Predicting Chemical-Protein Interactions for Drug Discovery

Undergraduate Thesis

Submitted in complete fulfillment of the requirements of
BITS F421T Thesis

By

Amitesh Badkul
ID No. 2018B2A30764H

Under the supervision of:

Dr. Lei Xie
&
Dr. Durba Roy

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, HYDERABAD CAMPUS

December 2022

"Aut inveniam viam aut faciam."

Seneca

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, HYDERABAD CAMPUS

Abstract

Masters of Science (Hons.) in Chemistry

Predicting Chemical-Protein Interactions for Drug Discovery

by Amitesh BADKUL

The development of new drugs through traditional methods is a lengthy and expensive process, often taking over a decade and costing millions or even billions of dollars. However, the use of artificial intelligence (AI) in drug discovery has grown significantly in recent years due to the availability of more data, improved algorithms, and reduced costs. AI has the potential to substantially reduce the time and cost of discovering new drugs. Prediction of chemical protein interactions has proven to be valuable in the study of potential novel drugs. This thesis aims to validate the significance of deep learning algorithms in the field of chemical protein interaction for drug discovery and subsequently encourage the utilization of deep learning algorithms. The thesis discusses and evaluates a novel deep learning framework, Portal Learning of Chemical Genomics (PortalCG), for predicting chemical protein interaction, which generalizes well on unseen data compared to other deep learning approaches. However, the central aim of the thesis is to explore residue-residue contact, as they contribute significantly in understanding the protein folding, which helps in learning more about the structure of the protein. Implementation of the residue-residue contact model achieves a high accuracy of **98.81%** in the classification of these residue-residue contacts.

Acknowledgements

Firstly I want to express my immeasurable and deepest gratitude to my primary supervisor, Dr. Lei Xie, Professor, Department of Computer Science, Hunter College, CUNY, who was actively involved in providing knowledge of various biological and deep learning concepts without which these results wouldn't have been possible. Moreover, his constant guidance and supervision throughout the project proved instrumental. I would also like to extend my gratefulness to Dr. Durba Roy, Associate Professor, Department of Chemistry, BITS Pilani Hyderabad, for her continual support through her teachings and conversations to improve my understanding of the vast world of sciences. Along with this, I would like to thank Tian Cai, Ph.D. candidate, Department of Computer Science, Hunter College, who provided me with resources and fruitful discussions to learn about deep learning models in bioinformatics.

I would also like to thank my lab mates - Shuo Zhang and Yoyo Wu, for their help in various aspects of the project. I would also like to convey my heartfelt thankfulness to my family, my parents - Dr. Alok Badkul, Dr. Madhulika Jain, and my sisters - Avani and Anika, and my friends - Disha, Aadarsh, Akshat, Ameya, Anant, Anjur, Ishita, Kshitij, Nitya, Pravar, Rishav and Subhransu for their support and insightful conversations.

Contents

List of Figures

List of Tables

Abbreviations

PortalCG	**P**ortal Learning Of **C**hemical **G**enomics
DISAE	**DI**stilled **S**equence **A**lignment **E**mbedding
CPI	**C**hemical **P**rotein **I**nteraction
PDB	**P**rotein **D**ata **B**ank
AI	**A**rtificial **I**ntelligence
DTI	**D**rug **T**arget **I**nteraction
DNN	**D**eep **N**eural **N**etwork
NAMD	**NA**noscale **M**olecular **D**ynamics
CASP	**C**ritical **A**ssessment of **S**tructure **P**rediction
CNN	**C**onvolutional **N**eural **N**etwork
STL	**S**tep-Wise **T**ransfer **L**earning
MAML	**M**odel-**A**gnostic **M**eta-**L**earning
OOC-ML	**O**ut-**O**f-**C**luster **M**eta-**L**earning
LP	**L**inear **P**robing
FT	**F**ine Tuning
GLASS	**G**PCR-**L**igand **Ass**ociation
IUPHAR	**I**nternational **U**nion of Basic and Clinical **Phar**macology
KEGG	**K**yoto **E**ncyclopedia of **G**enes and **G**enomes

Chapter 1

Introduction

1.1 Proteins

Proteins are often described as the building blocks of life, as they are essential for most cellular functions, including but not limited to structural support, organ functioning, and regulation. Since a cell contains roughly 42 million proteins, it is fair to say that they participate in a considerably large amount of cellular functions [43] . Proteins are spatially large complex macromolecules of smaller one-dimensional units known as amino acids [6]. These amino acids are often used to represent the structure of the proteins [33]. Four different protein structure levels exist: 1) Primary, 2) Secondary, 3) Tertiary, and 4) Quaternary. The primary structure is defined as the sequence of amino acids in the proteins. There are over 190 million unique sequences in the UniProt dataset [44]. Along with the sequences, UniProt also contains information on but not limited to, its function, organism, subcellular location, and protein-protein interaction. Fig. 1.1 shows an example of primary structure representation. The sequence of the amino acids is crucial and critical in determining the function of the corresponding unique protein. The secondary structure of proteins refers to the protein's backbone and local structure.

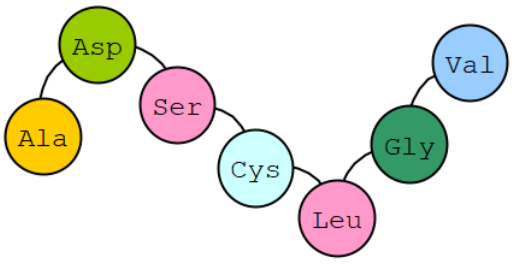

FIGURE 1.1: An illustration of a Primary Protein Structure

The secondary structure is further categorized as alpha helix, right-handed coiled strand, and beta sheets, consisting of nearly linear proteins connected via hydrogen bonding. Fig. 1.2A depicts the secondary structures. The tertiary structure refers to the three-dimensional structure of the proteins, which is the configuration the protein attains to obtain maximum stability. These 3D structures are present in the format of a PDB (Protein Data Bank) file, which consists of the x, y, and z of the various residues in the protein, which are obtained using methods like X-ray crystallography, Nuclear magnetic resonance, electron spectroscopy, and many others [3]. Fig. 1.2B is an example of a tertiary structure of the H-ras oncogene protein p21. Lastly, the quaternary structure combines numerous protein chains or subunits compactly packed [33]. Fig. 1.2C illustrates the packing of a quaternary structure.

