MULTIMODAL DATA FUSION AND MACHINE LEARNING FOR DECIPHERING PROTEIN-PROTEIN INTERACTIONS

An Undergraduate Research Scholars Thesis

by

ARGHAMITRA TALUKDER

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Dr. Yang Shen

May 2021

Major: Electrical Engineering

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

I, Arghamitra Talukder, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Research Faculty Advisor prior to the collection of any data used in this final thesis submission.

This project did not require approval from the Texas A&M University Research Compliance & Biosafety office.

TABLE OF CONTENTS

ABSTRACT

Multimodal Data Fusion and Machine Learning for Deciphering Protein-Protein Interactions

Arghamitra Talukder
Department of Electrical and Computer Engineering
Texas A&M University

Research Faculty Advisor: Dr. Yang Shen
Department of Electrical and Computer Engineering
Texas A&M University

Protein-protein interactions (PPIs) often underlie important biological processes. Due to the vast quantity of potential PPIs in living organisms, it can be an expensive if not daunting task to identify each PPI experimentally, thus computational methods have been developed in parallel to facilitate the task. Despite various experimental or computational methods to determine or predict PPIs, a knowledge gap is often there to understand the 3-dimensional interactions in atomic-level details. This research project aims to leverage the existing protein data and emerging tools of machine learning to both predict and explain protein-protein interactions. Specifically, using various modalities of protein data including 1D sequences and 2D structures, several hierarchical recurrent neural network (HRNN) and joint attention based models have been developed. These models predict whether two proteins interact (the probability of PPI) and, if they do, how they interact (the probabilities of their residue-residue contacts (RRC)). The prediction of PPI from model I (uses only 1D sequences) has Area under the Precision-Recall Curve (AUPRC) output of 0.738. In the comparative analysis of model I with state-of-the-art PPI-detect [1], the precision, sensitivity and accuracy increased 7.8%, 9.5% and 6.2% respectively setting the geometric mean as binary threshold. To predict inter-protein RRC map, a gradual improvement has been observed

from model I , model II (uses sequence pre-training and inter RRC maps to fine tune) and model III (uses both sequences and intra-protein RRC maps) in case of test set. As a result, the best AUPRC for test set reached 2.78e-3 (model III), from 2.51e-3 (model II) and 1.10e-3 (model I). Thus, model III showed 153% AUPRC improvement than model I and 11% than model II; additionally model II showed 128% improvement than model I. The performance evaluations of these models show that the advantage of big data for 1D modality alone is not good enough to predict inter-protein RRC maps; rather joint attentions supervised by training PPI structure data and pretraining sequence embedding by model I as done in model II give much better inter-protein RRC predictions. The further combination of sequences and intra-protein RRC maps in model III, two modalities of individual protein data, shows the best results.

ACKNOWLEDGMENTS

Contributors

I would like to thank my faculty advisor, Dr. Yang Shen, for his guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to two of my lab members Rujie Yin and Yuning You for their encouragement, help and support.

Some of the data analyzed for the thesis "MULTI-MODAL DATA FUSION AND MACHINE LEARNING FOR DECIPHERING PROTEIN-PROTEIN INTERACTIONS" were provided by Rujin Yin.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

This work was part of the "Faculty Early Career Development Program (CAREER)" fund from NSF obtained by my faculty advisor. As this project was also part of my necessary senior design coursework, I personally did not receive any funding

NOMENCLATURE

PPI	Protein-protein interaction
RRC	Residue-residue contact
MSA	Multi-sequence alignment
DCA	Direct coupling analysis
plmDCA	Pseudolikelihood maximization DCA
UniProt	Universal Protein Resource
PDB	Protein Data Bank
SOTA	State of the art
HRNN	Hierarchical recurrent neural network
GRU	Gated recurrent units
HPRC	High Performance Research Computing
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under Curve - Receiver Operating Characteristics

1. INTRODUCTION

1.1 Proteins and Protein-Protein Interactions

Living systems including the human body are organized in hierarchical levels such as molecules, cells, tissues, and organs. One of the most important cellular molecules is the protein. Proteins contribute to most biological processes including genetic expression, intercellular communication, morphology, nutrition absorption and so on.

Proteins are linear chains of amino acids bonded sequentially. The 1-dimensional (1D) amino-acid sequences determine proteins' identities; and they often "fold" into 3-dimensional structures to express their specific functions. As the mechanisms of the human body are revealed, proteins are often found to interact with each other (among other molecules). Analyzing protein behaviors from the perspective of protein-protein interactions would help understand the biological processes they underlie.

1.2 Current Methods for Studying Protein-Protein Interactions

For the binary outcome of PPIs (whether proteins interact), there are various experimental methods such as affinity purification [2], yeast two hybrid [3], co-immunoprecipitation [4]. These experimental methods are high-throughput and accumulating large amount of binary data on PPIs. However, considering the vast space of potential PPIs in living systems, computational methods of even higher throughput have been developed. Some computational methods adopted homology-based approaches like interolog search. Interolog search is based on the principle that interactions are conserved and interlogs are homologous pairs of protein interactions across different species. The homology-based method also includes phylogenetic similarities which relates to the common ancestor proteins among species [5]. The simulation-based methods include protein docking. Protein docking is molecular modeling which predicts the mutual orientation [6]. A lot of machine learning techniques have been also applied based on protein sequence, structure and function. The limitations with these approaches are the difficulties to model any conformational changes in docking and lack of thorough understanding of the binding mechanism in learning. [5]

For the 3D structures of PPIs (how proteins interact), experimental methods are of lower throughput compared to those to identify whether proteins interact. Such structure determination experiments include X-ray crystallography [7], nuclear magnetic resonance (NMR) [8], and cryo-EM [9], [10] . Besides the relatively lower throughput, they each have their limitations and are not directly amenable to all PPIs. Computational methods are thus developed in parallel, including principle-driven protein docking and data-driven machine learning methods. Protein docking methods find the best fit between two protein structures, with atom-level details, by following the principle of energy minimization, which faces the challenge to model protein conformational changes, to derive a powerful energy function, and to search the conformational space efficiently. Recently, data-driven methods have been developed to find coarser-grained, amino acid (or residue) level 3D contact patterns between proteins. Examples include unsupervised methods such as direct coupling analysis based on residue-residue co-evolution [11] as well as supervised machine learning methods [12]. These data-driven methods face the challenges from limited data especially evolution data and 3D structure data for protein pairs.

1.3 Project Overview

An accurate PPI prediction model will serve several objectives including pathways for unknown proteins, different binding modes, specificity of protein based multiple targets, effectiveness of drugs, and design of new protein etc. This project will take a data-driven approach to simultaneously predict whether and how proteins interact. And the method development directly addresses aforementioned challenges to current methods, especially the data limitation, by data fusion in machine learning.

Specifically, this project aims to address several aspects of protein-protein interactions. Given two proteins (protein A and protein B), the project will try to answer three questions with data-driven predictive models:

1. Are the proteins interacting? With a binary output (1 being yes and 0 being no) it would aim to predict if two proteins are interacting or not.

2. If they are interacting what are the specific positions of interactions? As mentioned before proteins are made of amino acids or residues. The second focus of the project would be to predict if the ith amino acid of protein A is interacting with the jth amino acid of protein B.

