a maximum accuracy of 75.32% and a maximum ROC-AUC of 85.03%, which are still below the performance delivered by my framework, set aside their coverage.

There is also a set of frameworks developed for single viruses other than HIV, but still, my model out-performs them. Probably the reason is that DeNovo-Human learns from many human and viral proteins while the other models are constrained with single virus interactions. This highlights the importance of pooling PPIs from different viruses in the learning process. Of paramount importance is comparing my work with the framework proposed in [21] for predicting the interactions of human proteins with those of human papillomaviruses (HPV) and hepatitis C virus (HCV). The importance of this model is that it uses a modified (and better performing) version of the third feature encoding method (Equation 4.4) I examined in my work [108]. The usage of the same features makes the comparison of my framework to theirs directly reflecting the value of pooling the PPIs from different viral proteins to train the classifier. The modified method achieved an average accuracy of 81.6% on HCV and 83.3% on HPV, while the feature extraction method I employed achieved 77.1% and 78.2% on the same data sets. My pooling approach still outperforms their accuracy, and moreover, offers to test against a much larger set of human proteins without the need for a large set of PPIs for each new virus. In the future, I may use the modified method this study propose on the hope of improving DeNovo-Human further.

4.4.3 Comparison to PPI prediction in Human

It has long been thought that inter-species PPI prediction is more difficult than intra-species prediction [85]. The inter-species case involves predicting for two proteins from different organisms, which might imply different protein properties or behavior. However, the results from DeNovo-Human suggest that it is not always the case. In performing my experiments on the virus-human case, the data set was prepared in the same way as the two sets of PPIs of human and yeast proteins reported in a study comparing top performing PPI prediction methods [91]. DeNovo-Human achieved about 10% higher performance in terms of accuracy, the area under the ROC curve, and the area under the PR curve for C1 prediction. Comparing with C2 and C3 prediction modes is not valid, for DeNovo-Human is designed to operate in the C1 mode by requiring PPI examples for each query protein.

Intra-species and inter-species PPIs can be viewed as undirected and directed edges, respectively. The difference in viral proteins and human proteins nature makes it unnecessary to compare viral proteins with human proteins while training or testing the SVM models. In the intra-species case, an SVM kernel may need to perform cross-comparisons between the two proteins in each pair with all other PPI pairs in the training set, to accommodate the undirected nature of the interaction. For example, if (h_1, v_1) is compared in the classification process to (h_2, v_2) , the human protein h_1 needs to be compared to the human protein h_2 , and similarly comparing the two viral proteins v_1 and v_2 . When (h_1, h_2) and (h_3, h_4) are compared in the intra-species case, four comparisons are needed (h_1 with h_3 and h_4 , and similarly, h_2 with h_3 and h_4). This is why the second kernel (Equation 4.3), which makes the four comparisons, performed the best in the single organism prediction [91], but not as well in the virus-human case (Table 4.1). The first kernel (Equation 4.2), which makes direct comparisons only (h_1 to h_2 and v_1 to v_2), achieves better performance in the human-virus case, for it does not overload the model with unnecessary connections. This is also apparent in the case of the third kernel (Equation 4.4), which performs better than the second in the virus-human context as well for it performs the peer comparisons only; see Table 4.1 [91].

4.4.4 DeNovo-Human advantage

The advantage DeNovo-Human proposes over previous solutions is severalfold. DeNovo allows prediction from only a small seed of the query viral protein PPIs (Figure 4.4), instead of the large number of PPIs (typically hundreds) needed in single-virus applications, making it possible to predict for any viral protein once such a seed is available. Although my previous solution DeNovo does not require any PPIs to predict for a novel viral protein, this comes at the cost of significantly reducing the testing library size (Figure 4.1).

DeNovo-Human only uses protein sequences without additional prior knowledge, which is suitable for less-studied and novel viruses. By pooling PPIs from different viruses, DeNovo achieves higher prediction accuracy than that achieved by training on single virus PPIs

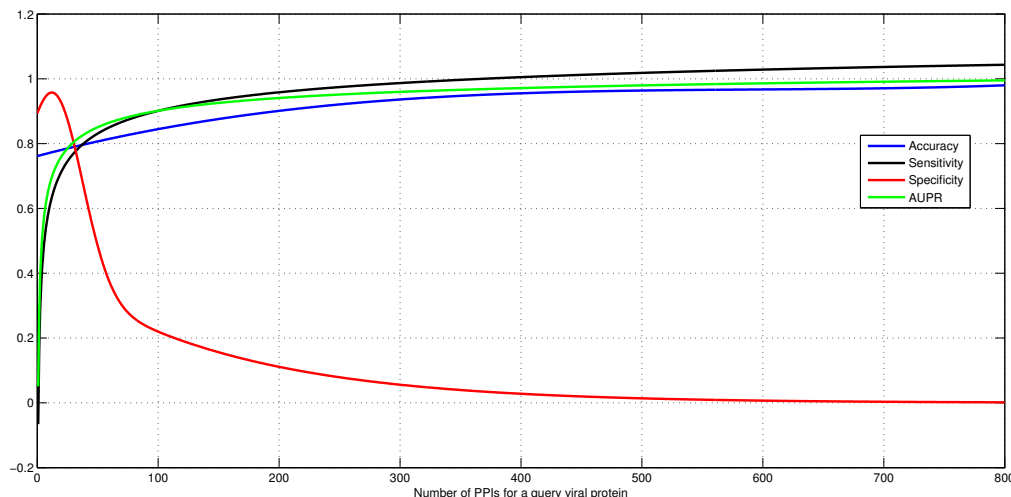


Figure 4.4: Estimated performance of DeNovo-Human with different number of PPIs given for a query viral protein. Accuracy, sensitivity, specificity, and AUPR are fitted to generate the tendency curves, with R^2 of 0.75, 0.78, 0.82, and 0.78, respectively. The balanced point occurs when a query viral protein has about 30 known PPIs with human proteins fed into DeNovo-Human. One can select different operating points according to their application tolerance and constraints.

(using the same features), while providing a larger human protein library to test against [6, 21]. I further used the same learning model of DeNovo-Human (but without pooling) to test the prediction performance of the single virus case using 37 virus-human networks for larger comparison quantification (Figure 4.3). The pooling allows DeNovo-Human to learn more discriminating features of the interactions between human and viral proteins. The average performance of my model for all viral proteins from sequence alone (see Table 4.1) is better than the single-virus-human PPI prediction models that use only protein sequences in extracting the features; see Section 4.4.2.

Not only does pooling achieve better accuracy than learning for single viruses using sequence features, it also achieves comparable accuracy to the models working on a single virus prediction using protein features more sophisticated than sequence (Section 4.4.2). The frameworks that learn to predict HIV protein interactions with human proteins take advantage of different sources of prior knowledge about the virus proteins. Features driven from GO terms, structural similarities, interaction interfaces, and co-localization are used in delivering more accurate performance [29, 33, 99, 113]. However, such prior knowledge is not available for almost all the other viruses. Experimentally collecting them is a time consuming and expensive process, making those methods not applicable for prediction for almost all other viruses.

4.4.5 Virus-human PPI network

I used DeNovo-Human to predict the entire virus-human interaction network. I further used the predicted network to estimate the positive to negative ratio for viral proteins connectivity in the network (how many human proteins a viral protein is connected to, in average). The estimated ratios are $1:5.90 \pm 0.0$, $1:5.87 \pm 0.0$, and $1:15.01 \pm 0.6$ for average, voting, and the strict networks. I used the same approach with human-human PPI network collected from String DB in December 2016 to estimate the positive to negative ratio for human PPI networks with the number of nodes in the subnetworks sampled 4000 times. The estimated ratio for the human PPI network case is $1:42.36 \pm 0.1$.

4.4.6 Redundancy

Does the performance improve if the model is trained on viral proteins similar in sequence to the query one? The results in Table 4.2 for the first similarity test show that a viral protein does not need to have similar proteins in the training set to have its PPIs predicted accurately, as long as a few of its own PPIs are used in training the model. The 40% similarity threshold corresponds to viral proteins from different families, which indicates that, even if a novel viral protein comes from a new viral family that has no other known PPIs, prediction for that viral protein is still almost as accurate as when there are several viral proteins in training with high similarities to the tested one. This means that once a few PPIs are available for a viral protein, accurate prediction can be generated for its interactions with the human protein library.

In the first similarity test above, I found that a viral protein does not need PPIs of related viral proteins in the training set to have its interactions predicted accurately as long as it has some of its own PPIs in the training set. In Similarity Test 2, I ask whether the presence of some viral proteins in the training that have some similarity to the viral protein v in testing can relax the condition of feeding the model with PPIs of v . I test whether the model can predict PPIs of a viral protein v when it has training examples for a set of viral proteins similar to v . The results (Table 4.3) suggest that some PPIs for each protein is needed to train the model in order to make a good prediction for that viral protein. Even 90% sequence similarity cannot compensate for the absence of those own PPIs. The model can recognize a true interaction only one-third of the time. Approximately 30 PPIs are needed to achieve a balanced performance for each query viral protein (Figure 4.4).

DeNovo-Human results challenge two notions in the PPI prediction domain. First, it demonstrates that prediction for virus-human PPIs is easier (or at least not more difficult) than predicting in the intra-species case, contrary to the mainstream notion. The performance obtained for the virus-human case outperforms the human-human case with approximately 10% when the same model is used [91]. Second, I used the framework to demonstrate that, at least in the virus-human case operating in the C1 mode, redundancy in sequence similarity

does not have a significant effect on the prediction quality; see Tables 4.2 and 4.3.

4.4.7 Predicting for a query viral protein

DeNovo-Human needs the sequence of a new viral protein (query protein) and few of its PPIs to predict which of the human proteins in the library can interact with that query. DeNovo-Human adds the query's PPIs to the data set, uses the entire data set for training, generates random negative PPIs of ratio 1:1, and trains the model with those positive and negative PPIs. This process is repeated 40 times to reduce the noise effect of the random negative sampling. Then the query viral protein is paired with all the human proteins in the library and the trained model is used to test whether they can interact or not. Human proteins that are predicted to be able to interact with the query viral protein are returned with the average confidence scores of the predicted interactions. With a PPI seed of size 30 for a query viral protein, DeNovo is estimated to achieve balanced performance measures around 80%, see Figure 4.4.

4.5 Conclusion

The DeNovo and DeNovo-Human frameworks address two closely related problems. DeNovo predicts PPIs of proteins of a novel virus with human proteins with constraints on the human testing library size, but it does not need any known PPIs for a query viral protein. DeNovo-Human predicts PPIs of a viral protein with few known interactions with human proteins and achieves considerable accuracy without the coverage problem. If a few PPIs can be obtained for a novel virus, then thousands of human proteins can be tested. However, if no PPIs can be obtained, a future possibility is to use DeNovo to generate a small seed of predicted interactions for the query viral protein, then the seed can be fed into DeNovo-Human training, thus bypassing the limitation of DeNovo-Human.

Chapter 5

Case Study: Zika Virus

5.1 Introduction

A global health emergency state was declared in 2015 in response to the breakout of the Zika virus (ZIKV) in the Americas. ZIKV is associated with microcephaly [78] and has the potential to spread worldwide via mosquito vectors. In response to that threat, ZIKV has jumped into the focus of many research labs in attempts to understand its pathogenesis, to determine why and how it causes birth defects, and hopefully to develop vaccines and drugs to overcome the virus. Unfortunately, at the time ZIKV re-emerged, little was known about it, requiring much scientific effort to fill this knowledge gap [68]. For example, no protein-protein interactions (PPIs) were known between ZIKV proteins and human proteins (recently, a few PPIs have been published [68]).

ZIKV is a positive-sense, single-stranded RNA virus species from the *Flaviviridae* family, *Flavivirus* genus [96]. *Flavivirus* includes some mosquito-transmitted viruses, such as the Dengue virus (DENV), the West Nile virus, and the yellow fever virus [74, 96]. The ZIKV genome polyprotein is cleaved into three structural proteins (C, M, and E), and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5); see Figure 5.1. ZIKV enters a human body through a mosquito bite, fights the immune system, travels to the fetal brain (in pregnant women), enters the fetal neural stem cells (fNSCs), and causes microcephaly [57].