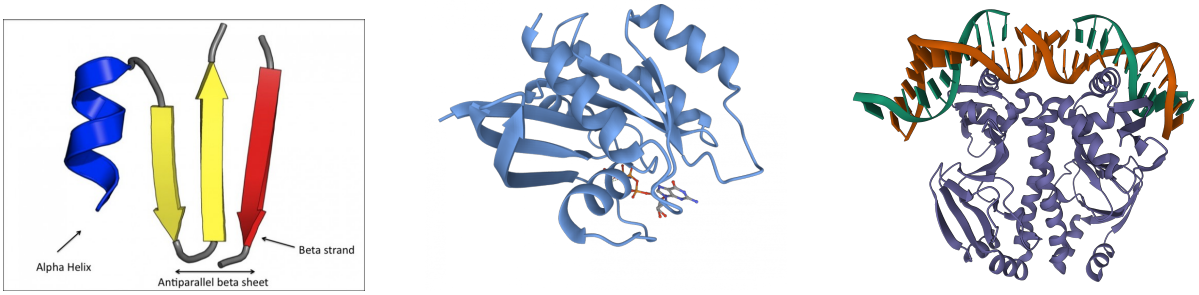

FIGURE 1.2: (A) An Example of representation of a Secondary Structure Protein; (B) Tertiary Structure of the H-ras oncogene protein p21; (C) An illustration of a Quaternary Structure. (From: https://www.ebi.ac.uk/)

Amino Acids	Three Letter Representation	Single Letter Representation
Glycine	GLY	G
Methionine	MET	M
Aspartate	ASP	R
Leucine	LEU	L

TABLE 1.1: A few examples of Amino Acids and their notations

```
>P05367
MKLLTSLVFCSLLLGVCHGGFFSFIGEAFQGAGDMWRAYTDMKEAGWKDGDKYFHARGNYDAAQRGPGGVWAAE
KISDARESFQEFFGRGHEDTMADQEANRHGRSGKDPNYYRPPGLPAKY
```

For quicker and more convenient notation purposes, each of the 20 amino acids is denoted with a single character and a three-letter characters-based representation. (ADD figure containing the codes and the amino acids). Generally, the FASTA format represents these single-letter notations [44].

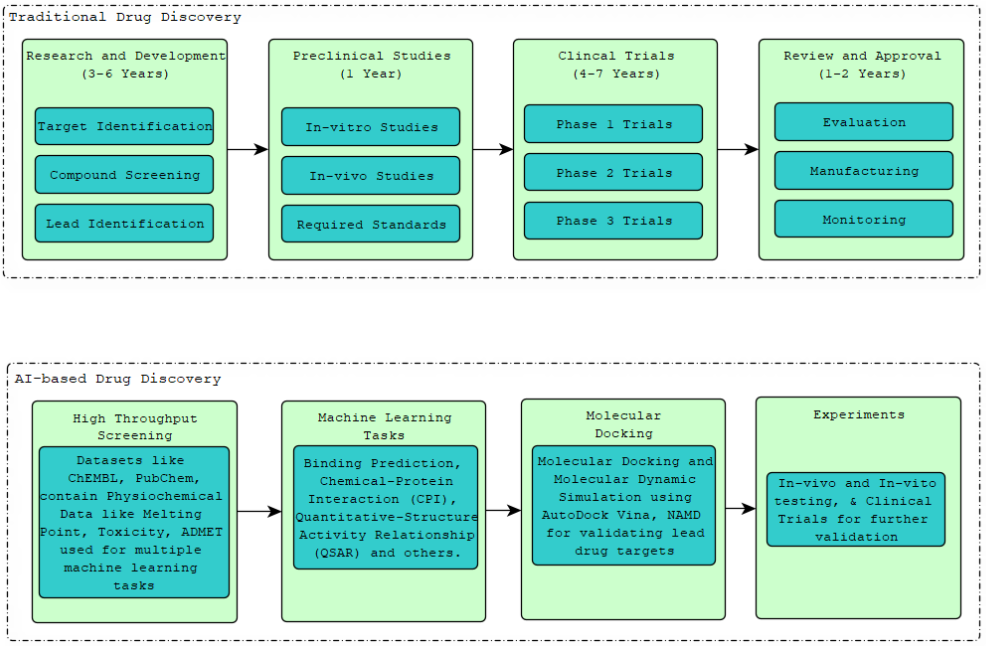

FIGURE 1.3: Comparison of Traditional Drug Discovery vs AI-based Drug Discovery

1.2 Drug Discovery

Drug discovery is the process of uncovering new pharmaceutical drugs. However, traditional drug discovery processes are very complex, take upto 15 years to finish, and cost millions, if not billions, of dollars [19]. The chemical space available for drug discovery is vast; hence, it is not feasible to test all of them. Multiple screenings of compounds from this expansive space are conducted through a comprehensive study, which includes screening the biochemical assay, physiochemical properties, and pharmacological properties [41]. Fig. 1.3 illustrates the general flow of drug discovery. Artificial Intelligence's (AI) application in drug discovery has expanded drastically due to increased available data, ever-improving algorithms, and reduced costs [35]. AI also substantially decreases the price and the duration of the drug discovery process [28]. The chemical space for developing novel drug compounds is enormous, and AI allows quicker and more efficient exploration [18]. Some of the notable works involve - AlphaFold, an attention-based deep neural network, is capable of predicting the structure of proteins from its protein sequences with high accuracy. Since most protein functions depend on their structural conformation, it is groundbreaking to predict structures accurately [20]. DeepChem is a python-based deep-learning library incorporating tools for drug discovery, biology, material sciences, and others. DeepChem has a wide range of diverse tools easily deployable for efficient deep learning [37]. Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC) is an effective reinforcement learning-based neural network for generating compounds with desired properties [39]. PotentialNet is a graph neural network-based model that can predict the binding affinity of proteins with ligands, thereby reducing the costs for physical experimentation [12].