3. Considering distance as continuous random variable, what is the distance distribution between residues of protein A and protein B.

4. Lastly the performance evaluation with respect to the SOTA methods.

To answer the above questions, several modalities would be used. Examples include the 1-D sequences of proteins or the strings of amino acids, the 2-D structure modality of two individual proteins, and cross modality embedding.

The rest of the thesis is organized as follows. Chapter 2 would describe data extraction and curation methodologies, different model functionalities and architectures and a baseline model; Chapter 3 includes comparative analysis and thorough description of the model performance followed by last chapter which includes conclusion and future directions.

2. METHODS

This chapter focuses on our methodology and is divided into three main sections: data curation and statistics, the model description and the baseline model. The data curation section would describe the source databases selected to collect protein interaction data from. The model description section would contain the machine learning tools used and their explanations. Finally, the baseline model section would include the reference models and the state of the art (SOTA) models used to compare the performance of the newly built model.

We aim at predictions for two different outputs: protein-protein interaction (PPI) and inter-protein RRC (residue-residue contact) map. PPI prediction is about if two proteins are interacting with each other or not; the output is either 0, when two proteins are not interacting or 1, when they are interacting. On the other hand, Inter-protein RRC map represents the 3D interaction at the residue level. If protein A has x residues (or amino acids) and protein B has y residues, inter protein RRC map shows which pairs of residues i–j are interacting or not $i = 1, \ldots, x$ and $j = 1, \ldots, y$. Currently both PPI and RRC outputs are treated binary and predicted with a probability for each protein pair or residue pair. In future, they can be extended to continuous outputs such as PPI affinities and RRC distances, with predicted probabilities over discretized ranges.

2.1 Data Curation and Statistics

Various data sources are used in the study for the lists of positive and negative PPIs, the sequences and the (predicted) structures of individual proteins involved, and the complex structures of selected positive PPIs.

2.1.1 PPIs and 1D Modality of Protein Sequences

Four source datasets which store and provide various PPI data (positive and negative PPIs as well as positive PPIs with 3D complex structures) were used to further extract protein sequences and structures. The description of the source databases is given below in **Table 2.1**.

The extraction of 1D modality of protein sequence data was done for both a benchmark

Table 2.1: Description of the source PPI databases

Name	Description
iRefWeb [13]	Contains proteins identities known to interact with each other
Negatome [14]	Contains proteins identities known to not interact with each other
INstruct [15]	Contains proteins identities known to interact with each other and their interacting structures are also available
3did [16]	Contains domain identities known to interact with each other and their interacting structures are also available

dataset of PPI prediction [1] and our own curated dataset of both PPI and RRC predictions. The protein sequences were collected using an automated python script

To evaluate and compare the performance of our model, a specific data set has been used from PPI-Detect [1]. The website associated with the paper provides 4,327 pairs of proteins (involved in 1,922 interacting and 2,405 non-interacting pairs) and their sequences in the FASTA format. The dataset was split into training and test sets following [1]. Specifically, the test set contained three subsets. (1) The easy subset has 426 pairs (150 positive and 276 negatives) of proteins A-B where both A and B are present in the training set but interacting with other proteins. (2) The mid-hard subset is made of 307 pairs (102 positive and 205 negative ones) of proteins A-B where either A or B (but not both) is present in the training set. (3) The very hard subset is made of 103 pairs (57 positive and 46 negative ones) of proteins A-B, where neither protein A nor protein B is found in the training set.

For our own dataset, 310,180 unique positive PPIs were collected from iRefWeb [13] and 5685 unique negative PPIs (protein pairs validated to not interact) were collected from Negatome [14]. Both positive and negative interaction lists had some unique proteins only present in either list. Including such proteins in a training set can make the resulting machine-learning model biased. Therefore a shorter protein list was made where each retained protein was present in both positive and negative interaction lists. After such a procedure to remove the exclusive bias, we had 4941 positive interactions and 4326 negative interactions made of 1031 unique proteins; 289 positive interactions and 120 negative interactions are homodimers. To remove the redundancy CD-Hit [17]

was used with a 40% cutoff and we had 991 unique proteins. Both positive and negative interaction data set were randomly divided as follows: The validation set 10%, test set 10% and the remaining 80% was included in the training set.

As the protein identities for these PPIs were available in UniProt ID [18] which also gives a formatted link of a sequence file in the FASTA format. A link has been made based on the ID to access the FASTA file. For example, if the UniProt ID of a protein is Q6ZNK6, the FASTA file would have a common link pattern of 'https://www.uniprot.org/uniprot/ Q6ZNK6.fasta'. The FASTA file contains the sequence which has been automatically read from the web and stored in a .csv file along with the label if they are interacting or not. Each sequence is made of amino acids or residues presented with different single letters; for example, Alanine is represented as 'A', Aspartic acid is represented as 'D', and so on. Using a custom python dictionary the letters are converted in numerical orders for the one-hot encoding input to neural networks. The data statistics of the PPI dataset and corresponding protein sequences are shown in **Table 2.2**. And the distributions of protein sequences for iRefWeb [13] (positive PPI) and Negatome (negative PPI) [14]are shown in **Figure 2.3**.

Table 2.2: PPI and protein sequence data statistics

Name	**Statistics**
iRefWeb [13]	Extracted unique positive interactions: 310180 Unique proteins: 67607 The number of homodimers: 6814 Average length of protein sequences: 548 Maximum length of protein sequences: 32759 Minimum length of protein sequences: 12
Negatome [14]	Extracted unique negative interactions: 5685 Unique proteins: 3214 The number of homodimers: 70 Average length of protein sequences: 339 Maximum length of protein sequences: 7074 Minimum length of protein sequences: 16

.