I asked how my work can be of help in a re-emerging virus case that has previously received little study attention but now is an alarming threat. The Zika virus (ZIKV) is a classical example of such a case. ZIKV has been known for a long time, but only when it represents a threat do we start to fight it, but without much knowledge about its behavior and pathogenesis. Starting from ZIKV sequences only, I tested ZIKV proteins against human proteins to find about their ability to interact. I used the list of potential interactions to target interesting questions about ZIKV behavior. Using the two DeNovo models developed

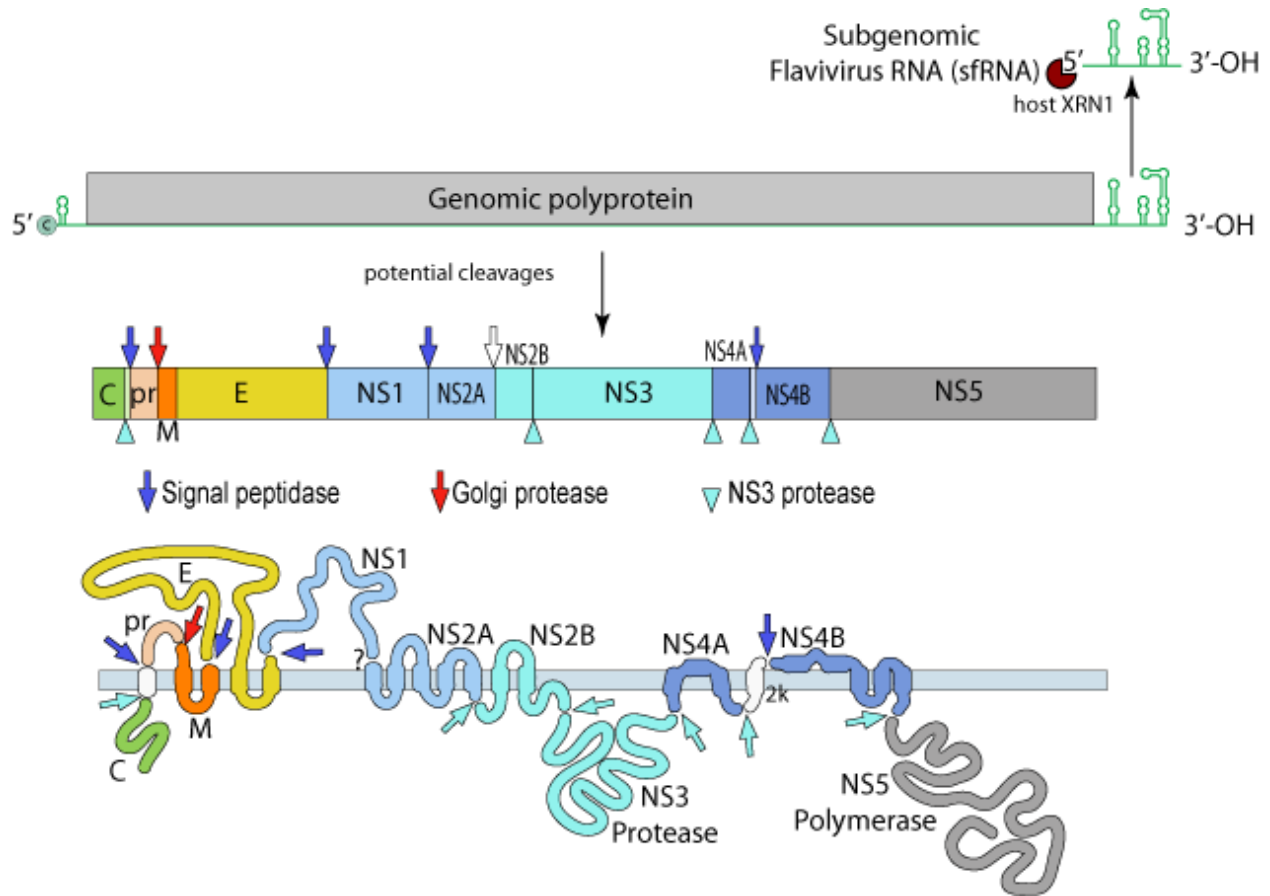


Figure 5.1: Zika virus genome and proteins (Source: ViralZone:www.expasy.org/viralzone, SIB Swiss Institute of Bioinformatics) [47]

in this thesis, a set of proposed answers can be obtained. It is for the biologists to exclude some and experiment with what they see as potential answers for verification. I further verified some of these answers by recently published studies. The goal of this case study is to demonstrate how the frameworks presented in this thesis are woven together to enrich our knowledge about virus behavior when only its sequence is available.

5.2 Research Questions

The research questions in this case study focus on revealing ZIKV behavior by examining predicted PPIs between ZIKV and human proteins and adding domain knowledge. Basically, the study answers which cell receptors may facilitate the virus entry, how the virus fights the interferon circuitry, and how it causes microcephaly. Additionally, I examined why Dengue virus does not cause microcephaly even though it is closely related to Zika virus.

Research Questions 1: Which cell entry receptors help ZIKV? ZIKV belongs to the viral family *Flaviviridae*, genus *Flavivirus*. A flavivirus enters the cell via receptor-mediated endocytosis or apoptotic mimicry, where the E protein of the virus binds to a cell receptor to enter the cell [96]. Here I want to identify which cell receptors help to internalize ZIKV into the cell. I use a cascade of my prediction models, DeNovo and DeNovo-Human, to identify which cell receptors ZIKV protein E can interact with. See Section 5.6.2.

Research Question 2: How does ZIKV fight the interferon machinery? Interferon (IFN) induction is the front-line of host defense against invading viruses. Viruses have evolved attacking strategies to suppress the antiviral state initiated and maintained by the IFN circuit, to be able to replicate in the cell without being degraded and to dampen the adaptive immune response [66]. The IFN circuit has three phases: IFN induction in response to recognizing pathogen-associated molecular patterns (PAMP), IFN signaling, and an antiviral effector phase led by many products of IFN-stimulated genes (ISGs) [98]. Each of these phases activates different signaling cascades in the cell. I asked which of these pathways is likely to be taken down by ZIKV proteins and at which points along the cascades ZIKV is most likely to interfere. See Section 5.6.3.

Research Question 3: How does ZIKV cause microcephaly? If a pregnant woman is infected with ZIKV during the first trimester, there is a considerable chance her fetus will develop microcephaly, causing the newborn's head to be smaller than normal due to brain development defects [78]. Here I want to investigate how ZIKV might be causing microcephaly, using only the ZIKV protein sequences. To use my PPI prediction models in this investigation, the research question can be stated as: Which PPIs of ZIKV proteins with human proteins cause microcephaly? See Section 5.6.4.

Research Question 4: Why does Dengue virus, a closely related virus to ZIKV, not cause microcephaly as well [42]? While the Zika virus causes microcephaly, other viruses in the same genus, *Flavivirus*, such as the Dengue virus (DENV), do not cause that disorder [57]. Specifically, I hypothesize that DENV does not cause microcephaly because it cannot produce the same effects that lead to microcephaly in the ZIKV case (which I examine in the third question). This test is a good chance to demonstrate the comparative power of the two DeNovo models; can they catch the difference in PPIs while the genomes of the two viruses have high sequence similarity? See Section 5.6.5.

5.3 Data

ZIKV proteins were retrieved from the most recent strain available at the time of the study (Rio-S1, sampled in January 2016) [97]. This strain is chosen to be sampled from Brazil so that the viral sequence is one of those associated with (or as close as possible to) the current ZIKV outbreak in Brazil. I also retrieved the sequence of the most recent isolated strain in Brazil for DENV (R/SJRP/2271/2014, isolated on February 10, 2014) from VIPR [97].

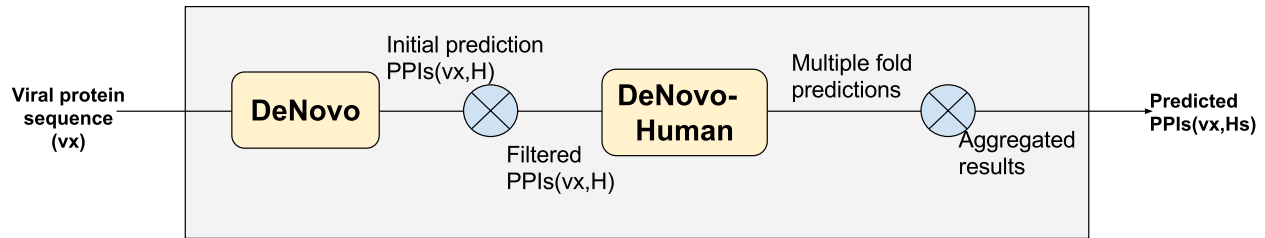


Figure 5.2: DeNovo Cascade. Prediction throughout this case study is prepared by cascading the two DeNovo models developed earlier. Positive and negative highly confident predictions from DeNovo (for a viral protein v_x with no known PPIs) are used as seed PPIs fed into DeNovo-Human.

I used human receptor proteins previously known to mediate some virus entry into human cells as potential receptors for testing against the ZIKV envelope protein (E protein). I retrieved the 269 protein identifiers from GO:0001618 (virus receptor activity, defined as “combining with a virus component and mediating entry of the virus into the cell”) [48]. Receptors were then filtered to exclude the non-human receptors. A set of 71 human entry receptor proteins remain. As we need examples of each human protein interactions to operate any of the DeNovo models, 26 out of the 71 human entry receptors are excluded, as they have not been reported as interacting with specific viral proteins. Moreover, the definition of GO:0001618 does not necessarily imply that each human protein reported is a receptor interacting with viruses. The list of the remaining 45 human proteins used in the prediction is the union of the null prediction and the average prediction sets in Table 5.1.

The interferon circuit pathways were aggregated from multiple sources [52, 66, 118]. The key human proteins participating in the IFN circuit and previously reported to interact with other viral proteins were labeled. The test set is thus composed of 41 human proteins participating in one or more signaling cascades in the IFN machinery. The mTOR pathway was also processed in the same way; it was collected mainly from KEGG [52] and extended from other sources [98, 100]. Multiple cascades triggered by irrelevant signals were not considered. The key human proteins previously reported to interact with viral proteins were labeled for testing against ZIKV and DENV proteins.

5.4 DeNovo-Cascade

DeNovo-Cascade works as follows. A set of highly confident positive and negative interactions of the viral protein of interest is generated using DeNovo, then fed into DeNovo-Human to produce a broader prediction against a larger set of human proteins. This process involves four steps: initial prediction using DeNovo, cascading DeNovo and DeNovo-Human by preparing a high-quality DeNovo prediction set to pass to DeNovo-Human (seed), performing prediction using DeNovo-Human, and finally aggregating the DeNovo-Human results. See

Figure 5.2.

The DeNovo model is trained with a PPI set between human and different viral proteins with different dissimilarity thresholds as discussed in DeNovo [35] (Chapter 3). The sequence of the viral protein of interest v is paired with all human proteins used in training using the feature extraction scheme of DeNovo. The trained models are then used to predict how likely it is that there are interactions between v and that library of human proteins.

The predictions from the different DeNovo models (at different dissimilarity threshold values) are aggregated to obtain two sets of human proteins: one that is highly likely to be able to interact with v and the other that is highly unlikely. For simplicity, I use PPI^+ and PPI^- to refer to the two groups.