1.3 Outline

The motivation for studying deep learning models for drug discovery is now evident. The rest of the thesis follows the understanding of the state-of-the-art deep learning framework PortalCG for drug discovery and chemical genomics. It describes experimental work done for refinements and improvements on PortalCG.

Chapter 2 describes chemical protein interactions, it's cellular perspective and quantification. It also illustrates the working of PortalCG and it's components briefly.

Chapter 3 covers the introduction, description of implementation and results of residue-residue contacts as a classification problem for improving PortalCG. It also concludes the thesis and discusses the future work.

Chapter 2

Chemical Protein Interaction

2.1 Chemical-Protein Interaction

Chemical-Protein Interaction or CPI refers to the interlinkage of proteins and chemicals. It is essential to understand the impact of these interactions as they affect biological processes like metabolism. Considering the high cost and the lengthy duration of the in-vitro and in-vivo experiments, the knowledge of CPIs, using artificial intelligence, thereby scaling down the costs and time, is paramount to drug discovery. Drug compounds work by interacting with proteins in our body, whether they bind or not, validating the drugs' effectiveness. Therefore, these chemical protein interactions are crucial in drug discovery, and traditional drug discovery involves in-vivo and in-vitro-based experiments, which are expensive and time-consuming. Instead, one can use deep learning models to predict the CPI. Another commonly used approach is molecular docking for predicting a suitable binding site. However, for the molecular docking approach, the spatial information of the protein structure is needed, which may not always be available. The table 2.1 depicts the works involving deep-learning based CPI models using protein sequence data. The review of these works indicates that these works utilize the following datasets:

1. GPCR-Ligand Association (GLASS) [7]
2. KIBA [42]
3. DrugBank [46]
4. International Union of Basic and Clinical Pharmacology (IUPHAR) [36]
5. Kyoto Encyclopedia of Genes and Genomes (KEGG) [21]
6. DAVIS [10]
7. ChEMBL [16]

8. BindingDB [17]

Almost all of the methods except for GraphDTA classify whether the chemical protein interact. However, GraphDTA attempts to predict the Binding Affinity of the interactions involved. Although these distinctive approaches have been successful to an extent, they also have limitations, which include the lack of generalizability on the vast protein families. In that respect, our new work, Portal Learning of Chemical Genomics (PortalCG) [4], successfully mitigates the problem of generalization using an end-to-end deep learning framework.

2.1.1 Chemical Protein Interaction: A cellular perspective

The process of chemical-protein interaction typically occurs through binding [14]. However, these interactions are of various kinds - ionic bonding, hydrogen bonding, hydrophobic interactions, and van der Waals forces. Hydrogen bonding generally forms when small molecules interact with proteins, consisting of hydrogen bonding sites (such as oxygen or nitrogen). Hydrogen bonding is critical in protein-ligand binding, as it facilitates cellular functions [8]. Ionic bonding refers to the interaction between oppositely charged species. In the case of chemical-protein ionic bonding, the bonding is formed between oppositely charged regions of chemical and protein. Hydrophobic interactions transpire when the water-repelling or non-polar regions of chemicals and proteins are attracted to each other and form a bond. It has been observed that hydrophobic interactions are a significant part of all the bonding interactions in high-efficiency ligands [13]. Van der Waals forces are short-range weak and attractive forces emerging between all molecules. The Van der Waals forces assist in stabilizing the protein-ligand complex [34].

All of these interactions, shown in the Fig. 2.1, partake in the chemical protein binding process. However, not all regions in a protein can bind with chemicals. Only certain specific regions are referred to as binding sites or active sites, as shown in Fig. 2.2(A), in proteins, where the binding occurs. These binding sites are formed due to the particular arrangement of amino acid residue in proteins. Moreover, binding sites may only be available under specific conditions of conformation or temperatures. Once the potential drug has bonded itself to the binding site on the protein, it can alter the protein's function by inhibiting or activating it and, in the process, change the structure of the protein, as shown in the figure 2.2(A).

2.1.2 Quantitative Measurement of Chemical-Protein Interaction

Now that chemical-protein interaction has been defined, it is crucial to understand how it is quantitatively measured. In the world of pharmaceuticals, a drug's effect on the human body and its proteins are often measured using various assays, which are laboratory-based experiments to quantify the biological activity of a drug or chemical. These experiments involve a biological

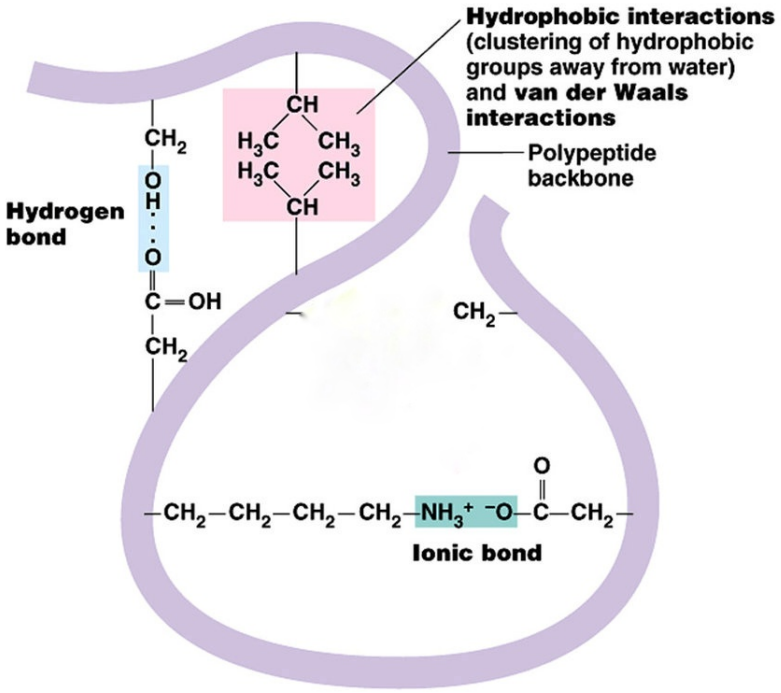

FIGURE 2.1: Various Bonding Interactions [15]

environment of the body, such as cells or enzymes, for measuring the drug's impact. In the case of the protein-chemical interaction, multiple biological assays provide vital information, and these include:

1. **Half Maximal Inhibitory Concentration (IC50)** - IC50 refers to the measurement of a chemical or drug's efficacy in inhibiting a specific target's activity. It is widely used in the field of pharmacology to assess the potency of drugs. The IC50 value conveys the concentration of a drug required to inhibit the target by 50% and is generally expressed in units of molarity. This measure helps compare different drugs' potency and determine the drug dose suitable for a given therapeutic effect.