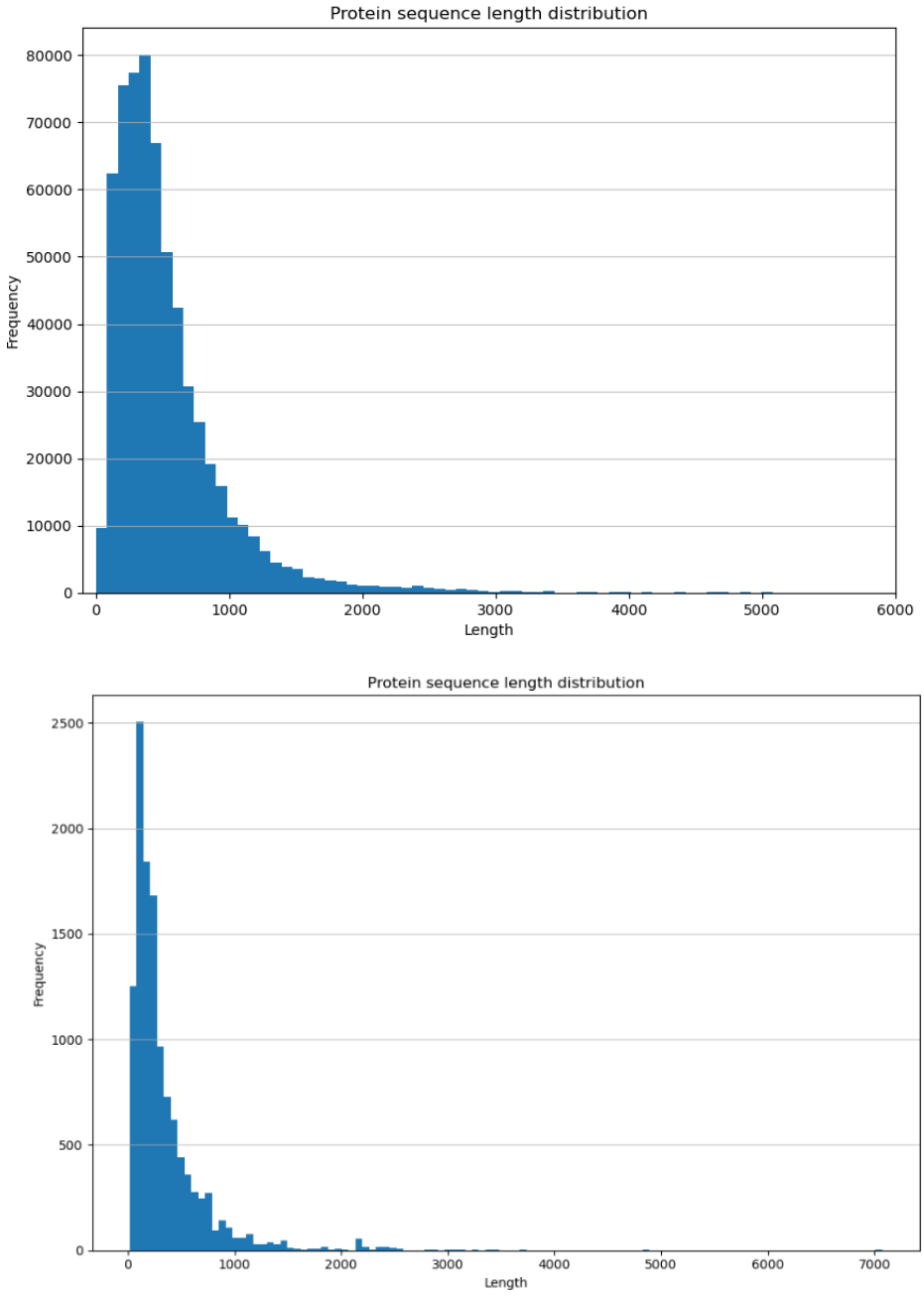

Figure 2.1: Protein sequence length distribution in iRefWeb [13] (top) and Negatome [14] (bottom).

2.1.2 2D Modality of Intra- and Inter-Protein 3D Structures

The 2D modality or protein structure data is gathered in two parts: inter-protein residue-residue contact (RRC) maps derived from 3D structures of bound protein-protein complexes and intra-protein RRC maps derived from unbound structures of individual proteins (or predicted from their sequences).

For inter-protein RRC data, it is the overlapping part between the previously mentioned positive PPI dataset (iRefWeb [13]) and a database of positive PPIs with known complex structures (INstruct [15]). Entries in INstruct [15] are based on two major sources of domain-domain interaction evidence: direct co-crystal structures and indirect inference from the co-crystal structures of homologous domains. Since the inter-protein contacts represent a major objective of this paper, their high precision is desired. Therefore we only used the entries evidenced by the direct co-crystal structures and found 2,422 positive PPIs with co-crystal structures. After removing the redundancy of the entries representing the same domain-domain interaction and the entries of the same protein pairs but representing different domain-domain interactions, 1,001 pairs were obtained, with protein data bank (PDB) IDs of their co-crystal structures provided.

To calculate inter-protein contact maps for the 1,001 positive PPIs in INstruct [13], their 3D structures were retrieved and analyzed in Python scripts using Biopython (a Python tool for molecular biology) and atomium [19]. For a given pair between protein A and B, either a homo-mer of identical proteins or a hetero-mer of different proteins, chains in the structure were aligned to corresponding protein UniProt sequences and residues in the structure are re-indexed with the corresponding residue index in the sequence. For each pair of residues i-j, we determined whether any distance between a heavy atom of residue i and a heavy atom of residue j was within 5Å and only assigned a nonzero value of 1 to the (i, j) element of the inter-protein contact map when that is the case. We note that the zero elements of the inter-protein contact maps could indicate either a non-contact between the two residues or the missing 3D structural information of at least one residue. We also removed 47 homo-mers because no information about inter-protein residue-residue contacts can be obtained from the PDB structure (often originating from the lack of stoichiometry

information in the biological assembly files). In total, 954 inter-protein RRC maps were curated for INstruct [15], and 68 of them overlap with the curated 1D modality sequences.

For intra-protein contact maps for the 79 unique proteins involved in the 68 overlapping pairs, the predicted intra-protein residue-residue contact results from RaptorX [20] are used which is a reasonably reliable source of intra-protein contact information.

Removing proteins which had more than 1,000 residues the resulting dataset of 57 positive PPIs with co-crystal structure information, a subset of our curated positive PPIs with just binary information, was split into 30 training, 8 validation, and 19 test pairs. **Figure 2.2** shows the data division among PPIs (positive and negative PPIs as well as PPIs with structures).

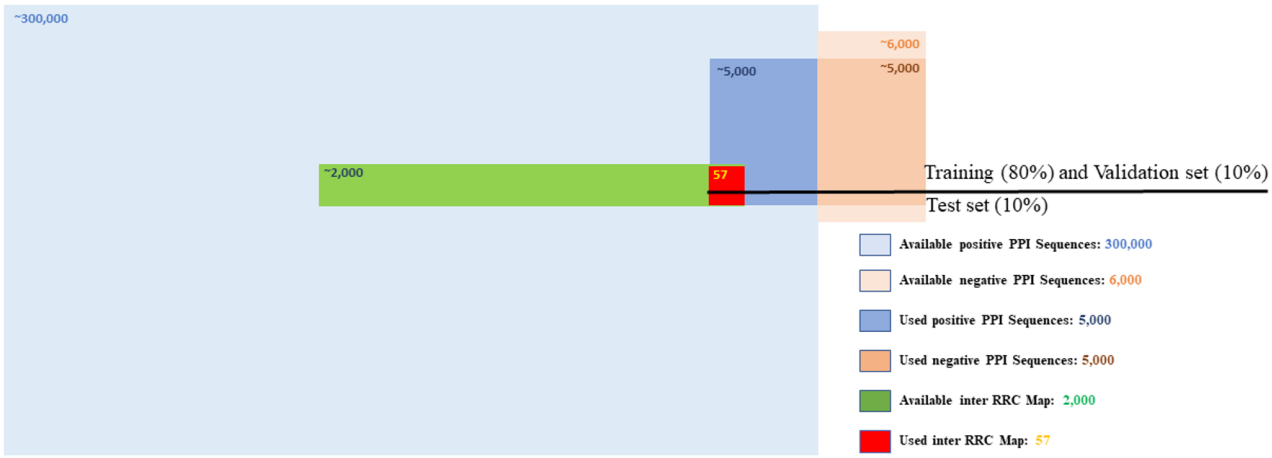

Figure 2.2: Data Division:1D and 2D modality

2.2 Model Description

2.2.1 Prediction of PPI Using 1D Sequence: Model I

Model I is designed to take protein sequences or 1D modality as input and predict PPIs as binary outcomes. The input of the model includes two protein sequences and the output of the model is a probability vector of PPI. To train the model, interacting labels for training protein pairs are also available (if they are interacting label = 0, otherwise label = 1). To implement model I

the idea of hierarchical recurrent neural networks (HRNN) [21] has been used to encode protein sequences. The mechanism of HRNN is well used for modeling sequential data - its application to natural language processing exploits the relations among word embedding in a sentence embedding. The concept to use HRNN to detect PPI was very similar: residues in a protein sequence interact locally to fold into k-mers and secondary structures and globally to fold into tertiary structures before they interact across proteins to form quaternary structures. A sequence of proteins has been padded to reach length 1000 (the maximum length of proteins in the dataset) and divided into k-mers.