The models considered here for prediction aggregation are those with dissimilarity thresholds of 0.5, 0.6, and 0.7. These thresholds are not the same as the suggested optimal threshold T^* in DeNovo ($T^* = 0.8$). T^* is selected theoretically to remove all possible false negative interactions from training. However, using T^* in negative sampling of the current data set introduces a new source of bias. Some proteins appear much more frequently as negative partners than the others at $T = 0.8$ and $T = 0.9$, even with the outliers removed, causing them to be predicted as negative partners all the time. The highly confident predictions (as positive or negative PPIs) are selected by considering only the pairs that do not fall into the margin around the separating hyperplane of the trained SVM model. This is achieved by considering the pairs with a distance of at least 1 (governed by the C parameter value) for positive PPIs or -1 for negative PPIs from the hyperplane.

The positive and negative PPIs predicted from DeNovo for a viral protein v (PPI^+ and PPI^-) are combined with the set of known virus-human PPIs and generated negative PPIs (for that known set) to train the DeNovo-Human model. These two (PPI^+ and PPI^-) sets are dealt with as if they are known positive and negative PPIs for v because they are selected to be of high prediction confidence from DeNovo. A subset of the predicted PPIs is randomly selected for multiple rounds of training to reduce the effect of any noise that might be present in the initial DeNovo predictions. The viral protein v is then paired with all the human proteins used in training DeNovo-Human (using the feature scheme of DeNovo-Human). The trained DeNovo-Human models are then used to predict how likely to interact those pairs are.

The results are aggregated on three different bases: complete agreement (all-1's for all the models, strict prediction), confidence averaging, and majority voting. In complete agreement (strict) aggregation, a tested pair is considered likely to engage in a positive PPI if all the prediction models (trained with different negative samples) predict the pair as positive. In averaging aggregation, the mean of the confidence scores produced by the different models is what determines whether the pair is considered a positive PPI based on the conventional threshold of the prediction model. As I use a LibSVM implementation, the threshold is 0. Majority voting aggregation considers a pair as a positive PPI if more than half of the models predict it as positive.

Category	Human proteins
Strict prediction	LDLR, CD4, ITGB1, HSPA1A, HSPA1B, IDE, ITGA2, MRC1, SCARB2, ACE2, CLEC4M, CD209
Average prediction	LDLR, CD4, TFRC, ITGB3, ITGB1, CD55, HSPA1A, HSPA1B, IDE, ITGA2, ITGB5, MRC1, DPP4, AXL, CD80, CD86, EFNB2, SCARB2, ACE2, CLEC4M, CD209
Null prediction	NPC1, SIVA1, CLDN1, ICAM1, ITGAV, ITGA5, RPSA, LAMP1, PVR, CD46, CCR5, CXCR4, CXADR, HYAL2, SLAMF1, DAG1, NECTIN1, SLC1A5, EFNB3, SCARB1, NECTIN2, TNFRSF14, NECTIN4, F11R

Table 5.1: Predicted interactions of Zika virus protein E with human cell entry receptors. The strict prediction reduced the search space (45 human cell receptor proteins previously reported to internalize some viruses) to 12 candidate receptors, while the average prediction reduced it to 21.

For the ZIKV case study, I considered the complete agreement (all-1's) criteria when strict prediction is needed and the averaging criteria for a more relaxed prediction. For each question about ZIKV involving one (or more) of the ZIKV proteins with a specific set of human proteins, the whole DeNovo-Cascade four steps are run, then the prediction in the last step is filtered to contain only those human proteins of interest.

5.5 Results

Out of the 71 human cell entry receptor proteins, only 45 could be tested (for they have known PPIs with some viral proteins). Out of the 45 tested ones, the ZIKV E protein is predicted to be able to interact with 12 and 21 entry receptors in the strict (all 1's) and average prediction settings respectively. A list of the predicted receptors is provided in Table 5.1.

The interactions predicted between ZIKV proteins and the members of the IFN circuit are summarized in Figure 5.3 and Table 5.2. Those predicted between ZIKV proteins and Dengue virus proteins and members of the mTOR pathway are summarized in Figure 5.4 and Table 5.3.

Human Protein	Zika proteins
TRAF3	E,NS3,NS4A,NS4B,NS5
TBK1	C,E,NS2B,NS3,NS4A,NS4B,NS5
IRF3	C,M,NS4B
RIP1	E,NS3,NS4A,NS4B,NS5
TRAF6	M,E,NS3,NS4A,NS4B,NS5
TAK1	
MKK7	C,E,NS3,NS4A,NS4B
MKK3	E,NS3,NS4A,NS4B
JNK	NS4A
P38	E,NS3,NS4A,NS4B
AP-1	E,NS4A
IKKA	E,NS3,NS4A,NS4B,NS5
IKKE	NS3,NS4A,NS4B
IKKB	E,NS2B,NS3,NS4A,NS4B,NS5
IKKG	E,NS3,NS4A,NS4B,NS5
ERK	NS4A
IKBA	E,NS3,NS4A,NS4B,NS5
NFKB1	C,M,NS3,NS4A,NS4B
NFKB2	C,M
TLR7	E,NS3,NS4B
MYD88	C
IRAK1	C,M
IRF7	C,M
MDA-5	C,E,NS2B,NS3,NS4A,NS4B,NS5
RIG-I	E,NS2B,NS3,NS4A,NS4B,NS5
MAVS	C,M
TRIM25	E,NS3,NS4A,NS4B,NS5
TANK	E,NS3,NS4B
NAP1	
FADD	
CASP8	E,NS2B,NS3,NS4A,NS4B,NS5
CASP10	E
TRAF2	C,M,E,NS2A,NS3,NS4A,NS4B,NS5
IFNAR2	C,M,E,NS4A,NS4B,NS5
JAK1	M,E,NS2B,NS3,NS4A,NS4B,NS5
TYK2	C,M,E,NS3,NS4A,NS4B,NS5
STAT1	E,NS3,NS4A,NS4B,NS5
STAT2	E,NS3,NS4A,NS4B,NS5
SOCS1	M
p58IPK	E,NS3,NS4A,NS4B
PKR	E,NS3,NS4B,NS5

Table 5.2: Predicted interactions of Zika virus proteins with the interferon circuit proteins.

mTOR protein	Zika virus	Dengue virus
AKT1	NS3, NS4A	
AKT2		
AKT3	NS3	
PI3K	C,E,NS3,NS4A,NS4B,NS5	NS3,NS4A,NS4B,NS5
TSC1	E,NS3,NS4A,NS4B	NS3,NS5
TSC2	C,M,NS4A	
TBC1D7		
AMPK	NS4A	
MTOR	C,M,E,NS3,NS4A,NS4B	NS3,NS4A
R	C,M,E,NS3,NS4A,NS4B	NS3,NS4A
GRB2	E,NS2B,NS3,NS4A,NS4B	NS3,NS4A,NS4B,NS5
HRAS		
NRAS		
RAF	C,M,E,NS3	NS3
MEK1		
MEK2		
ERK1		
ERK2	NS4A	
PTEN	E,NS3,NS4A,NS4B	NS3,NS4A,NS5
IFNAR2	C,M,E,NS4A,NS4B,NS5	
IRS1	C,M	C
PDK1	C,M,E	NS3
IKKA	E,NS3,NS4A,NS4B,NS5	NS3,NS5
TYK2	C,M,E,NS3,NS4A,NS4B,NS5	C,NS2A,NS3,NS4A
JAK1	M,E,NS2B,NS3,NS4A,NS4B,NS5	NS3,NS4A,NS4B,NS5

Table 5.3: Predicted interference of Zika virus and Dengue virus with the mTOR pathway, which is a key pathway whose interruption in the fetus may cause microcephaly.

5.6 Discussion

5.6.1 Viral perturbation to human signaling pathways

Viruses depend on the cellular machinery of the host to survive their life cycles [24]. The virus-host interplay is mediated by many interactions between the virus and host cellular proteins. On the other hand, these interactions have driven the evolution of the compact viral genomes to adapt to the virus-host interplay [24]. Viruses evolve shared mechanisms to interfere with host cellular pathways. They target different pathway types, particularly the signaling pathways, whose subversion will help viruses hijack cellular regulation for their

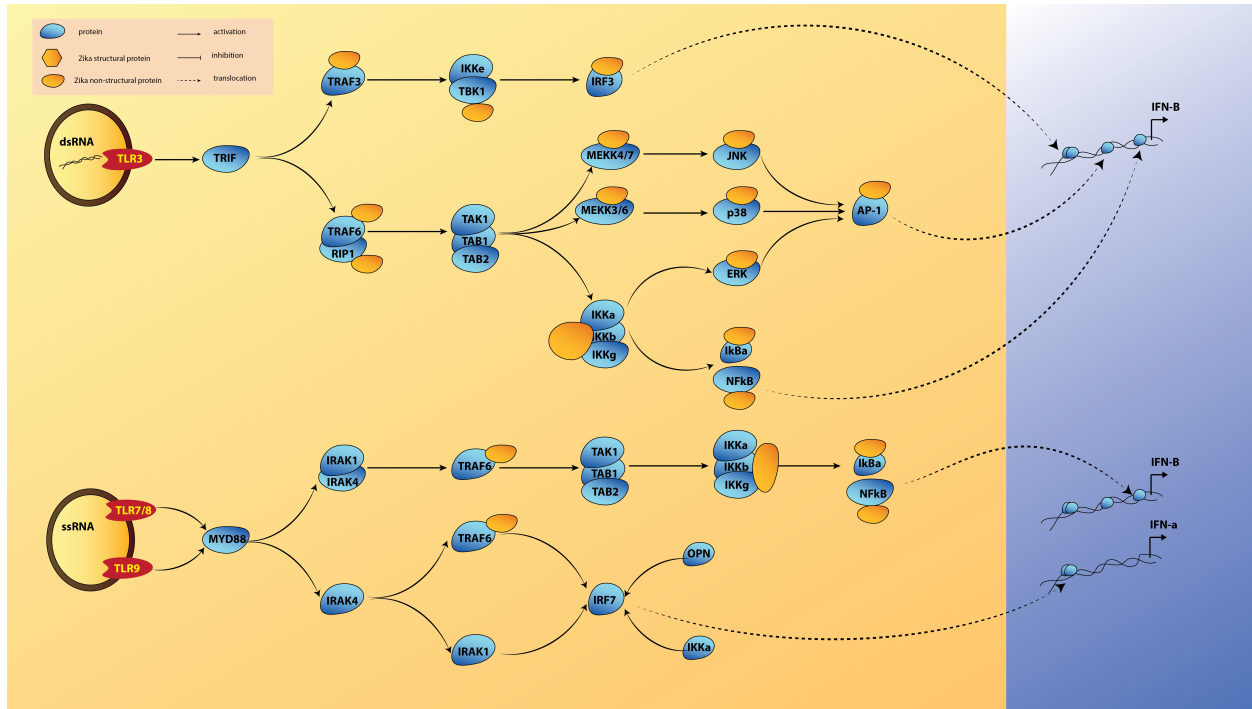


Figure 5.3: Summary of Zika interference with IFN pathways.

benefits [127].

Given that viruses have compact genomes (and consequently constrained proteins), how can their proteins interact extensively with multiple cell pathways and processes [24]? Now it is believed that most of these virus-host interactions are mediated with mimicry of short linear motifs (SLiMs) of the host proteins [24]. Given that human hub proteins also interact mostly through SLiMs, mimicry of their interfacing SLiMs can reinforce the ability of a virus to control a wide spectrum of cellular functions and processes [120]. This aligns with the observation that viral proteins tend to target host hubs and bottlenecks [127].

Some studies have attempted to deduce and quantify some of the rules of how viruses perturb human pathways from different points of view [32, 50, 127, 128]. Here, I give the reader a short account of my observations analyzing virus interference with human signaling pathways presented in KEGG. I observed that a virus is likely to target pathways of interest through interactions with multiple cellular proteins along each signaling cascade. Moreover, a virus may target a cascade with multiple viral proteins. This observation is also supported by the quantification study in [127].