2. **Dissociation Constant (Kd)** - The dissociation constant (Kd) estimates the binding strength of two substances. It is usually used to characterize the binding of a ligand to an enzyme. A lower Kd value implies a more powerful binding between the two substances. This is because a lower Kd value signifies that the concentration of the bound ligand is lower, indicating a stronger binding between the two substances.

3. **Half Maximal Effective Concentration (EC50)** - EC50 is another estimate to measure the potency of a potential drug. It is the concentration of the drug/compound at which it can produce a response that is 50% of the maximum response. In other words, EC50 measures a compound's effectiveness in producing a biological response. A lower EC50 value hints at a more potent compound, as it can produce a response at a lower concentration.

EC50 is typically expressed in units of concentration, such as milligrams per liter (mg/L) or micromoles per liter (μM).

4. **Inhibition Constant (Ki)** - The inhibition constant (Ki) is a measure of the affinity of a compound for its target. It measures how tightly the compound binds to the target and is typically used to describe the binding of an inhibitor to an enzyme. A lower ki value indicates a stronger binding between the inhibitor and the enzyme and, thus, a more potent inhibitor. The Ki value can be used to compare the potencies of different inhibitors for a given enzyme. It is typically expressed in units of concentration, such as millimolars (mM) or micromolars (μM).

2.1.2.1 Calculating $IC_{50}, K_i, K_d, EC_{50}$

Let P be the protein, C be the chemical. Let their simplified interaction be defined as follows:

$$\mathrm{P} + \mathrm{C} \rightleftharpoons \mathrm{PC} \tag{2.1}$$

Therefore, their association constant is calculated using the concentration of protein, chemical and their complex denoted as [P], [C], and [PC].

$$K_a = \frac{[PC]}{[P][C]} \tag{2.2}$$

Their dissociation constant or K_d is defined as the inverse of K_a.

$$K_d = \frac{[P][C]}{[PC]} \tag{2.3}$$

K_i is often defined using the same formula (1.3), when the protein is an enzyme and the chemical is a drug and the inhibition is competitive in nature.

IC_{50} and K_i are defined in the same scenario, therefore, for a case of competitive binding the Michealis-Menten kinetics and K_i is used to estimate IC_{50}.

$$IC_{50} = K_i(1 + \frac{[S]}{K_m}) \tag{2.4}$$

Model	Description	Dataset	Threshold
TransformerCPI[9]	Uses a transformer-based architecture with a modified self-attention mechanism for better sequence information processing.	GLASS, KIBA	IC_{50} <1 μM
DeepConv-DTI [25]	Utilizes a CNN to capture patterns from the protein sequences on generalized protein families, thereby successfully using the raw protein sequence.	DrugBank, IUPHAR, KEGG	$K_d < 10\mu M$
GraphDTA [31]	Employs a graph neural network-based model to obtain drug representation and a convolution-based approach on the protein sequence for obtaining the protein sequence representation.	DAVIS, KIBA	-
DeepDTI [26]	Implements a Deep-belief network-based model and can perform reasonably well on unseen data.	BindingDB	$IC_{50} < 50\mu M$
DNN-DTI [47]	Deploys a deep neural network using transcriptome data from gene and drug data.	GSD	$IC_{50} < 10\mu M$
DISAE [5]	Implements a state-of-the-art natural language processing (NLP) algorithm for whole genome chemical protein interaction.	ChEMBL23, BindingDB24, GLASS25, DrugBank	$IC_{50} < 5\mu M$

TABLE 2.1: Noteworthy Works in CPI

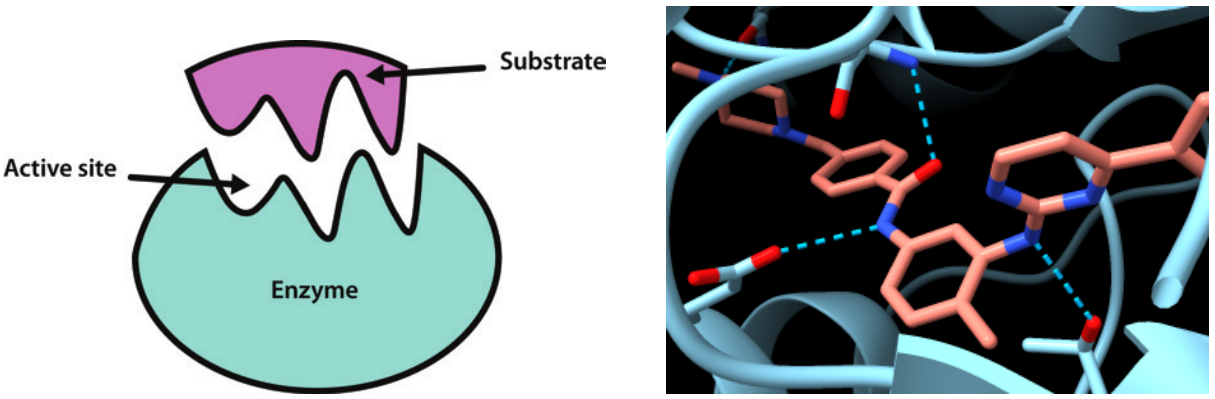

FIGURE 2.2: (A) Binding Site/ Active Site, (B) Interaction of Chemical (red) Protein (blue)

2.2 PortalCG

This section briefly describes PortalCG and it's various deep learning components [4]. PortalCG is a deep learning framework utilizes novel algorithms to learn improved protein representations and effectively predict dark protein family ligands for the discovery of new drugs. PortalCG consists of three major components: 1) Structure-enhanced Sequence Pre-Training, 2) End-to-End Sequence-Structure-Function Step-Wise Transfer Learning and 3) Out-of-cluster Meta-Learning.