Mathematically a vector of length 1000 has been converted into a matrix of dimension 25×40 assuming when proteins fold in a 3-dimensional space, each row interacts with each other. The model architecture is made of amino acid or residue embedding. First, each residue has been embedded into a vector of length 128. Two interacting proteins have converted into matrices of dimension 25×40×128 and have passed two hierarchical stages of GRUs (a form of RNN) first horizontally and then vertically. **Figure 2.3** shows the entire conversion of interacting protein sequences. Two proteins with length L1 and L2 are interacting with proteins of length L3 and L4; In the second stage they are padded with 0 so that all of the proteins have a new length Lmax = 1000; in the last stage of conversion vectors of length Lmax are transformed into a matrix of 25×40 (25 k-mers of length $k = 40$). The activation function ReLU is used on both of the protein embeddings and passed through the linear transformation of joint attention, dimension of 128×128. Both of the attention are concatenated by the Pytorch Einsum function and the activation function Sigmoid is finally used to get the inter-protein RRC map. In the following mathematical expressions, x is the input, y is the output, b is bias and w is the weight vector.

$$
\begin{aligned}
ReLU(x) &= (x)^{+} = max(0, x); \\
Sigmoid(x) &= \sigma(x) = \frac{1}{1 + exp(-x)}; \\
y &= xA^{T} + b
\end{aligned}
\qquad \text{(Eq. 1)}
$$

To get the PPI predictions, both protein embeddings are concatenated and passed through the Tanh activation function. The joint protein embedding is concatenated with an inter RRC map and passed through two different sequential models: one was made of one 1 dimensional convolution, LeakyReLU activation function, and max pooling; and the other used two similar layouts one after another (these layouts include one layer of linear transformation followed by LeakyReLU with a dropout rate of 0.7). Lastly, one last layer of linear transformation gives out the prediction of PPI.

$$LeakyReLU(x) = \max(0, x) + negative\_slope * \min(0, x) \quad \text{(Eq. 2)}$$

BCE loss or binary cross entropy loss has been used between PPI prediction and PPI labels to calculate PPI cross entropy.

$$\begin{aligned} BCEloss = ll(x, y) = L = l_1, ...{l_N}^T, \\ l_n = -w_n[y_n * log(x_n) + (1 - y_n) * log(1 - x_n)] \end{aligned} \quad \text{(Eq. 3)}$$

To optimize the model performance, the optimization step size (learning rate), the dropout rate and other hyperparameters have been tuned over the validation set and finally got the best value for a step size of 1e-3, dropout 0.7, 200 epochs, and batch size of 8. The ADAM algorithm has been for backpropagation training. The best AUPRC value of the validation dataset was used to select the epoch and save optimal parameters of the neural network model. The overall picture of protein embedding through GRU and the output format is shown in **Figure 2.4**.

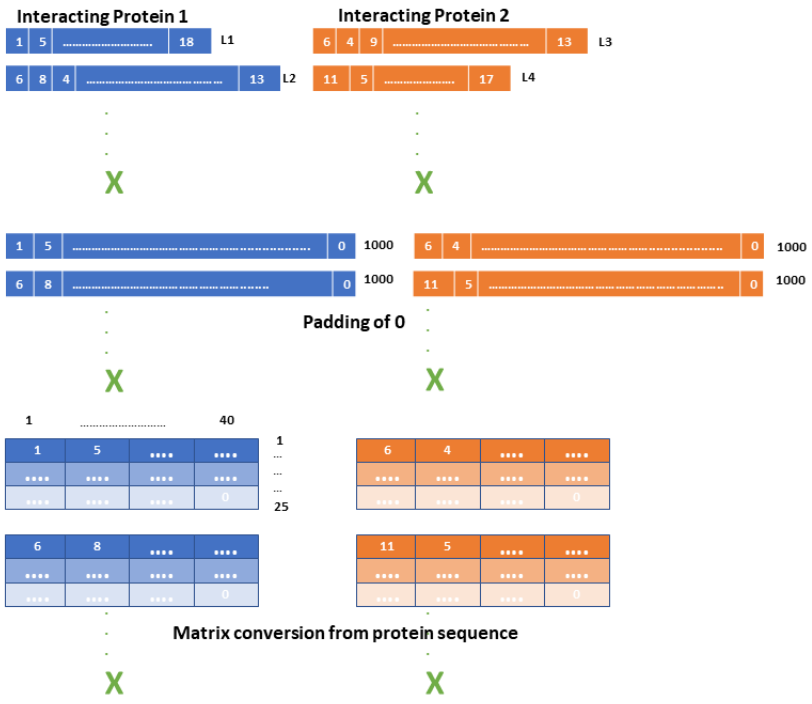

Figure 2.3: Protein conversion from 1D sequence to k-mer matrix

As of software CUDAtoolkit of version 10.2 and PyTorch version 1.6 have been used so that the training of the model can be done effectively through GPU computing. All the experiments are done using the Terra cluster at Texas A&M High Performance Research Computing facility (TAMU HPRC).

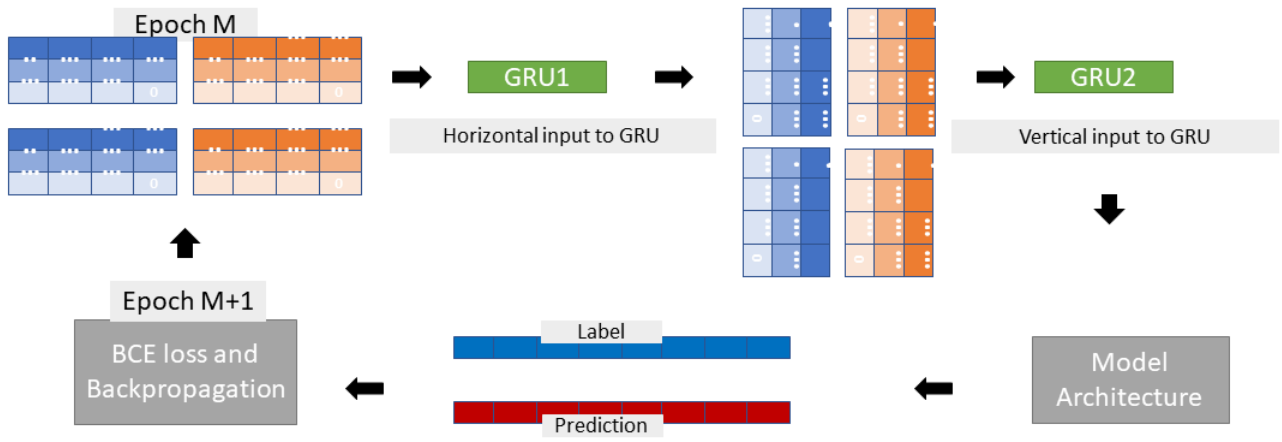

Figure 2.4: Model I: Protein embedding and output format

2.2.2 *Prediction of Inter-Protein RRC Using 1D Sequence: Model II*

Model II takes interacting protein sequences and predicts the inter RRC map. The inputs of model II are two sequences of interacting proteins and their inter RRC maps as the label to calculate the loss. The architecture of model II is very similar to model I; additionally, the concept of pretraining and fine-tuning is used. As the output of model II is different from model I, new loss calculations also have been introduced.