I also observed that virus-host PPIs over a single pathway are not of equal strength in subverting the pathway. We can classify the interference (through PPI) into primary and secondary types. In primary interference, a virus targets one or more hubs of the pathway, so it controls the pathway with few interactions, for a pathway hub mediates several signals at

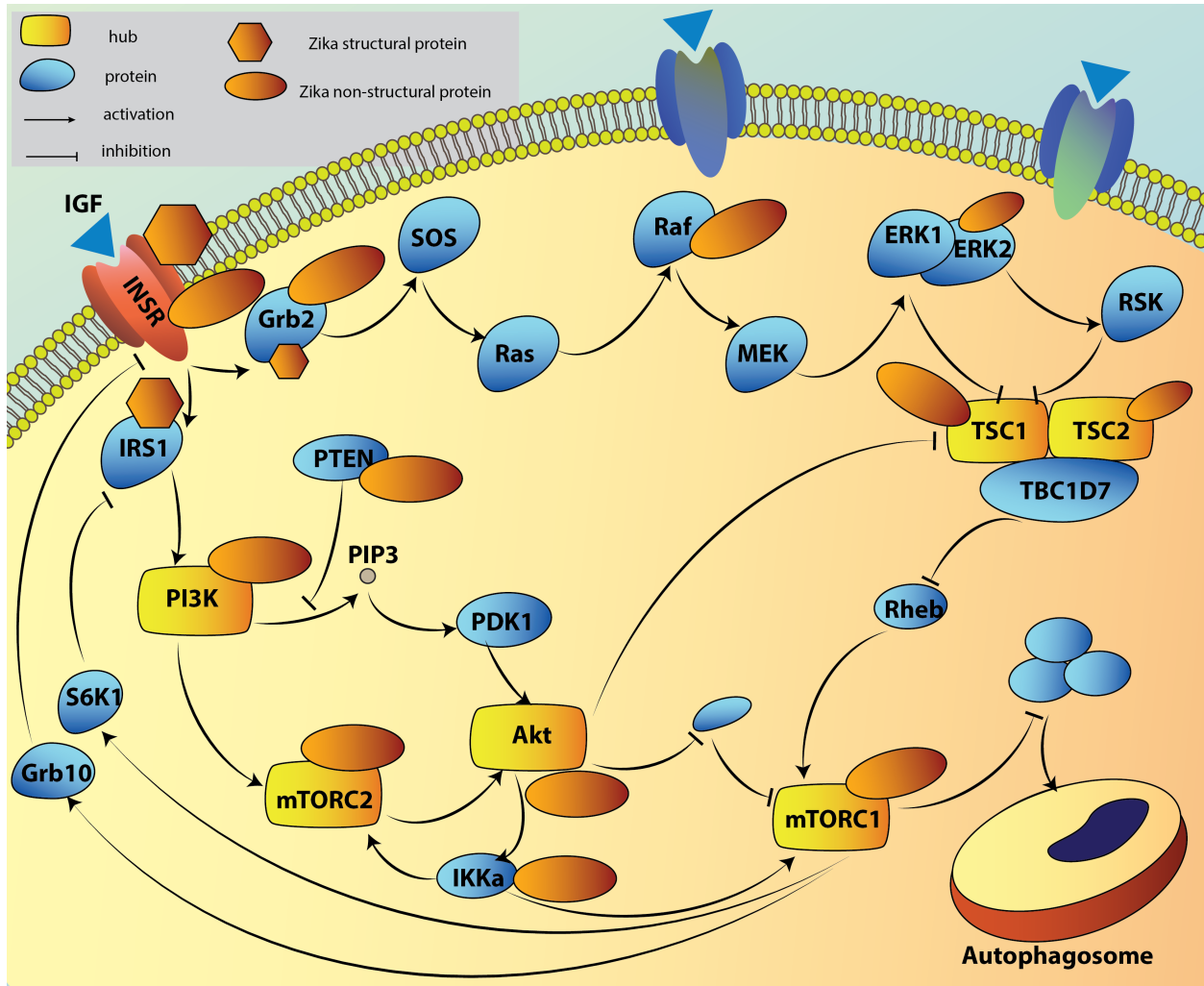


Figure 5.4: How Zika virus causes microcephaly. ZIKV predicted interactions with the mTOR pathway show high interference potential, especially by ZIKV NS4A. The predicted suppression of mTORC1 signaling should lead to increased autophagy in fNSC, which in turn can lead to neurogenesis defects and consequently fetal microcephaly.

once. On the other hand, to take down a pathway without involving a hub, in the secondary scenario, a virus must interact with members of several signaling cascades to prevent signal propagation or force its own signal to propagate.

ZIKV, as with any other human virus, must be interacting with human cellular pathways to attenuate the IFN circuit and to cause the disorders associated with its interactions. In answering the research questions of this case study, I investigate how ZIKV interacts with the different parts of the IFN circuit. I also examine its interactions (as well as Dengue virus interactions) along the pathways that (most probably) lead to fetal microcephaly.

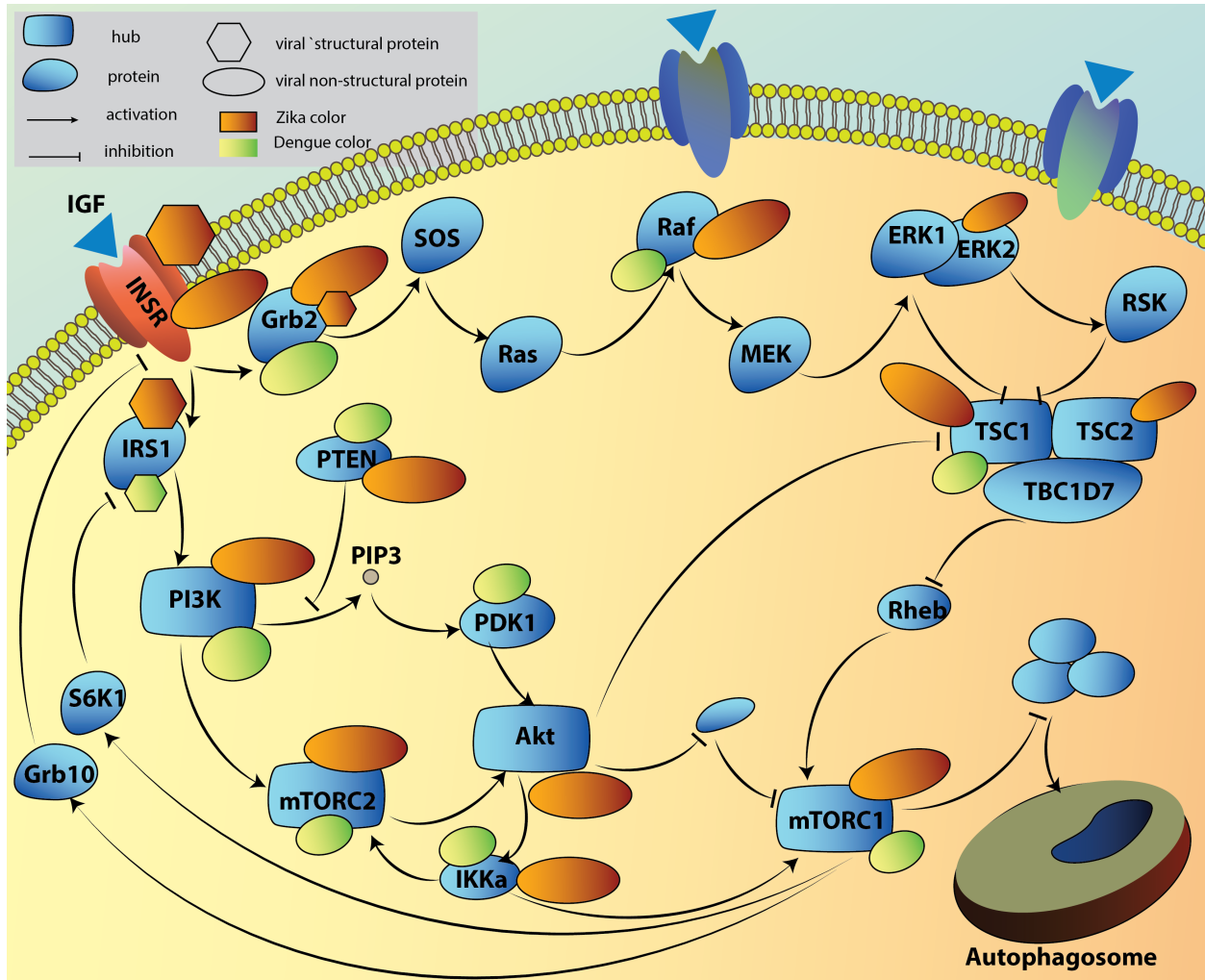


Figure 5.5: Dengue virus (DENV) potential interactions with the mTOR pathway demonstrate its interference capabilities with mTOR signaling and its potential to cause autophagy. The difference in ZIKV and DENV predicted PPIs suggests that either DENV cannot efficiently block the IGF to maintain upregulated autophagy or that DENV can induce autophagy but cannot reach the fetal cells at the first place to cause the downstream microcephaly.

5.6.2 Research Question 1

Which cell receptors are responsible for ZIKV entry? As a flavivirus, ZIKV enters the cell via receptor-mediated endocytosis, where the Envelope protein (E) binds with an entry receptor [96]. E is the major surface protein, which is responsible for directing the virus particles to host cell receptors and then fusing the viral and cell membranes to mediate the virus entry. Here I want to identify which cell entry receptors help ZIKV. Using DeNovo-

Cascade, I tested the ZIKV E protein against the human receptors previously identified as entry receptors for at least a single virus and having at least one PPI known with a human protein.

The strict and averaging predictions (between ZIKV E protein and the 45 candidate human cell entry receptors) suggest that 12 and 21 proteins play the role of entry receptors for ZIKV. The strict prediction includes DC-SIGN and the averaging prediction includes the DC-SIGN and AXL receptors. DC-SIGN and AXL are two of three receptors recently confirmed to mediate ZIKV cell entry in skin cells [45]. The third receptor, Tyro3, is not considered for testing here as it has no PPIs with any viral protein reported. AXL is now strongly thought to mediate the entry of ZIKV into neural stem cells as well [87].

Although the reduction in the candidate receptors for ZIKV is not dramatic (33 and 24 are excluded out of 45 candidates in the two prediction sets), it is still interesting that more than half of the candidate proteins are excluded. A biologist may prefer to experiment with a reduced set of candidate receptors for cost and risk considerations. Such a reduction can be more appreciated given that the identification of flavivirus entry receptors is an experimental challenge to date [96]. The reduction of the candidate entry receptor list by half using my prediction is thus a considerable achievement.

It is useful to note that, after all, the prediction means that the viral protein can interact with the predicted human protein if there is no interaction barrier, like localization or blockage of the interaction interface. On the other hand, not all the receptors computationally tested here were examined experimentally [45], thus the predicted receptors may present an opportunity for further investigation.

It is even feasible to down-size the predicted receptor list to the ones expressed in placental cells (thought to mediate passage of ZIKV through the placenta) to learn more about ZIKV crossing the placental barrier. Also, down-sizing to those receptors expressed in fetal neural stem cells (fNSCs) or other known (or candidate) ZIKV tropism (human cells and tissues supporting the virus growth) can shed light on which cells might be susceptible to ZIKV infection [57].

Still, the presence of other virus-dependency factors in the host cell is essential for viral replication, even if the cell has the necessary receptors to internalize the virus [57]. Thus, receptor prediction is sufficient to down-size the search space, but still lab work or more prediction experiments and further reasoning are needed to account for factors not explainable by PPI predictions.

5.6.3 Research Question 2

How does ZIKV fight Interferon (IFN)? Interferons (IFNs) are widely expressed cytokines with potent antiviral effects that form a first-line of defense against viral infection [98]. The IFN circuit is a three-phase cascaded amplifier. The first phase is triggered by the

recognition of a virus (or a pathogen) nucleic acid or proteins by Toll-like receptors or other cytoplasmic sensors like RIG-I and MDA-5 [118]. The second phase is IFN transcription, which takes place by the propagation of the recognition signal to activate transcription factors that in turn translocate to the nucleus to drive IFN-I transcription [118]. In the third phase, the antiviral effector phase, hundreds of IFN-stimulated genes (ISGs) are upregulated in response to the stimulation of IFN receptor (IFNAR) upon binding of the secreted IFN in the same or nearby cells and activation of STAT1/2.