PortalCG is accurately able to predict the interaction of ligands with similar scaffold to Dopamine D1 and D3 receptors, which have been suggested to counter the Opioid use disorder (OUD), which could be used as potential drugs.

2.2.1 Dataset

The multiple components of the PortalCG are trained on multiple widely-used datasets in bioinformatics. List of datasets utilized:

1. DISAE: Pfam [29]
2. STL: BioLP [48] and PDB [2]
3. OOC-ML: ChEMBL [16]

2.2.2 Structure-enhanced Sequence Pre-Training

The protein sequences and human languages are similar in aspects to the deep learning models. Hence many researchers have attempted to utilize Natural Language Processing (NLP) models to develop pre-trained protein language models such as TAPE [38], and ProtTrans [11]. The process of pre-training refers to a deep learning model's approach to learning a language representation via self-supervised learning.

DIASE [5], a deep learning framework, is deployed in the first step of Structure-enhanced Sequence Pre-Training. It utilizes the multiple sequence alignments (MSA) and explicitly uses the evolutionarily significant MSA positions and protein sequences to understand protein structure-function relations without its structural information effectively. DISAE utilizes A Lite BERT (ALBERT) model [24], a highly effective model for learning language representations, as its architecture. Fig. 2.3 depicts the general pre-training process on the protein sequences. The distilled triplet representation is obtained using the MSA and the protein sequence. The pre-training task is a masked language modeling (MLM) task; therefore, around 15% of the distilled triplet sequences are masked. The DISAE model attempts to predict these to improve its

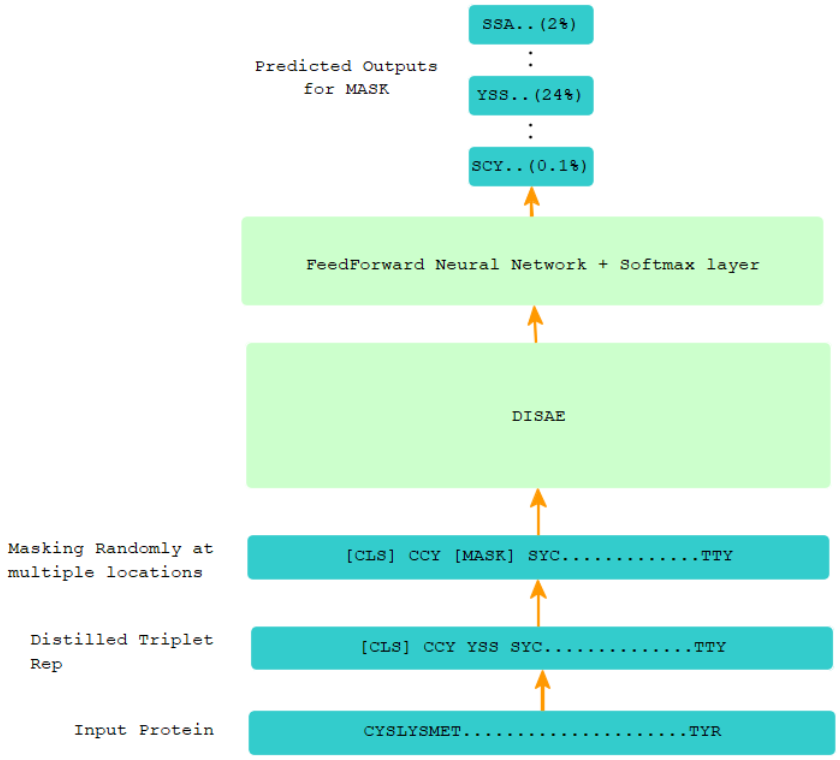

FIGURE 2.3: DISAE Architecture

protein representation knowledge. DISAE uses about 40 million protein sequences for pre-training on the ALBERT model.

2.2.3 End-to-End Sequence-Structure-Function Step-Wise Transfer Learning (STL)

The second step incorporates 3-Dimensional information about the proteins by attempting to predict atom-residue interaction as a binary classification problem. The pre-trained protein descriptor model from the first component is fine-tuned on the Atom-Residue data. This helps the protein descriptor understand the physical space of proteins. Around 30,000 proteins and their potential ligand binding sites at residues are obtained from the BioLP dataset, and approximately 13,000 ligands are used. The ligands are represented as the simplified molecular-input line-entry system (SMILES) [45], and their embedding is obtained by pre-training on a graph isomorphism network (GIN) model. Lastly, the ligand and protein embeddings are multiplied to get a final matrix input through a multi-layer perceptron and a softmax function to classify whether they ultimately bind. Fig. 2.4 illustrates the working of the atom-residue fine-training model.