Model I has a huge advantage of using big data of binary information for PPI. As mentioned earlier, the available dataset of PPI is much bigger than inter RRC map or inter protein structure. To use this advantage of big data, the parameters of protein embedding in model II are initialized with the optimum parameters obtained from model I (**Figure 2.5**). This pre-training of embedding parameters in model II connects the 1D modality with the 2D modality. This embedding part consists of amino-acid or residue embedding and two layers of GRU.

As shown in **Figure 2.5**, the outputs of model II are inter RRC maps or matrices of dimension 1000×1000. After pre-training model II, the available inter RRC map or labels have been used to fine-tune the model. That is why the loss calculation includes the weighted sum of two different losses: the L_1 sparsity of neural network parameters and a BCE loss. But this time the

BCE loss is calculated between predicted and true inter RRC maps; this can be also considered as inter-protein RRC cross-entropy. As the maximum protein length is 1000, each protein is padded to make length 1000 while the actual length of protein A is x and the actual length of protein B is y ($x, y < 1000$).

Though the output, inter RRC map, is a matrix of 1000×1000 where the rows are protein A residues and column are protein B residues. To calculate the loss, the padded portion of the inter RRC map needs to be discarded. Thus the matrix is multiplied with a row vector and column vector of length 1000 made of 1 and 0. The row vector has the first x number of 1 and the next $1000 - x$ number of 0; similarly, the column vector has y number of 1 and $1000 - y$ number of 0. The sparsity regularization has been adjusted with a weight factor. Four different models were trained with sparsity regularization factor of 1e-3, 1e-4 and learning rate of 1e-3, 1e-4. The best performance of among these four different combinations were considered and mentioned in the result section.

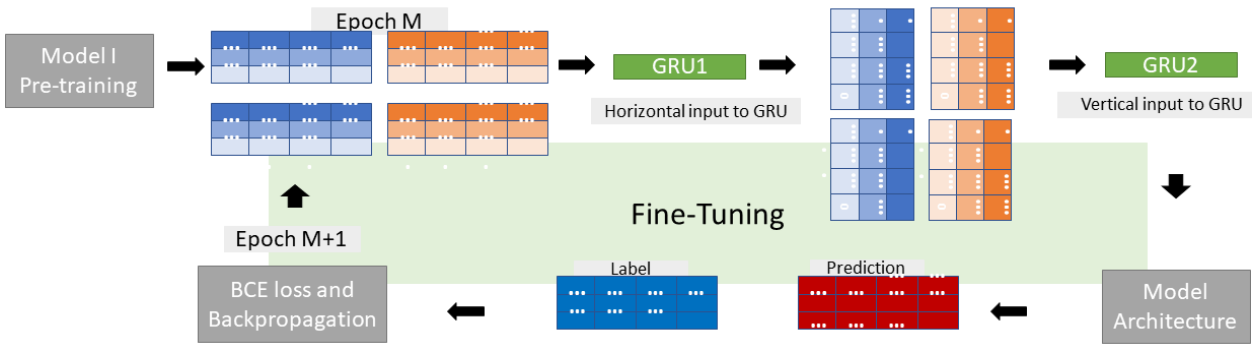

Figure 2.5: Model II: Pre-training from model I and fine-tuning through inter RRC map labels

2.2.3 Prediction of Inter-Protein RRC Using Both 1D Sequence and 2D Intra-Protein RRC: Model III

Model III can be considered as an expansion of model II. This model takes two different modalities of input: protein sequences as 1D modality and intra RRC map or intra-protein struc-

ture as 2D modality. And it predicts inter-protein RRC map. Explicitly, as input, there are two sequences and two intra protein structures of the interacting proteins. The intra-protein RRC map represents which of the residue positions within a protein interact with each other. Like model II, model III also takes inter-protein RRC map from the training data to evaluate the loss **Figure 2.6**.

The embedding of protein sequences in model III follows the same architecture as that in model I. Additionally, the intra-protein RRC maps are processed as protein graphs where each node is a residue and each edge represents a spatial contact. The embedding of intra-protein RRC maps follows another neural-network architecture named Graph Attention Networks (GAT) which is made of layers of linear transformation and ReLU activation function. The protein sequence embeddings and protein graph embeddings are concatenated. The prediction of inter-protein RRC is very similar to the calculation mentioned for model I and model II.

Similar to model II the loss calculation is done through PyTorch BCE to calculate intra RRC cross-entropy. Hyper-parameter tuning was done for similar parameters and values as model II.

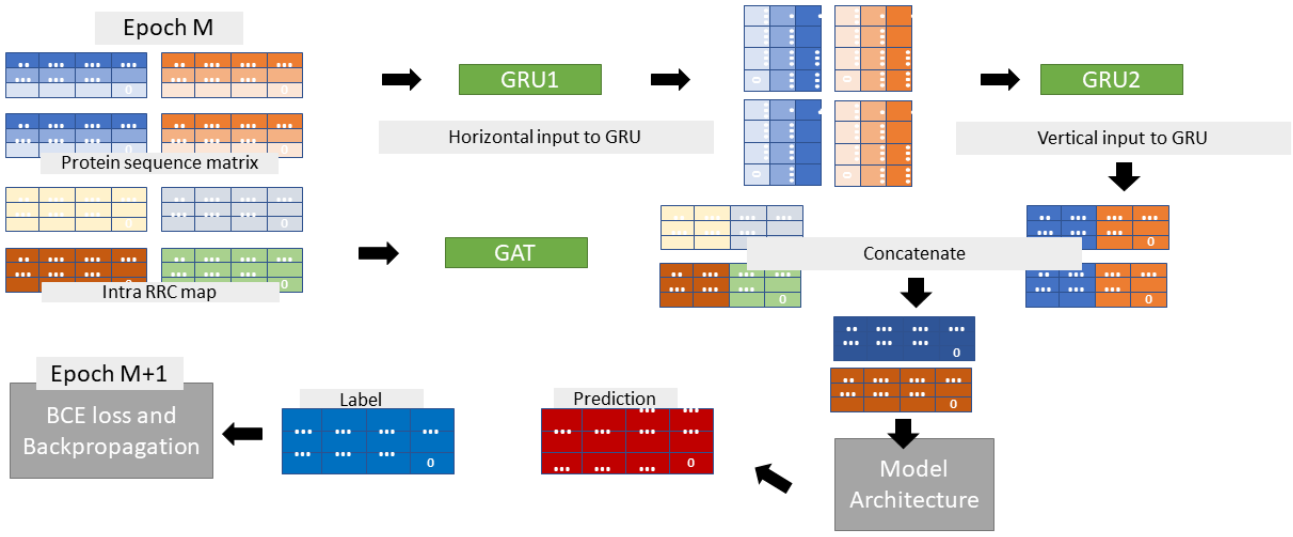

Figure 2.6: Model III: Multi-modality protein input and inter-protein RRC map output

2.3 Baseline Model

It is also a very important aspect of the project to choose and work on the baseline models or the SOTA methods to compare or evaluate the performance of the new model. The baseline method for the PPI prediction is PPI-detect[1]. PPI-detect is a website to predict PPI. The underlying algorithm is based on numerical encoding procedure to develop a support vector machine model.