Almost all viruses interfere with the IFN circuit at different phases in an attempt to take down the innate immune circuit before the adaptive immune circuit is launched to threaten the virus survival and replication [118]. However, viruses adopt different strategies and interference mechanisms to attack the IFN circuit activities. Understanding the virus-IFN subversion mechanism is key to designing live vaccines and antiviral drugs.

I asked how Zika virus antagonizes the IFN circuit. Here I use my models to specify which pathways ZIKV targets in its effort to take down the innate immune system. I thus used DeNovo-Cascade to test ZIKV proteins against the members of the pathways involved in the three phases of the IFN circuit.

Examining the prediction results, I found that ZIKV heavily targets the cascades activated by TLR3 but not those by TLR7, TLR8, or TLR9, except for shared proteins. Some predicted PPIs with members of non-TLR3 triggered cascades are mediated by ZIKV structural proteins that most likely cannot be co-localized with the members of those cascades; see Figure 5.3. This is an interesting observation since ZIKV is found experimentally to activate the TLR3-mediated innate immune response [23].

Recognition of a pathogen by TLR induces IFN- β , while recognition by cytoplasmic sensors (MDA-5 or RIG-I) induces both IFN- β and IFN- α through different cascades. ZIKV is predicted here to extensively interact with the IFN- β induction cascades (triggered by TLR and some cytoplasmic sensors) but not the IFN- α cascades (except for some shared proteins with other targeted cascades and structural protein predicted interactions). This observation aligns with the recent finding that ZIKV induces TLR3, RIG-I, and MDA-5, beside ISGs associated with IFN- β expression [45].

On the IFN third phase of ISGs transcription, multiple ZIKV proteins, including NS5, are predicted to have the ability to interact with JAK-1, TYK2, STAT1, and STAT2. Indeed ZIKV NS5 is found to bind to STAT2 to inhibit type I IFN signaling [42]. More interestingly, ZIKV is predicted to have a strong potential to bind to IFNAR2 (one of the IFN-I receptor two subunits), both with structural and non-structural proteins. If such interaction is verified *in vivo*, it can be a strong attempt to block the interferon signal propagation by cutting the signal at the source (the interferon receptor).

Besides being supported by experimental evidence, the above findings correlate with the precise evolution of the virus to meet its needs with optimization of its compact genome. An RNA virus, like ZIKV, has a small genome that must encode all functionality to hijack the



Figure 5.6: Reasoning about how Zika virus might be causing microcephaly. Going backward from microcephaly, it is caused by disturbing neurogenesis by autophagy. Autophagy, in turn, is induced by suppressing mTOR. Thus, ZIKV is probably causing microcephaly by suppressing mTOR in fetal NSCs.

host cell regulations and survive the host defense. This compactness along with selection pressure should concentrate the viral effort to fight its opponent proteins directly or indirectly and limit fights that are not necessary to virus survival. ZIKV here is predicted to interact extensively with those proteins found to have a direct fighting effect over the virus. DeNovo models can thus be used in the future to learn more about virus behavior by examining which cascades are heavily targeted. However, these comparisons need careful examination of human pathways, which for now is performed manually. I hope future researchers will help improve the accuracy of the DeNovo models and will consider the automation of pathway interference labeling.

5.6.4 Research Question 3

How does ZIKV cause microcephaly? Infection with ZIKV in pregnant women is correlated with fetal microcephaly [16, 20, 78]. Here I use DeNovo-Cascade to find out how ZIKV causes this brain disorder. Microcephaly, as well as other developmental neurological disorders, are caused by some defect in the differentiation of neural stem (or progenitor) cells; this process is called neurogenesis [41, 79, 123]. Neurogenesis defects are attributed mostly to induced autophagy [124]. Beside its catabolic role in degrading cellular damaged material, autophagy is evolved as a strong immune response against pathogenic infection by degrading the invading pathogen proteins [65]. Thus, several viruses work on suppressing autophagy to maintain replication. Surprisingly, some other viruses induce autophagy to aid their survival [46]. The mTOR pathway, and specifically signaling through mTORC1, is the major player in regulating autophagy [51], thus, it can act as the major player in causing microcephaly [64]. Please see the reasoning chain in Figure 5.6. From another perspective, mTOR signaling is proven to be a key pathway in brain development, especially the neuronal and glial differentiation, and maintaining the stemness of NSCs [64]. Disturbing the mTOR activity in NSCs leads to several brain disorders. Additionally, mTOR is explicitly linked to microcephaly [18]. A study containing another reasoning chain and referencing other experimental studies linking mTOR to fetus microcephaly has been recently published [68].

The mTOR pathway is activated by multiple nutritional and environmental signals, for example, energy defect and oxygen defect [63]. Most of these signals are irrelevant to our case study, thus, I assume their cascades do not have effect on ZIKV (even if ZIKV can

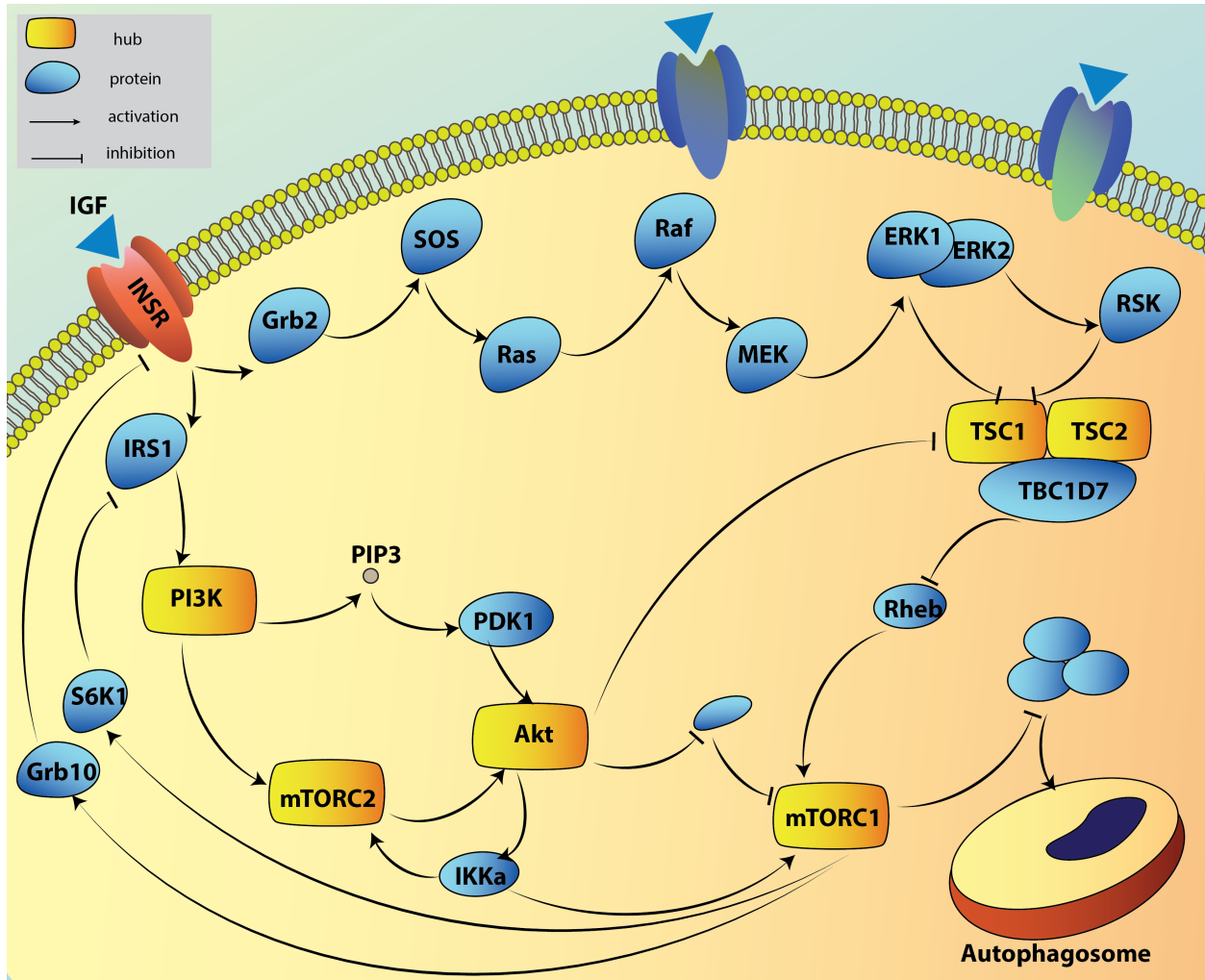


Figure 5.7: Relevant cascades in the mTOR pathway.

interact with them). Consequently, they are not considered for PPI prediction and further investigation of ZIKV interference with the mTOR pathway. On the other hand, the insulin-like growth factor (IGF) axis has a critical role in fetal growth [100]. IGF binds to insulin receptor (INSR) and activates mTOR in two cascades: one through the PI3K/Akt axis and the other through MAPK signaling, see Figure 5.4. Activation of mTOR by IGF should suppress autophagy. Thus, I focus on these two cascades in this study, see Figure 5.7.

As a result of mTOR activation through IGF, autophagy is downregulated to allow for developmental growth [100]. If we identify how ZIKV interferes with the mTOR pathway to down-regulate mTORC1 signaling, we can sketch scenarios of how ZIKV may cause microcephaly. We, thus, used DeNovo-Cascade to test ZIKV proteins against the key proteins in the mTOR relevant cascades.

My results show that ZIKV is capable of strongly inducing autophagy by interacting with

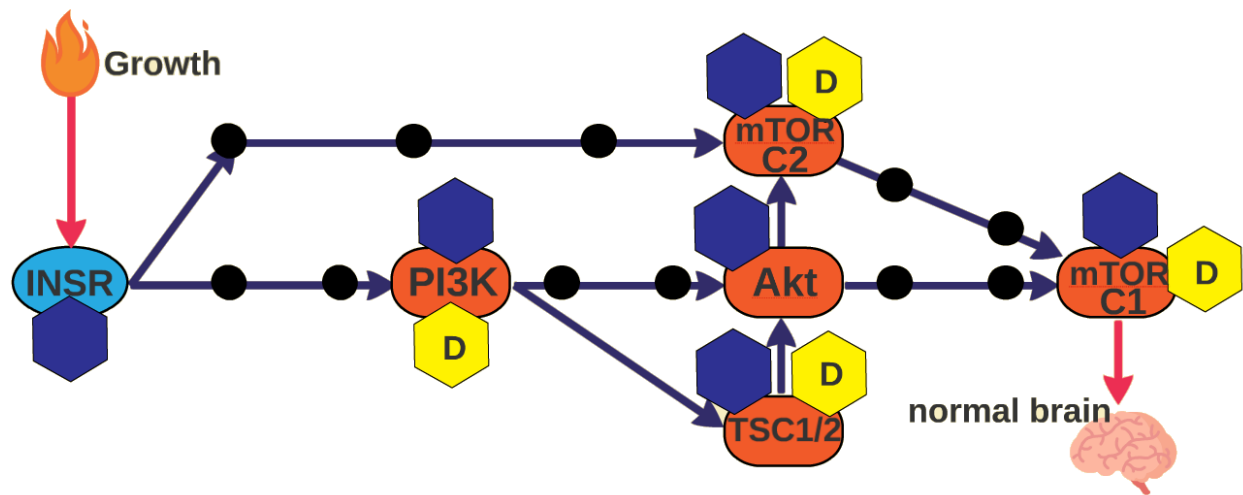


Figure 5.8: Simplified diagram of the key interference points of Zika virus and Dengue virus with the mTOR pathway. The blue and yellow hexagons represent proteins of Zika virus and Dengue virus, respectively.

multiple key proteins in the mTOR pathway; see Figures 5.4 and 5.8. This conclusion suggests that ZIKV needs autophagy for its replication. Following the above reasoning and published experimental evidence linking autophagy induction in fNSCs to microcephaly [64, 124] and, given the strongly supported observation that ZIKV crosses the placental barrier [16, 20, 78], this predicted interference to the mTOR pathway can be the mechanism by which ZIKV causes microcephaly.