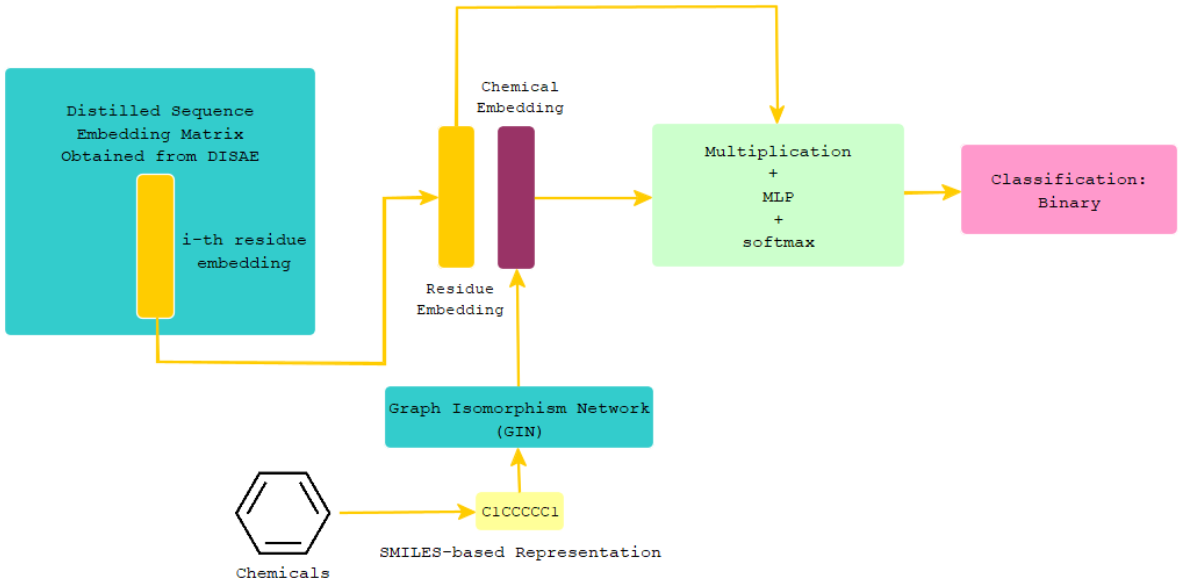

FIGURE 2.4: Atom-Residue Architecture

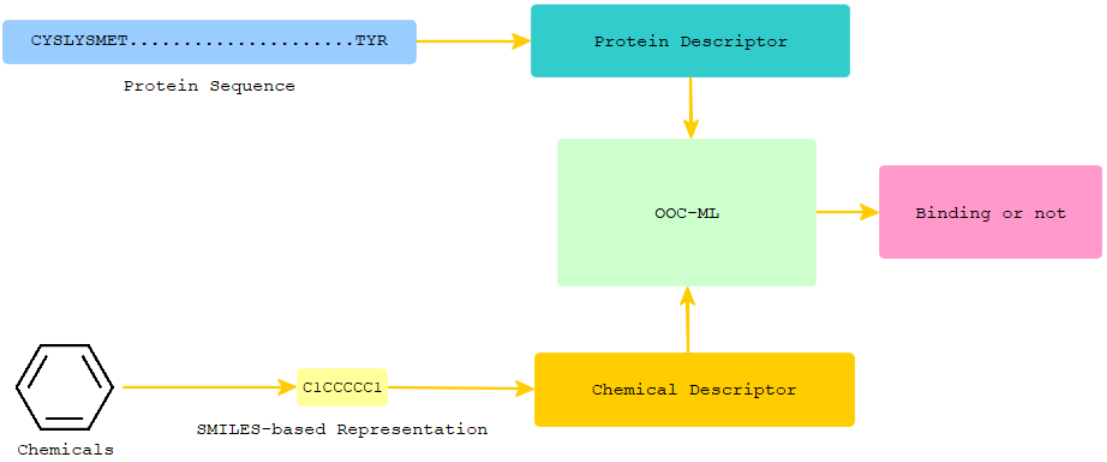

FIGURE 2.5: OOC-ML Architecture

2.2.4 Out-of-cluster Meta-Learning (OOC-ML)

In the last step we perform Out-of-cluster Meta-Learning for predicting CPI, the OOC-ML algorithm gets trained on a small batch of various proteins from the same pfam family for all the pfam families present in the dataset. This helps the deep learning models to generalize well on out-of-distribution data (that refers to data different from training data). Fig. 2.5 illustrates the working of the same.

Chapter 3

Residue-Residue Contacts

3.1 Introduction

Protein Residue-Residue contact, or inter-residue contact, refers to the residue pairs that are within a certain pre-defined proximity of each other. Since the protein's 3D structural information can be obtained from their respective PDB files, the residue's location can be pinpointed using the protein's x, y, and z coordinates. More precisely, two residues are said to be in contact if the distance between the $C\beta$ atoms is less than **8Å**. The Critical Assessment of protein Structure Prediction (CASP) [22] defines the proximity for residue-residue contact less than 8 Angstrom and the sequence's degree of separation of the residues. A larger sequence separation threshold would mean lesser residue contacts [1].

Generally, contacts are classified into three ranges: short-range contact, medium-range contact, and long-range contact. Fig. 3.1 illustrates the categorization of residue-residue contacts. Out of the three categories of contacts, long-range contacts are highly beneficial because of the spatial constraint information learned [30]. Due to this, most of the methods focus on the prediction of these long-range contacts.

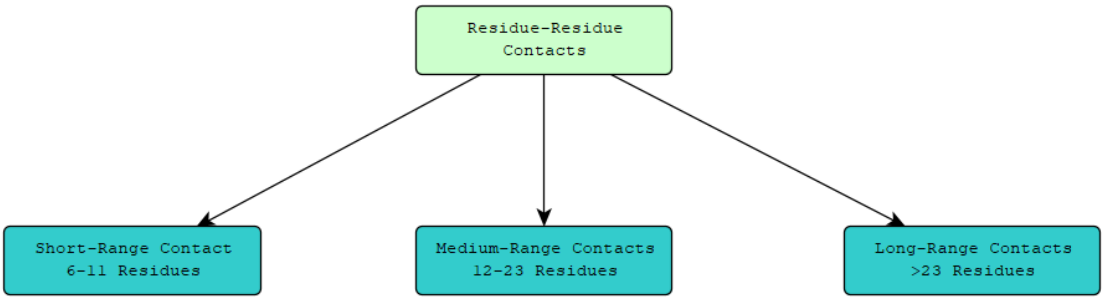

FIGURE 3.1: Residue-Residue Contacts

There are five major methods that are used to predict the long-range contacts:

1. Coevolution-Derived Information-based approach [32]
2. Machine Learning-based approach [49]
3. Template-based approach [50]
4. Physiochemical Information-based approach [27]
5. Hybrid approach [40]

3.1.1 Why are residue-residue contacts of interest?

A vital aspect of CPI is a suitable representation of proteins for deep-learning algorithms, ensuring high accuracy of the predicted CPIs. A commonly used method for obtaining an appropriate representation includes learning patterns from residue-residue contact information. Moreover, these residue-residue contacts provide information on proteins' likely chemical binding sites and their physical properties. The utilization of residue-residue contact prediction has vastly reduced computational complexity in predicting protein structures as it restricts the possible number of conformations the residue could have within the protein [30]. AlphaFold incorporates a component involving training a deep learning model on residue-residue contact information and can ultimately predict the 3D structures of proteins with remarkable accuracy [20].