To evaluate inter-protein RRC prediction, SOTA methods can be divided into two categories: unsupervised and supervised.

One of the unsupervised model is plmDCA or Pseudo-likelihood Maximization DCA [11]. This model works on the principle of coevolution among homologous proteins. Given two strings of protein, the model firstly performs MSA or Multi Sequence Alignment. The homologous proteins are arranged in each row in a way so that their correlations can be visualized in the best way and that process is called MSA. Based on the Pseudo-likelihood Maximization algorithm, the model calculates DCA which is described as Direct Coupling Analysis score between the amino acid positions. As the positions get higher scores, the probability of their interaction is higher.

From the PDB IDs of the proteins, the actual distal distances between these two residues' C_β atoms are extracted. If the distance is less than 8 Angstrom, the contact between two protein positions is considered positive. The output matrix of the unsupervised model as plmDCA can be considered and used as an important feature for the newly developed model. (**Figure 2.7**) shows the output matrix with DCA score between residues of protein A placed as rows and protein B placed as columns.

One of the supervised models which is filterDCA already uses the output of the plmDCA as one of the features; filterDCA takes the output of the plmDCA which is an adjacency matrix and filters it using several filters [12]. These filters are made from the secondary structure of the proteins. These filters can be considered as the graph modality or protein. (**Figure 2.7**) demonstrates the output of filterDCA.

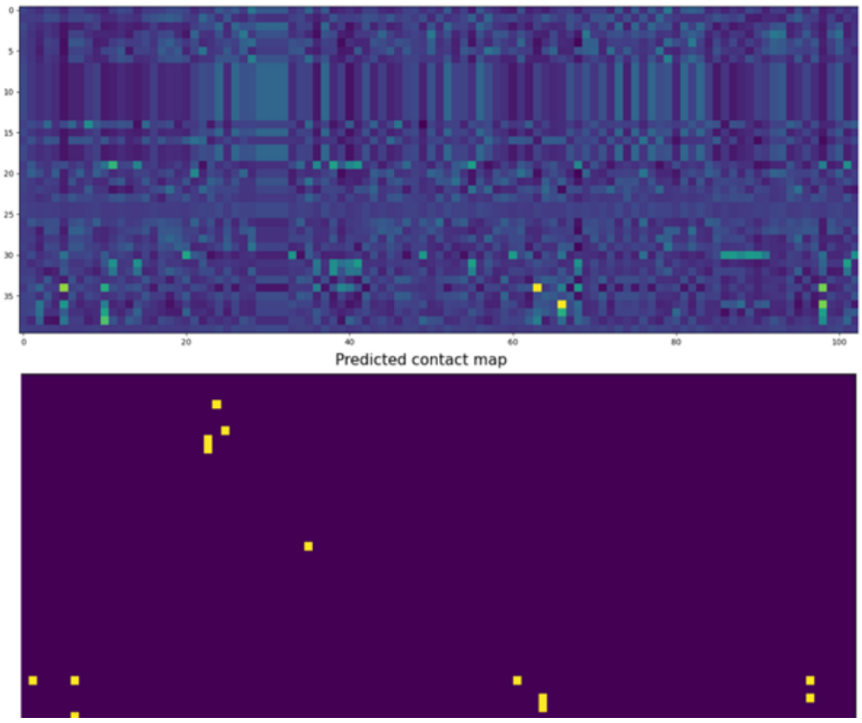

Figure 2.7: plmDCA output (up), filterDCA Output Matrix (down)

3. RESULTS

This chapter covering results is divided into two sections. In the first part, we would compare the PPI-prediction performance of model I with PPI-detect [1] over their dataset. In the second part, the PPI- and RRC-prediction performance of our models (model I, model II, and model III) would be evaluated on our curated datasets.

3.1 Comparative Evaluation of Model I with PPI-detect [1]

To evaluate the performance of our model we first use the area under the precision-recall curve (AUPRC) and the Area Under the Receiver Operating Characteristics (AUROC). Though PPI-detect [1] has published Precision-recall curves (PRC) but not the numerical values for the areas under the PRC. We thus compare our model I to PPI-detect using accuracy next. The AUPRC and AUROC of model I are still reported in the following **Table 3.1**.

Table 3.1: AUPRC and AUROC values of model I

Name	**Training**	**Validation**	**Easy test**	**Mid hard test**	**Hard test**
AUPRC	0.9981	0.769	0.847	0.640	0.769
AUROC	0.9983	0.691	0.912	0.768	0.691

Besides PRC curve PPI-detect [1] has also published Precision (Pr), Sensitivity(Sn) and Accuracy (Acc) of their mid-hard and hard test set. The calculations of these quantities are done as follows:

$$Pr = \frac{TP}{TP + FP}, \qquad Sn = \frac{TP}{TP + FN}, \qquad Acc = \frac{TP + TN}{TP + FP + TN + FN} \qquad \text{(Eq. 1)}$$

where TP = True positive and FP = False positive, FN = False negative, TN = True negative and FN = False negative.

The comparative values are shown below **Table 3.2**:

Table 3.2: Comparison of performance measures for model I with PPI-Detect hard test set (mid-hard + very-hard subsets)

Name	Precision	Sensitivity	Accuracy	Threshold to binarize
PPI-detect	0.554	0.648	0.661	0.5
Model I	0.529	0.779	0.646	0.5
Model I	0.597	0.710	0.702	g-mean