I recognize multiple primary interference scenarios (strong, one-shot interference) of ZIKV with the mTOR signaling pathway triggered by IGF; see Figures 5.4 and 5.8. One primary scenario is the blockage of PI3K signaling by interacting with PI3K, upstream to Akt. This scenario would possibly shut down signaling through Akt that otherwise would activate mTORC1. A stronger primary scenario is the predicted ability of ZIKV to interact with Akt itself, for Akt is the central hub of the mTOR pathway. Another strong blockage primary scenario would be to interact with mTOR itself to prevent its activation and complex formation. A fourth primary scenario is the ability of ZIKV structural and some non-structural proteins to interact with the insulin receptor (INSR) itself, suggesting that ZIKV can block the growth signal at the source (receptor). Considering these strong blockage candidate scenarios, it is apparent that ZIKV is capable of taking down all the cascades triggered by IGF. Thus, autophagy during fetal development might be caused by complete blockage of IGF signaling to mTORC1 (to inhibit its activation).

If ZIKV induction of autophagy *in vivo* requires complete (not partial) blockage of the IGF signal along mTOR, it would be interesting to sketch the possible complete-blockage scenarios. There are two primary scenarios (mentioned above) that can take down the signal

cascade on their own. There are the blockage of the insulin receptor to prevent the signal from the source and the inhibition of mTOR itself to prevent mTORC1 complex formation or functioning at the sink. The other two primary scenarios are not sufficient to take down the two cascades at once (PI3K cascade and MAPK cascades) unless they operate simultaneously to block Akt or PI3K and TSC1/2. However, blockage of Akt alone might be sufficient to cause downstream autophagy and microcephaly, while interacting with PI3K is not [80].

The results also show that NS4A can interact with most key proteins in the mTOR signaling pathway. ZIKV NS4A and NS4B have been recently found to collaboratively induce autophagy in human fNSCs as well as other cell types [68]. The refereed study found that NS4A and NS4B suppress the activation of Akt by inhibiting upstream PI3K signaling and thus reducing Akt phosphorylation. This inhibition of Akt is thought to cause increased autophagy and disturbance of the neurogenesis in fNSCs.

5.6.5 Research Question 4

Why do not closely related viruses cause microcephaly as well? While the four Dengue virus serotypes share 65%-70% of sequence similarity, Zika virus is within the 54%-59% range of similarity [27]. DENV and ZIKV infections are similar serologically and symptomatically [14, 30, 61]. According to the twilight zone of sequence similarity concept [104], proteins sharing at least 40% of sequence identity are likely to be similar in structure and consequently function and interactions. ZIKV causes fetal microcephaly, while DENV does not [39]. Here I do a comparative study to find out why. This question has first been raised in a recent experimental study [68].

I examined the ability of DENV proteins to interact with the human proteins in the mTOR pathway, using DeNovo-Cascade as I did in testing ZIKV. The results show that my anticipation of DENV not causing microcephaly because it cannot induce autophagy is not absolutely correct. As shown in Figure 5.5, DENV has a high potential to induce autophagy by interfering with the PI3K key protein, with a possibility of interfering also with the mTOR complex itself. NS4A is also the major player for DENV to induce autophagy as NS4A of ZIKV. NS4A of DENV is believed to up-regulate autophagy for replication benefit in epithelial cells [46, 74].

I also found that ZIKV and DENV are likely to use different mechanisms of interfering with the mTOR pathway and induction of autophagy. While ZIKV can lock down the pathway by interfering with its central protein Akt, none of the DENV proteins can interact with Akt and instead DENV interacts with PI3K (upstream of Akt). A difference in autophagy induction mechanisms between ZIKV and DENV is also suggested in a recent experimental study [68].

I suggest three possible reasons for why DENV does not cause fetal microcephaly as ZIKV does. The first possibility is the difference in blocking strength of the IGF signal propagation.

ZIKV has strong potential to interact with multiple members of the two cascades triggered by IGF, which is sufficient (theoretically) to block the IGF signal and induce autophagy. On the other hand, DENV can down-regulate mTORC1 by blocking PI3K, but the prediction suggests that DENV is not strongly interacting with the other cascade (MAPK). Thus, according to this possibility, DENV might be able to induce autophagy in non-developmental cells where IGF is not triggered but cannot block the IGF signal in developmental cells like fNSCs.

The second possibility is the difference in the autophagy induction mechanisms between ZIKV and DENV. While DENV disturbs the PI3K-Akt-mTOR signaling upstream to Akt, ZIKV targets Akt itself. Non-functional Akt mutations have been reported to cause microcephaly [80]. Thus, Akt blockage might be sufficient to cause downstream autophagy and microcephaly, while interacting with PI3K is not.

The third possibility is that DENV-induced autophagy can fully function in fNSCs (if DENV targeting mTOR is sufficient to block mTORC1 functioning or formation *in vivo*), but DENV does not enter the fNSCs to cause this disorder in the first place. ZIKV is strongly believed to cross the placental barrier [16, 20, 57, 78], while there has been no report about the ability of DENV to cross the placental barrier. Thus, it might be the case that DENV does not have the right localization to cause microcephaly.

Interested biologists can test whether DENV can cause microcephaly if injected into fNSCs. If proven, then DENV can be confirmed not to cause fetal microcephaly because of localization constraint while it can cause microcephaly if it reaches the fNSCs. However, if DENV cannot cause microcephaly in the fetus, then examining the difference in sequence between ZIKV and DENV proteins should reveal key interaction interfaces for ZIKV that can be targeted for therapeutics. An interesting result from this comparative study is that DeNovo-Cascade is sensitive enough to identify such differences in interactions for closely related sequences.

5.7 Conclusions

This case study of ZIKV addresses questions under some theories about the virus. The prediction models are tools that can be used to investigate hypotheses quickly and help design experiments. The findings here, which also match published experimental results, are for demonstration of the capabilities of the DeNovo models, which can be used to examine other theories as well.

The DeNovo models are demonstrated as promising tools for predicting the interactions between virus and human proteins. DeNovo can catch the sensitive difference in PPIs for related proteins with high sequence similarity, thus, they can be used for comparative studies and for answering fundamental questions about viral behavior differences. Yet future research is needed to identify the effect of each interaction (whether it is activation or inhibition).

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Touched by the importance of disease networks, I decided to study the networks of protein-protein interactions (PPIs) between viruses and their hosts. Astonished at the scarcity of interactions, I found out that computational methods for predicting those interactions are only available for HIV and human proteins. Those methods need hundreds or thousands of known PPIs for each virus to be able to predict more interactions for the same virus. I thus realized the need to design a computational approach to predict PPIs between any virus and its host without requiring that there be a large number of known PPIs to learn from.

Realizing that the classical computational solutions for this problem are inadequate, I believed that a biology-inspired solution is a key. I thus returned to the molecular virology field to find inspiration. By employing several concepts of virology regarding similarity within and among viral families and how viral proteins interact with their host proteins, I was able to design three frameworks to predict the protein interactions between any virus and human once the virus protein sequences are available. The key solution has been to reverse the learning focus of the learning model from learning with respect to the viral proteins to learning with respect to the host proteins, making use of a shared host.

The DeNovo-Human framework is designed to predict the PPIs for a viral protein that has some known interactions with human proteins (Chapter 4), while the DeNovo framework is designed for the case when such interactions are not available (Chapter 3). However, DeNovo-Human achieves better accuracy and higher coverage than DeNovo. Restricting the prediction to the human host at this stage is for lack of sufficient data for other hosts. Overall, from the two frameworks, I observed that when a viral protein has no representative PPIs in the training set, the learning model makes predictions based on the PPIs of the other viral proteins with high similarity to it (in the same family). If the viral protein has representative PPIs, the model does not consider interactions of similar proteins in making

the prediction decision. I further proposed to cascade the two frameworks into the third one, DeNovo-Cascade. Using DeNovo, a seed of PPIs can be generated for a novel viral protein, then amplified using DeNovo-Human to make predictions against its large library of human proteins (Chapter 5). DeNovo-Cascade benefits from the two frameworks' coverage and accuracy; however, propagating the error through the two models reduces the overall performance.

This work on predicting PPIs between viruses and their host pushes the boundaries of viral bioinformatics. Previously, such prediction has been limited to HIV-human, but now prediction is possible for any novel virus if sufficient PPIs are known for its host with multiple viruses (which is the case with human). Besides providing the prediction *de novo* from viral protein sequence, the accuracy of the prediction is comparable to the best achieved in a single virus case (the HIV-human models) without requiring sophisticated prior knowledge (as in the HIV case). An additional advantage is a high coverage (larger human protein library to test against) compared to the single-virus case. As these are the first solutions to different aspects of the problem, I hope they open the door for future improvements that make the prediction possible without restrictions.

One goal of my research is to make my frameworks usable by convincing the biologists of their value. With the recent viral outbreaks, it was a great time to answer puzzling questions about one of the threatening viruses, namely Zika virus (Chapter 5). Using my frameworks and the biological context of the problem, I was able to specify which human cell receptors potentially internalize the virus into the cell. Further, I was able to identify the association between the Zika virus infection and interferon- β . Additionally, the prediction helped me to obtain insight into how the virus causes the brain disorder microcephaly in a fetus and why a closely related virus does not cause the same disorder. This case study demonstrates that the proposed frameworks can be extended beyond large-scale prediction to small scale and comparative studies to learn more about virus behavior. In collaboration with other labs, two comparative studies are in progress beside several single-virus studies. I wish to see more studies utilizing my frameworks to help expand the virus-host knowledge.

Surfing through these interleaved problems and solutions discussed in this work, the take-home message is that computational prediction models for biological problems should not act in isolation from the biological principles governing the system. The biology can guide the computational work to tailor it to the problem at hand. In my case, respecting the virus family similarities and differences enabled me to design a restricted negative sampling method that aims at selecting less likely interactions to use as negative examples for training the model. These similarities also were employed to improve the prediction accuracy when predicting for a novel virus and to design a practical procedure of how to test the frameworks to simulate the situation of predicting for new viruses. The strategies repeatedly employed by viruses to attack the host were also employed here to select an appropriate feature extraction scheme without the need to require sophisticated prior knowledge about each new viral protein.

Looking deeper into the learning models, I realized that, unlike what had been thought, predicting virus-human PPIs can be easier, not more difficult, than predicting for human-human PPIs. Considering the machine learning models I experimented with, a model trained to predict inter-species PPIs tries to find the relation between all-versus-all proteins in the training PPI examples. In the virus-human case, the viral proteins are dissimilar to the human proteins at the sequence and feature levels so that the learning model does not consider a relation between them. The models in the virus-human case are less complex, faster, and more accurate than for the human-human case.

It was also interesting to use the prediction framework to predict the whole virus-human PPI network. This gives insight into the high connectivity of a viral protein with the human proteins, which comes in agreement with the viral compact genomes and the wide range of functionality its proteins must perform. This also explains why using the restricted negative sampling method works to improve the prediction for virus PPIs with human while it has no significant effect when applied to the human-human prediction case.

6.2 Future Work

As these are the first solutions to the problem of predicting the PPI between any virus and its host, there is a large room for future improvements. Beyond improving their accuracies by experimenting with different computational learning models, employing more biologically insightful observations about virus-host interactions can be key to delivering better and more generalized prediction tools, probably for hosts with no known interactions as well. Considering more biologically meaningful sequence descriptors that make use of viral family similarities and locality of interaction interfaces is one possibility. Another possibility is to think of insightful ways to transfer the biology-based methods designed for single-species PPI prediction to be used in the virus-host prediction case.

It would be impactful to leverage the predicted interactions for inferring the short linear motifs associated with the interactions between viral and human proteins. One possibility is to mask sequence stretches iteratively and perform the prediction to locate the critical sequence for identifying the interactions.