3.2 Residue-Residue Contact Training-based fine-training

While PortalCG generalizes well on unseen data and can illuminate the dark protein space of proteins, we implement the residue-residue contact fine-training-based method to assist in improving the protein descriptors for PortalCG. Residue-residue contact incorporates the euclidean distances and provides physical constraints to the protein descriptor to identify from and potentially improve it. As part of fine-training, the residue-residue contact prediction problem is posed as a binary and multi-class classification. Fine-tuning uses the DISAE-based protein embedding from PortalCG's initial step. The dataset used for fine-tuning consists of approximately 6000 proteins, each composed of 80 residue contacts. As part of fine-tuning, the residue-residue contact prediction problem is posed as a binary and multi-class classification, as explained below:

1. Binary Classification: To classify whether the residues are in contact or no.
2. Multi-class Classification: To classify in various classes, ranging from 3 to 10 classes, based on the distances.

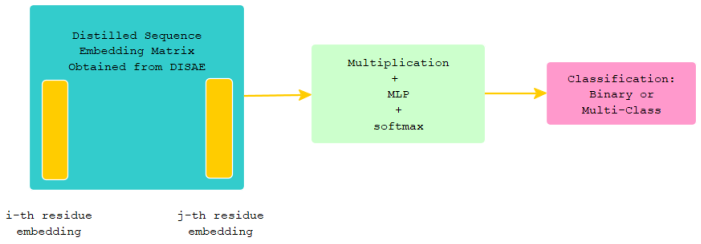

FIGURE 3.2: Residue-Residue Architecture Implementation

3.2.1 Implementation

The model obtained by the first-step training is used as a starting point for the pre-training. Since the DISAE-based protein descriptor output position-specific embeddings of a distilled protein sequence and pair-wise interaction features of the residues were produced with a simple vector operation: a matrix multiplication was used to select embedding vectors of each residue. This matrix was fed into a straightforward feed-forward layer for either binary or multi-class classification. We utilize the "accuracy" evaluation metric for measuring the performance of the deep learning models. Fig. 3.2 depicts the architecture of Residue-Residue model.

3.2.2 Results and Discussion

Any machine learning or deep learning model consists of parameters that control the model training process, and these are referred to as hyperparameters. Unlike the model parameters that are updated when training the model, the hyperparameters can't be modified during the process. For example, model parameters would be the weights of the various neurons of the network, which impact the final prediction. In contrast, model hyperparameters would be the number of neurons in the model architecture. The model parameters are optimized through the process of training. However, the hyperparameters are optimized (or 'tuned') using experimentation, which includes measuring the changes to the evaluation metric by modifying the hyperparameters. Another commonly used approach is searching through literature for similar works and using the hyperparameters deployed by these works. The model training involved hyperparameter-tuning to obtain the best result.

The various hyperparameters changed include:

1. Classification Type: Binary or Multi-class
2. Learning Rate: Decides the rate of modifying the weights of neurons
3. Frozen Layers: Layer freezing refers to the process of restraining the weight updation of neurons to reduce computational costs and preventing the data from overfitting on the

training data and failing to generalize on unseen data. In our case, the frozen options include - whole, none, or partial freezing.

4. Epochs: It is the amount of times a machine learning or deep learning model is trained for on the entire training dataset

Model	Epochs	Gloabl Steps	Frozen	Classification Type	Accuracy (%)
Model 1	1000	50	Whole	Binary	67.5
Model 2	40000	200	Whole	Binary	80.49
Model 3	40000	200	Whole	Multi - 10 Class	43.09
Model 4	40000	200	Whole	Multi - 4 Class	60.24
Model 5	40000	200	Whole	Multi - 6 Class	49.31
Model 6	40000	200	Whole	Multi - 8 Class	44.12
Model 7	40000	200	None	Binary	95.39
Model 8	40000	200	Partial	Binary	96.08
Model 9	40000	200	Partial-Whole	Binary	92.92
Model 10	40000	200	Partial-Whole	Multi - 10 Class	64.05
Model 11	**100000**	**500**	**LPFT**	**Binary**	**98.81**

TABLE 3.1: Results of Various Models

Fig. 3.3 (top) dictates the need for more training as the loss and accuracy keep decreasing and increasing, respectively. The number of epochs is increased to 40000, totaling 200 global steps. The increase in accuracy and the decrease in loss are evident. Fig. 3.3 (bottom) reflects the same.

Fig. 3.4 (top) compares multi-class vs. binary classification. It validates the hypothesis that multi-class classification is relatively more complex for the model to perform because of the fine divisions of data and because of the complexity of mathematically modelling the classification. Fig. 3.4 (bottom) also accredits the previous hypothesis.

Fig. 3.5 (top) illustrates the effect of unfreezing all the model layers. It is noted that the model fails to generalize on the testing dataset, implying that the model is overfitting on the training dataset. Fig. 3.5 (bottom) compares no-layer freezing, partial-layer freezing, and whole-layer freezing. The partial-layer freezing includes freezing layers from 1 to 15 of the model, which has a similar effect to the no-layer freezing model. It overfits the training data but not as much as the no-layer model. To circumvent the problem of overfitting, a novel dynamic layer freezing algorithm is implemented, which incorporates switching of layer freezing from none to the whole, or partial to the whole when the training accuracy is more than the testing accuracy to avoid overfitting. Fig. 3.6 (top) illustrates the effect of including dynamic freezing. The testing and training accuracy closely follow each other. Fig. 3.6 (bottom) compares binary vs. multi-class classification with the implementation of dynamic freezing.