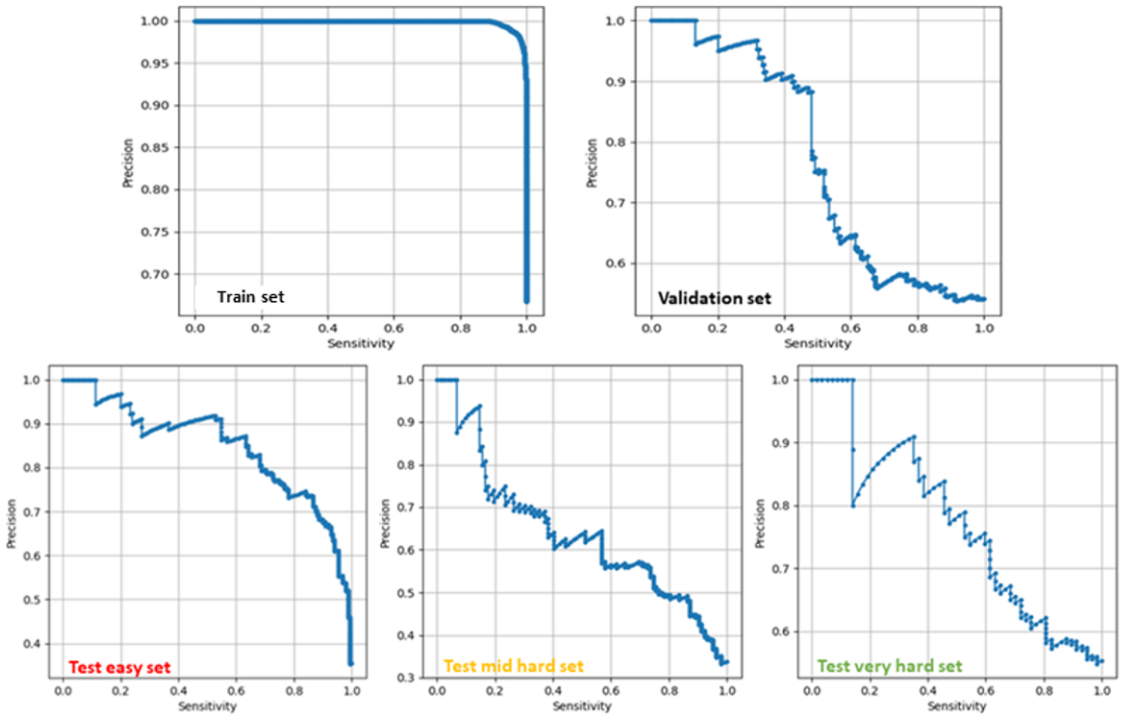

Figure 3.1: PRC curve obtained from model I on different dataset

3.2 Performance of Models for Our Curated Dataset

PPI-prediction AUPRC and AUROC of Model I are in **Table 3.3**. Even better accuracy was achieved for this larger PPI dataset compared to the benchmark from PPI-detect.

Before explaining the performance on inter RRC map, the ratio of positive interactions or average minority in the labeled inter RRC map is very important because it reflects the performance of a random classifier:

$$Random = \frac{\sum_{n=1}^{X} \frac{(PP)_n}{(L1)_n * (L2)_n}}{X} \qquad \text{(Eq. 1)}$$

Table 3.3: AUPRC and AUROC of Model I

Name	Training	Validation	Test
AUPRC	0.8850	0.738	0.776
AUROC	0.8858	0.734	0.797

where X is the number of interactions present in any dataset, $(PP)_n$ is the number of positive RR contacts in interaction n, $(L1)_n$ is length of protein 1 in interaction n and $(L2)_n$ is length of protein 2 in interaction n. In the case of PPI prediction, the ratios of positive and negative interactions were almost 0.5 and 0.5. However, in the case of RRC prediction, the ratio of residue contacts 0.11e-3, 0.0404e-3, and 0.0807e-3 in test, validation, and training sets of the 57 PPIs with structures, respectively. That is why the AUPRC and AUROC values in the case of PPI are much more different than those in the case of inter RRC map to be seen next.

The summary of inter RRC map AUPRC of our three models is given in **table 3.4**. We make the following observations.

Table 3.4: Comparative AUPRC of inter RRC map among Model I, II and III

Dataset	Average minority	Model I	Model II	Model III
Training	0.081e-3	1.12e-3	0.4	0.065
Validation	0.0404e-3	0.33e-3	6.56e-3	4.08e-3
Test	0.11e-3	1.09e-3	2.51e-3	2.78e-3

Firstly, we tried to also evaluate the performance of model I to predict the inter-protein RRC map. For model I, the AUPRC for validation and test were 0.33e-3 and 1.10e-3, almost 7.17 and 9 times better than random values (average minority) respectively **(Figure 3.2)**. The selected hyper-parameters are: sparsity regularization factor 0.01 and learning rate 1e-3.

Secondly, model II showed better RRC prediction results as expected. The AUPRC for test and validation are 6.56e-3, 2.51e-3 (161.38 and 21.82 times better than average minority) **(Figure 3.2)**. The selected hyper-parameters are: sparsity regularization factor 0.1e-3 and learning rate 1e-

3. These results show that supervised joint attentions are more accurate in predicting RRC maps than unsupervised ones in Model I.

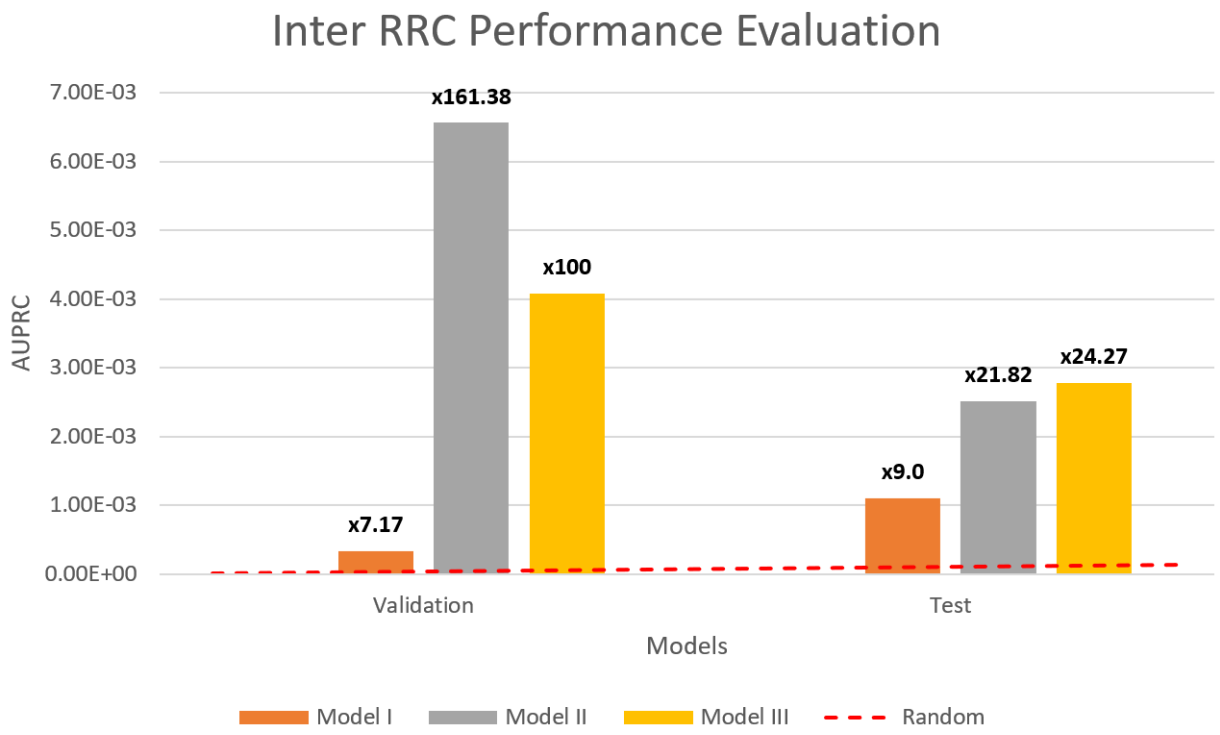

Figure 3.2: Inter RRC Performance of Model I, II, III and their performance improvement compared to random values (average minority)

Lastly, model III showed the best results for test set and significant improvement than average minority; the AUPRC are 4.08e-3 and 2.78e-3, respectively; again 100 and 24.27 times better performance than random values respectively (**Figure 3.2**). These results show that combining both modalities of protein sequences and structures, i.e., data fusion, can further improve the accuracy of RRC prediction.The selected hyper-parameters are: sparsity regularization factor 0.1e-3 and learning rate 0.1e-3.