Hosts other than human have few known interactions with their viruses reported in online databases, to the best of my knowledge. While understanding virus behavior in plants, for example, is crucial to crop production, we do not have enough known interactions to operate any of the proposed frameworks in this study. One opportunity is to use the interactions between proteins of a specific plant to train a model to recognize how those plant proteins interact. Then, we can use the model to predict potential interactions between the plant proteins and proteins of viruses infecting that plant.

What can be more beneficial and a breakthrough is to develop a generalized computational PPI prediction framework that can be used to determine if two proteins can physically

interact given their sequences. Working on such a project should be careful with the unlikely interactions to be used for training and testing the model. Probably using anomaly or novelty detection techniques (such as 1-class SVM model) to model the positive interactions can be a solution. It is also important for performance reporting to consider the three situations of prediction, namely predicting for a pair for which the model has seen examples of their interactions, only for one of them, or neither. Considering the sequence similarity to determine whether the model has seen examples for a specific protein or its similar ones is also an important issue for designing such a generalized system. When reporting the performance, it is also important to account for the positive to negative ratio and the connectivity in the PPI networks.

After studying the interactions of some viral proteins with human proteins, I believe there are patterns of interference to the human pathways that viruses repeatedly are engaged in while hijacking the cell processes. It would be beneficial to identify those patterns to learn more about the shared strategies that viruses develop (or share) to subvert the host cell. A starting point can be graph mining of the available virus-host PPI networks with alignment to the known pathways.

Upon identifying such interference patterns, it would be possible to use predicted PPIs between proteins of a specific virus and human proteins, for example, to propose possible interference scenarios. A framework that does this task should consider possibly involved pathways and how effective interference or perturbation to them can take place. Considering the amount of effort put into the work performed here in the Zika virus case study (Chapter 5), it will be time-saving for studying viruses to have such an automatic framework that maps the predicted interactions into possible scenarios. Until such an automated framework is developed, interested researchers may start applying the three frameworks to other threatening viruses to answer critical questions or examine hypothesis quickly at the time of viral outbreaks. However, contributions by biologists will be needed to form those questions and hypotheses.

Bibliography

- [1] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences*, 99(9):5896–5901, 2002.
- [2] P. Aloy and R. B. Russell. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, 19(1):161–162, 2003.
- [3] M. G. Ammari, C. R. Gresham, F. M. McCarthy, and B. Nanduri. HPIDB 2.0: A curated database for host–pathogen interactions. *Database*, 1:9, 2016.
- [4] A. S. Aytuna, A. Gursoy, and O. Keskin. Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21(12):2850–2855, 2005.
- [5] S. Bandyopadhyay, S. Ray, A. Mukhopadhyay, and U. Maulik. A review of *in silico* approaches for analysis and prediction of HIV-1-human protein–protein interactions. *Briefings in Bioinformatics*, 16(5):830–851, 2015.
- [6] R. K. Barman, S. Saha, and S. Das. Prediction of interactions between viral and host proteins using supervised machine learning methods. *PloS one*, 9(11):e112034, 2014.
- [7] A. Ben-Hur and W. S. Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38–i46, 2005.
- [8] A. Ben-Hur and W. S. Noble. Choosing negative examples for the prediction of protein–protein interactions. *BMC bioinformatics*, 7(Suppl 1):S2, 2006.
- [9] A. Ben-Hur and J. Weston. A users guide to support vector machines. In *Data Mining Techniques for the Life Sciences*, pages 223–239. Springer, 2010.
- [10] T. Berggård, S. Linse, and P. James. Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842, 2007.
- [11] P. Blohm, G. Frishman, P. Smialowski, F. Goebels, B. Wachinger, A. Ruepp, and D. Frishman. Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research*, 42(D1):D396–400, 2014.

- [12] J. R. Bock and D. A. Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.
- [13] D. D. Boehr and P. E. Wright. How do proteins interact? *Science*, 320(5882):1429–1430, 2008.
- [14] R. Buathong, L. Hermann, B. Thaisomboonsuk, W. Rutvisuttinunt, C. Klungthong, P. Chinnawirotpisan, W. Manasatienkij, A. Nisalak, S. Fernandez, I.-K. Yoon, et al. Detection of Zika virus infection in Thailand, 2012–2014. *The American Journal of Tropical Medicine and Hygiene*, 93(2):380–383, 2015.
- [15] A. Calderone, L. Licata, and G. Cesareni. VirusMentha: A new resource for virus–host protein interactions. *Nucleic Acids Research*, 43:D588–D592, 2014.
- [16] G. Calvet, R. S. Aguiar, A. S. Melo, S. A. Sampaio, I. de Filippis, A. Fabri, E. S. Araujo, P. C. de Sequeira, M. C. de Mendonça, L. de Oliveira, et al. Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: A case study. *The Lancet Infectious Diseases*, 16(6):653–660, 2016.
- [17] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
- [18] D. Cloëtta, V. Thomanetz, C. Baranek, R. M. Lustenberger, S. Lin, F. Oliveri, S. Atanasoski, and M. A. Rüegg. Inactivation of mTORC1 in the developing brain causes microcephaly and affects gliogenesis. *The Journal of Neuroscience*, 33(18):7799–7810, 2013.
- [19] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [20] F. R. Cugola, I. R. Fernandes, F. B. Russo, B. C. Freitas, J. L. Dias, K. P. Guimarães, C. Benazzato, N. Almeida, G. C. Pignatari, S. Romero, et al. The Brazilian Zika virus strain causes birth defects in experimental models. *Nature*, 534(7606):267–271, 2016.
- [21] G. Cui, C. Fang, and K. Han. Prediction of protein–protein interactions between viruses and human by an SVM model. *BMC Bioinformatics*, 13(Suppl 7):S5, 2012.
- [22] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–328, 1998.
- [23] J. Dang, S. K. Tiwari, G. Lichinchi, Y. Qin, V. S. Patil, A. M. Eroshkin, and T. M. Rana. Zika virus depletes neural progenitors in human cerebral organoids through activation of the innate immune receptor TLR3. *Cell Stem Cell*, 19(2):258–265, 2016.
- [24] N. E. Davey, G. Travé, and T. J. Gibson. How viruses hijack cell regulation. *Trends in Biochemical Sciences*, 36(3):159–169, 2011.

- [25] E. De Clercq. Strategies in the design of antiviral drugs. *Nature Reviews Drug Discovery*, 1(1):13–25, 2002.
- [26] J. De Las Rivas and C. Fontanillo. Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, 6(6):e1000807, 2010.
- [27] W. Dejnirattisai, P. Supasa, W. Wongwiwat, A. Rouvinski, G. Barba-Spaeth, T. Duangchinda, A. Sakuntabhai, V. Cao-Lormeau, P. Malasit, F. Rey, et al. Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with Zika virus. *Nature Immunology*, 17(9):1102, 2016.
- [28] H. Dinkel, K. Van Roey, S. Michael, N. E. Davey, R. J. Weatheritt, D. Born, T. Speck, D. Krüger, G. Grebnev, M. Kubań, et al. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Research*, 42(D1):D259–66, 2014.
- [29] J. M. Doolittle and S. M. Gomez. Structural similarity-based predictions of protein interactions between HIV-1 and homo sapiens. *Virology Journal*, 7(1):1, 2010.
- [30] M. R. Duffy, T.-H. Chen, W. T. Hancock, A. M. Powers, J. L. Kool, R. S. Lanciotti, M. Pretrick, M. Marfel, S. Holzbauer, C. Dubray, et al. Zika virus outbreak on Yap Island, federated states of Micronesia. *The New England Journal of Medicine*, 2009(360):2536–2543, 2009.
- [31] M. D. Dyer, T. Murali, and B. W. Sobral. Computational prediction of host-pathogen protein–protein interactions. *Bioinformatics*, 23(13):i159–i166, 2007.
- [32] M. D. Dyer, T. Murali, and B. W. Sobral. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathogens*, 4(2):e32, 2008.
- [33] M. D. Dyer, T. Murali, and B. W. Sobral. Supervised learning and prediction of physical interactions between human and HIV proteins. *Infection, Genetics and Evolution*, 11(5):917–923, 2011.
- [34] M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. Murali, and B. W. Sobral. The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. *PLoS One*, 5(8):e12089, 2010.
- [35] F.-E. Eid, M. ElHefnawi, and L. S. Heath. DeNovo: Virus-host sequence-based protein–protein interaction prediction. *Bioinformatics*, 32(8):1144–1150, 2016.
- [36] A. Emamjomeh, B. Goliaei, J. Zahiri, and R. Ebrahimpour. Predicting protein–protein interactions between human and hepatitis C virus via an ensemble learning method. *Molecular BioSystems*, 10(12):3147–3154, 2014.

- [37] P. Evans, W. Dampier, L. Ungar, and A. Tozeren. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Medical Genomics*, 2(1):27, 2009.
- [38] W. Fu, B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt, and R. G. Ptak. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Research*, 37(suppl 1):D417–D422, 2009.
- [39] P. P. Garcez, E. C. Loiola, R. M. da Costa, L. M. Higa, P. Trindade, R. Delvecchio, J. M. Nascimento, R. Brindeiro, A. Tanuri, and S. K. Rehen. Zika virus impairs growth in human neurospheres and brain organoids. *Science*, 352(6287):816–818, 2016.
- [40] S. K. Gire, A. Goba, K. G. Andersen, R. S. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.
- [41] M. Götz and W. B. Huttner. The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology*, 6(10):777–788, 2005.
- [42] A. Grant, S. S. Ponia, S. Tripathi, V. Balasubramaniam, L. Miorin, M. Sourisseau, M. C. Schwarz, M. P. Sánchez-Seco, M. J. Evans, S. M. Best, et al. Zika virus targets human STAT2 to inhibit type I interferon signaling. *Cell Host & Microbe*, 19(6):882–890, 2016.
- [43] A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 29(17):3513–3519, 2001.
- [44] Y. Guo, L. Yu, Z. Wen, and M. Li. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Research*, 36(9):3025–3030, 2008.
- [45] R. Hamel, O. Dejarnac, S. Wichit, P. Ekchariyawat, A. Neyret, N. Luplertlop, M. Perera-Lecoin, P. Surasombatpattana, L. Talignani, F. Thomas, et al. Biology of Zika virus infection in human skin cells. *Journal of Virology*, 89(17):8880–8896, 2015.
- [46] N. S. Heaton and G. Randall. Dengue virus-induced autophagy regulates lipid metabolism. *Cell Host & Microbe*, 8(5):422–432, 2010.
- [47] C. Hulo, E. De Castro, P. Masson, L. Bougueleret, A. Bairoch, I. Xenarios, and P. Le Mercier. ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Research*, 39(suppl 1):D576–D582, 2011.
- [48] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, and C. O’Donovan. The GOA database: Gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063, 2015.