As shown in **Table 3.4**, model II outperforms model III in validation data set AUPRC (61% better performance). This result can be explained by the training AUPRC values of both models. Model II training AUPRC is 0.4 which almost 5000 times greater than the random value and 5

times greater than model III training value. This big difference of training AUPRC between the two models shows, for the selected hyper parameter model II is more overfitting than model III. That is why, in future model II can be trained with a bigger value of sparsity regularization factor and model III can be trained with a smaller value. Though the inter RRC map predictions have not been compared with SOTA methods such as unsupervised plmDCA [11], filterDCA [12], and supervised ComplexContact [22], this is a part of the upcoming expansion of the project.

4. CONCLUSIONS AND FUTURE DIRECTIONS

From the comparative evaluation of the models, we can conclude that the layer embedding of HRNN works much better than the SOTA which is a vector machine based algorithm. On the other hand, the combinations of the sequences with structures have the best performance to predict inter-protein RRC maps.

This project has multiple future directions such as evaluating the prediction of PPI using only intra-protein RRC maps and the effect of pre-training (warm start). Another direction can be reforming the fine-tuning with structures with generalization and clustering. For further evaluation, the inter-protein RRC prediction can be compared with the SOTA methods (unsupervised and supervised). This project can be extended to predict inter-protein RRC distances and angles as well.

REFERENCES

[1] S. Romero-Molina, Y. B. Ruiz-Blanco, M. Harms, J. Münch, and E. Sanchez-Garcia, "Ppi-detect: A support vector machine model for sequence-based prediction of protein–protein interactions," *Journal of Computational Chemistry*, vol. 40, no. 11, pp. 1233–1242, 2019.

[2] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Séraphin, "The tandem affinity purification (tap) method: a general procedure of protein complex purification," *Methods*, vol. 24, no. 3, pp. 218–229, 2001.

[3] K. Young, "Yeast two-hybrid: so many interactions,(in) so little time...," *Biology of reproduction*, vol. 58, no. 2, pp. 302–311, 1998.

[4] J.-S. Lin and E.-M. Lai, "Protein–protein interactions: co-immunoprecipitation," in *Bacterial Protein Secretion Systems*, pp. 211–219, Springer, 2017.

[5] W. A. Abbasi and F. u. A. A. Minhas, "Problems in protein-protein interactions (a literature review)," 08 2018.

[6] G. Tradigo, F. Rondinelli, and G. Pollastri, *Algorithms for Structure Comparison and Analysis: Docking*. 01 2018.

[7] J. Drenth, *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

[8] D. S. Wishart, B. D. Sykes, and F. M. Richards, "Relationship between nuclear magnetic resonance chemical shift and protein secondary structure," *Journal of molecular biology*, vol. 222, no. 2, pp. 311–333, 1991.

[9] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali, "Protein structure fitting and refinement guided by cryo-em density," *Structure*, vol. 16, no. 2, pp. 295–307, 2008.

[10] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein–protein interactions. part i. experimental techniques and databases," *PLoS Comput Biol*, vol. 3, no. 3, p. e42, 2007.

[11] M. Ekeberg, T. Hartonen, and E. Aurell, "Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences," *Journal of Computational Physics*, vol. 276, pp. 341–356, 2014.

[12] M. Muscat, G. Croce, E. Sarti, and M. Weigt, "Filterdca: interpretable supervised contact prediction using inter-domain coevolution," *bioRxiv*, 2019.

[13] B. Turner, "irefweb: interactive analysis of consolidated protein interaction data and their supporting evidence," *Database : the journal of biological databases and curation*, vol. 2010, p. 023.

[14] P. Blohm, "Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis," *Nucleic acids research*, no. ue, p. 396–400.

[15] M. J. Meyer, J. Das, X. Wang, and H. Yu, "Instruct: a database of high-quality 3d structurally resolved protein interactome networks," *Bioinformatics*, vol. 29, no. 12, pp. 1577–1579, 2013.

[16] R. Mosca, A. Ceol, A. Stein, R. Olivella, and P. Aloy, "3did: a catalog of domain-based interactions of known three-dimensional structure," *Nucleic acids research*, vol. 42, no. D1, pp. D374–D379, 2014.

[17] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "Cd-hit: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[18] U. Consortium, "The universal protein resource (uniprot)," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D190–D195, 2007.

[19] S. M. Ireland and A. C. R. Martin, "atomium—a Python structure parser," *Bioinformatics*, vol. 36, pp. 2750–2754, 02 2020.

[20] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLoS computational biology*, vol. 13, no. 1, p. e1005324, 2017.

[21] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 1480–1489, Association for Computational Linguistics, June 2016.

[22] H. Zeng, S. Wang, T. Zhou, F. Zhao, X. Li, Q. Wu, and J. Xu, "ComplexContact: a web server for inter-protein contact prediction using deep learning," *Nucleic acids research*, vol. 46, no. W1, pp. W432–W437, 2018.

APPENDIX

All necessary codes and extracted data can be found in this github repository: https://github.com/Arghamitra/senior-design

For 2D modality interactions are:

Training set: (UniProt1 UniProt2 PDB)

P25963 Q04206 1nfi; O75531 O75531 1qck; Q96B26 Q9NQT4 2nn6; Q99497 Q99497 2r1v; Q14186 P06400 2aze; P32321 P32321 2w4l; Q13616 Q13616 1ldk; P05412 P01100 1s9k; Q13309 P63208 1fs1; P06730 P06730 2v8w; P12755 Q13485 1mr1; Q13309 Q13309 1fs2; P17676 P17676 1gtw; P06400 P06400 2r7g; P46063 P46063 2wwy; P12004 P12004 1vyj; Q13616 Q13309 1ldk; P36894 P36894 2goo; P31785 P60568 2b5i; P46527 P24941 1jsu; P16070 P16070 1uuh; P13861 P13861 2izx; P51149 Q96NA2 1yhn; Q9Y5X1 Q9Y5X1 2rai; P13010 P12956 1jeq; P22681 P22681 1b47; P49763 P49763 1fzv; P10415 P10415 2w3l; P84022 P84022 1mhd; P27487 P27487 3kwf;

Validation set: (UniProt1 UniProt2 PDB)

O95786 O95786 2qfd; P12830 P12830 3ff8; P45973 P45973 3i3c; O14745 P26038 1sgh; P19838 P19838 1svc; P24864 P24941 1w98; Q12933 Q15628 1f3v; P05067 P05067 1aap

Test set: (UniProt1 UniProt2 PDB)

Q8WXD5 Q9H840 1y96; Q13541 P06730 2v8w; P01730 P01730 1wio; P12956 P13010 1jeq; P37108 P49458 1e8o; P08670 P08670 1gk4; P35222 Q9NSA3 1t08; P25963 P19838 1nfi; P20963 P20963 2hac; Q13263 Q13263 2yvr; P61981 P61981 2b05; Q9Y6D9 Q9Y6D9 1go4; Q13363 Q13363 1mx3; P43351 P43351 1h2i; Q13485 Q13485 1g88; Q07817 Q07817 2p1l; P60709 P60709 3lue; O43187 O43187 3mop; P04156 P04156 3hj5