- [49] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.
- [50] Z. Jin, M. Kotera, and S. Goto. Virus proteins similar to human proteins as possible disturbance on human pathways. *Systems and Synthetic Biology*, 8(4):283–295, 2014.
- [51] C. H. Jung, S.-H. Ro, J. Cao, N. M. Otto, and D.-H. Kim. mTOR regulation of autophagy. *FEBS Letters*, 584(7):1287–1295, 2010.
- [52] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2016.
- [53] O. Keskin, B. Ma, and R. Nussinov. Hot regions in protein–protein interactions: The organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, 345(5):1281–1294, 2005.
- [54] O. Keskin, N. Tuncbag, and A. Gursoy. Predicting protein–protein interactions from the molecular to the proteome level. *Chemical Reviews*, 116(8):4884–4909, 2016.
- [55] I. Kim, Y. Liu, and H. Zhao. Bayesian methods for predicting interacting protein pairs using domain information. *Biometrics*, 63(3):824–833, 2007.
- [56] A. M. King, M. J. Adams, E. J. Lefkowitz, and E. B. Carstens. *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*, volume 9. Elsevier, 2012.
- [57] Z. A. Klase, S. Khakhina, A. D. B. Schneider, M. V. Callahan, J. Glasspool-Malone, and R. Malone. Zika fetal neuropathogenesis: Etiology of a viral syndrome. *PLOS Neglected Tropical Diseases*, 10(8):e0004877, 2016.
- [58] M. Kshirsagar, J. Carbonell, and J. Klein-Seetharaman. Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks. In *NIPS Workshop on Machine Learning for Computational Biology*, 2013.
- [59] R. Kumar and B. Nanduri. HPIDB - A unified resource for host-pathogen interactions. *BMC Bioinformatics*, 11(6):1, 2010.
- [60] S. K. Kwofie, U. Schaefer, V. S. Sundararajan, V. B. Bajic, and A. Christoffels. HCVpro: Hepatitis C virus protein interaction database. *Infection, Genetics and Evolution*, 11(8):1971–1977, 2011.
- [61] R. S. Lanciotti, O. L. Kosoy, J. J. Laven, J. O. Velez, A. J. Lambert, A. J. Johnson, S. M. Stanfield, and M. R. Duffy. Genetic and serologic properties of Zika virus associated with an epidemics, Yap State, Micronesia, 2007. *Emerging Infectious Diseases*, 14(8):1232–1240, 2008.

- [62] M. Lapinsh, P. Prusis, T. Lundstedt, and J. E. Wikberg. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Molecular pharmacology*, 61(6):1465–1475, 2002.
- [63] M. Laplante and D. M. Sabatini. mTOR signaling at a glance. *Journal of Cell Science*, 122(20):3589–3594, 2009.
- [64] D. Y. Lee. Roles of mTOR signaling in brain development. *Experimental Neurobiology*, 24(3):177–185, 2015.
- [65] B. Levine, N. Mizushima, and H. W. Virgin. Autophagy in immunity and inflammation. *Nature*, 469(7330):323–335, 2011.
- [66] D. E. Levy, I. J. Marié, and J. E. Durbin. Induction and function of type I and III interferon in response to viral infection. *Current Opinion in Virology*, 1(6):476–486, 2011.
- [67] T. Li, R. Wernersson, R. B. Hansen, H. Horn, J. Mercer, G. Slodkiewicz, C. T. Workman, O. Rigina, K. Rapacki, H. H. Stærfeldt, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods*, 14(1):61–64, 2017.
- [68] Q. Liang, Z. Luo, J. Zeng, W. Chen, S.-S. Foo, S.-A. Lee, J. Ge, S. Wang, S. A. Goldman, B. V. Zlokovic, et al. Zika virus NS4A and NS4B proteins deregulate Akt-mTOR signaling in human fetal neural stem cells to inhibit neurogenesis and induce autophagy. *Cell Stem Cell*, 19(5):663–671, 2016.
- [69] L. Lu, H. Lu, and J. Skolnick. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins: Structure, Function, and Bioinformatics*, 49(3):350–364, 2002.
- [70] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.
- [71] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86, 1999.
- [72] J. A. Marsh, H. Hernández, Z. Hall, S. E. Ahnert, T. Perica, C. V. Robinson, and S. A. Teichmann. Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell*, 153(2):461–470, 2013.
- [73] S. Martin, D. Roe, and J.-L. Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics*, 21(2):218–226, 2005.

- [74] J. E. McLean, A. Wudzinska, E. Datan, D. Quaglino, and Z. Zakeri. Flavivirus NS4A-induced autophagy protects cells against death and enhances virus replication. *Journal of Biological Chemistry*, 286(25):22147–22159, 2011.
- [75] S. Mei. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PLoS One*, 8(11):e79606–e79606, 2013.
- [76] S. Mei and K. Zhang. Computational discovery of Epstein-Barr virus targeted human genes and signalling pathways. *Scientific Reports*, 6:30612, 2016.
- [77] S. Mei and H. Zhu. A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. *Scientific Reports*, 5:8034–8034, 2015.
- [78] J. J. Miner, B. Cao, J. Govero, A. M. Smith, E. Fernandez, O. H. Cabrera, C. Garber, M. Noll, R. S. Klein, K. K. Noguchi, et al. Zika virus infection during pregnancy in mice causes placental damage and fetal demise. *Cell*, 165(5):1081–1091, 2016.
- [79] G.-l. Ming and H. Song. Adult neurogenesis in the mammalian brain: Significant answers and significant questions. *Neuron*, 70(4):687–702, 2011.
- [80] G. M. Mirzaa, J.-B. RIVIÈRE, and W. B. Dobyns. Megalencephaly syndromes and activating mutations in the PI3K-AKT pathway: MPPH and MCAP. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 163, pages 122–130. Wiley Online Library, 2013.
- [81] V. Y. Muley and A. Ranjan. Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. *PLoS One*, 7(7):e42057, 2012.
- [82] P. R. Murray, K. S. Rosenthal, G. Kobayashi, and M. Pfaller. *Medical Microbiology*. Elsevier Health Sciences, 2015.
- [83] H. S. Najafabadi and R. Salavati. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biology*, 9(5):R87, 2008.
- [84] I. M. Nooren and J. M. Thornton. Diversity of protein-protein interactions. *The EMBO Journal*, 22(14):3486–3492, 2003.
- [85] E. Nourani, F. Khunjush, and S. Durmuş. Computational approaches for prediction of pathogen-host protein-protein interactions. *Frontiers in Microbiology*, 6:94–94, 2015.
- [86] I. Nouretdinov, A. Gammerman, Y. Qi, J. Klein-Seetharaman, and C. Learning. Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method. In *Pacific Symposium on Biocomputing*, volume 311, pages 311–322. World Scientific, 2012.

- [87] T. J. Nowakowski, A. A. Pollen, E. Di Lullo, C. Sandoval-Espinosa, M. Bershteyn, and A. R. Kriegstein. Expression analysis highlights AXL as a candidate Zika virus entry receptor in neural stem cells. *Cell Stem Cell*, 18(5):591–596, 2016.
- [88] Y. Ofra and B. Rost. Protein–protein interaction hotspots carved into sequences. *PLoS Computational Biology*, 3(7):e119, 2007.
- [89] U. Ogmen, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy. PRISM: Protein interactions by structural matching. *Nucleic Acids Research*, 33(suppl 2):W331–W336, 2005.
- [90] Y. Park and E. M. Marcotte. Revisiting the negative example sampling problem for predicting protein–protein interactions. *Bioinformatics*, 27(21):3024–3028, 2011.
- [91] Y. Park and E. M. Marcotte. Flaws in evaluation schemes for pair-input computational predictions. *Nature Methods*, 9(12):1134–1136, 2012.
- [92] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering*, 14(9):609–614, 2001.
- [93] P. Pei and A. Zhang. A topological measurement for weighted protein interaction network. In *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*, pages 268–278. IEEE, 2005.
- [94] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.
- [95] N. Perdigão, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, et al. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, 2015.
- [96] M. Perera-Lecoin, L. Meertens, X. Carnec, and A. Amara. Flavivirus entry receptors: An update. *Viruses*, 6(1):69–88, 2013.
- [97] B. E. Pickett, E. L. Sadat, Y. Zhang, J. M. Noronha, R. B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu, et al. ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, 40(D1):D593–D598, 2012.
- [98] L. C. Platanias. Mechanisms of type-I-and type-II-interferon-mediated signalling. *Nature Reviews Immunology*, 5(5):375–386, 2005.
- [99] Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics*, 26(18):i645–i652, 2010.

- [100] R. Randhawa and P. Cohen. The role of the insulin-like growth factor system in prenatal growth. *Molecular Genetics and Metabolism*, 86(1):84–90, 2005.
- [101] V. S. Rao, K. Srinivas, G. Sujini, and G. Kumar. Protein-protein interaction detection: Methods and analysis. *International Journal of Proteomics*, 2014:147648–147648, 2013.
- [102] J. Reimand, S. Hui, S. Jain, B. Law, and G. D. Bader. Domain-mediated protein interaction prediction: From genome to network. *FEBS letters*, 586(17):2751–2763, 2012.
- [103] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 2014.
- [104] B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, 1999.
- [105] R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein–protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19(6):756–763, 2003.
- [106] M. Sarmady. *HIV protein sequence signatures for crosstalk with host proteins*. PhD thesis, Drexel University, 2010.
- [107] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, 2007.
- [108] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, 2007.
- [109] B. A. Shoemaker and A. R. Panchenko. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, 3(4):e43, 2007.
- [110] M. Singhal and H. Resat. A domain-based approach to predict protein-protein interactions. *BMC Bioinformatics*, 8(1):199, 2007.
- [111] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692, 2001.
- [112] S.-H. Tan, Z. Zhang, and S.-K. Ng. ADVICE: automated detection and validation of interaction by co-evolution. *Nucleic Acids Research*, 32(suppl 2):W69–W72, 2004.

- [113] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman. Prediction of interactions between HIV-1 and human proteins by information integration. In *Pacific Symposium on Biocomputing*, pages 516–527. NIH Public Access, 2009.
- [114] N. Tokuriki, C. J. Oldfield, V. N. Uversky, I. N. Berezovsky, and D. S. Tawfik. Do viral proteins possess unique biophysical features? *Trends in Biochemical Sciences*, 34(2):53–59, 2009.
- [115] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.
- [116] UniProt Consortium and Others. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42(D1):D191–D198, 2014.
- [117] J. M. Urquiza, I. Rojas, H. Pomares, J. Herrera, J. Florido, and O. Valenzuela. Selecting negative samples for PPI prediction using hierarchical clustering methodology. *Journal of Applied Mathematics*, 2012, 2012.
- [118] G. A. Versteeg and A. García-Sastre. Viral tricks to grid-lock the type I interferon system. *Current Opinion in Microbiology*, 13(4):508–516, 2010.
- [119] J.-P. Vert, J. Qiu, and W. S. Noble. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics*, 8(10):S8, 2007.
- [120] A. Via, B. Uyar, C. Brun, and A. Zanzoni. How pathogens use linear motifs to perturb host cell networks. *Trends in Biochemical Sciences*, 40(1):36–48, 2015.
- [121] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.
- [122] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35(suppl 1):D5–D12, 2007.
- [123] C. G. Woods. Human microcephaly. *Current Opinion in Neurobiology*, 14(1):112–117, 2004.
- [124] X. Wu, A. Fleming, T. Ricketts, M. Pavel, H. Virgin, F. M. Menzies, and D. C. Rubinsztajn. Autophagy regulates Notch degradation and modulates stem cell development and neurogenesis. *Nature Communications*, 7:10533, 2016.
- [125] J. Zahiri, J. Hannon Bozorgmehr, and A. Masoudi-Nejad. Computational prediction of protein–protein interaction networks: Algorithms and resources. *Current Genomics*, 14(6):397–414, 2013.

- [126] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–560, 2012.
- [127] Z. Zhao, J. Xia, O. Tastan, I. Singh, M. Kshirsagar, J. Carbonell, and J. Klein-Seetharaman. Virus interactions with human signal transduction pathways. *International Journal of Computational Biology and Drug Design*, 4(1):83–105, 2011.
- [128] L. Zheng, C. Li, J. Ping, Y. Zhou, Y. Li, and P. Hao. The domain landscape of virus-host interactomes. *BioMed research international*, 2014:867235–867235, 2013.
- [129] H. Zhou, S. Gao, N. N. Nguyen, M. Fan, J. Jin, B. Liu, L. Zhao, G. Xiong, M. Tan, S. Li, et al. Stringent homology-based prediction of *H. sapiens-M. tuberculosis* H37Rv protein-protein interactions. *Biology Direct*, 9(5), 2014.
- [130] H. Zhou, J. Rezaei, W. Hugo, S. Gao, J. Jin, M. Fan, C.-H. Yong, M. Wozniak, and L. Wong. Stringent DDI-based prediction of *H. sapiens-M. tuberculosis* H37Rv protein-protein interactions. *BMC Systems Biology*, 7(Suppl 6):S6, 2013.