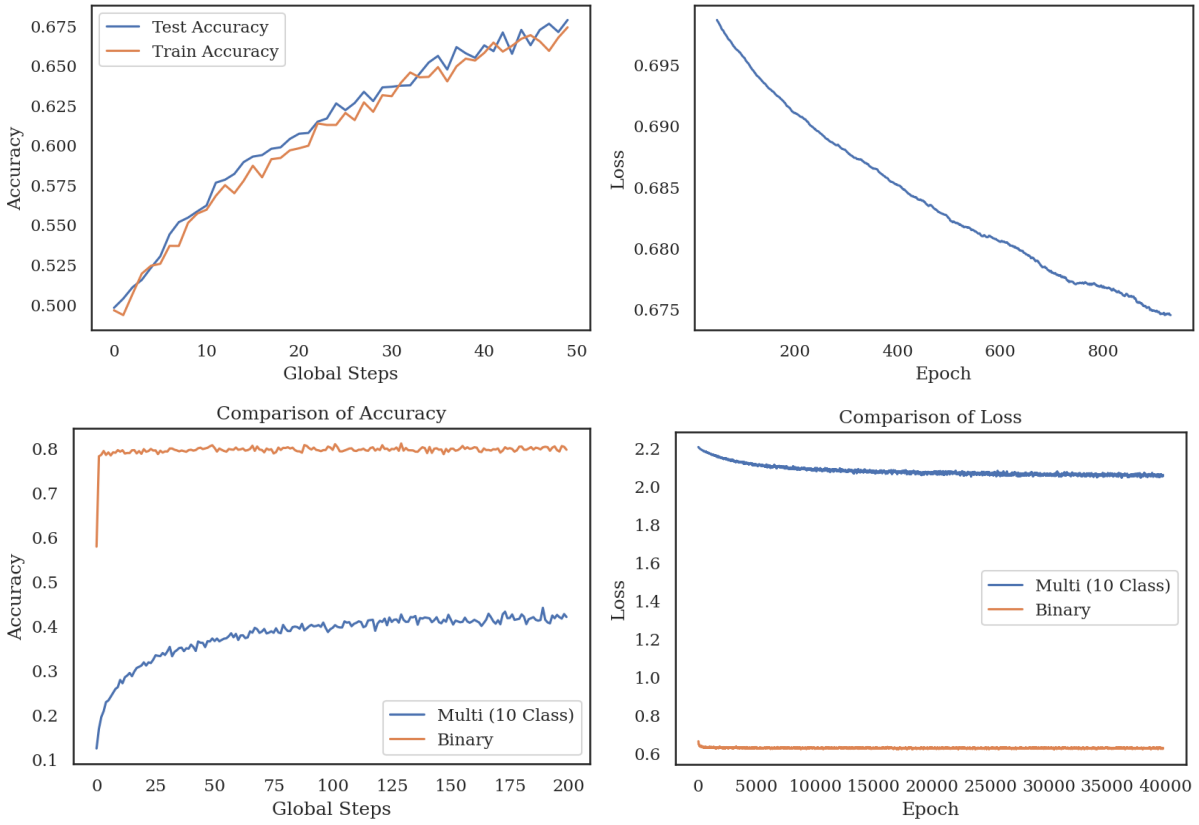

FIGURE 3.3: Effect of varying epochs

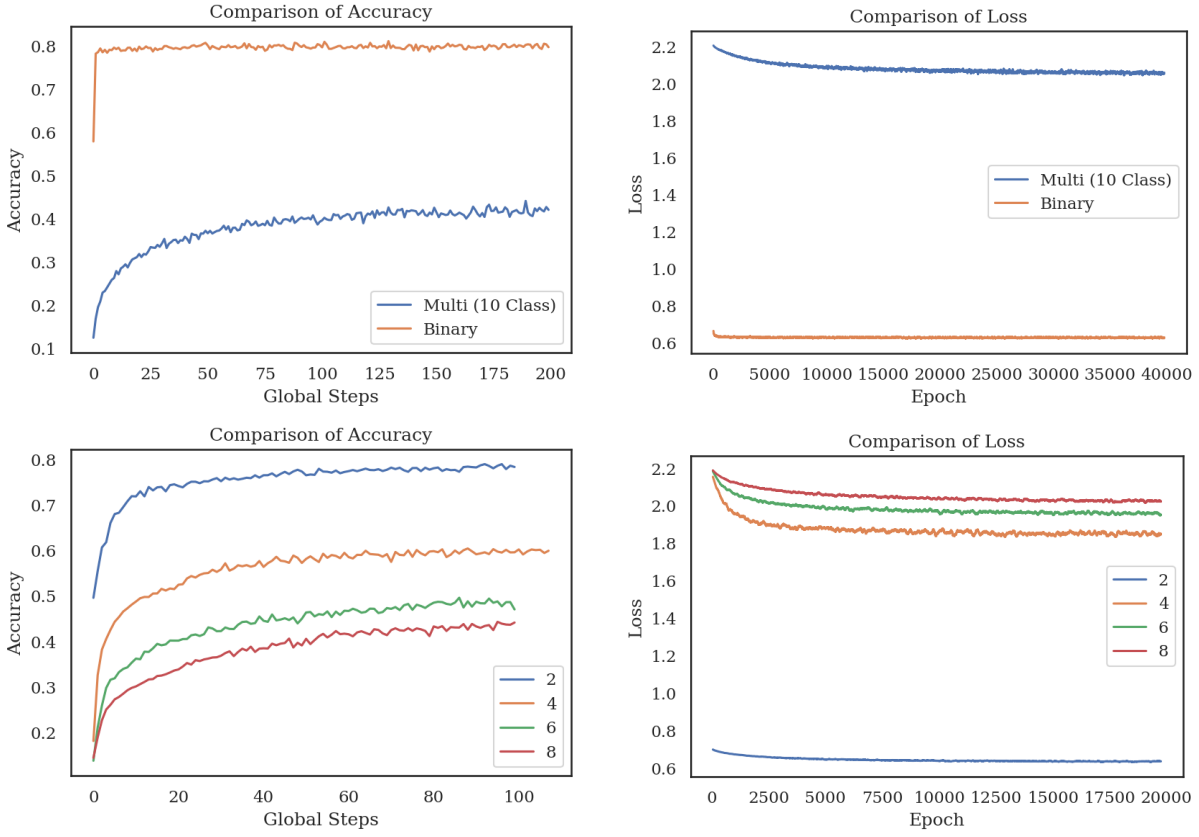

FIGURE 3.4: Effect of multi-class classification

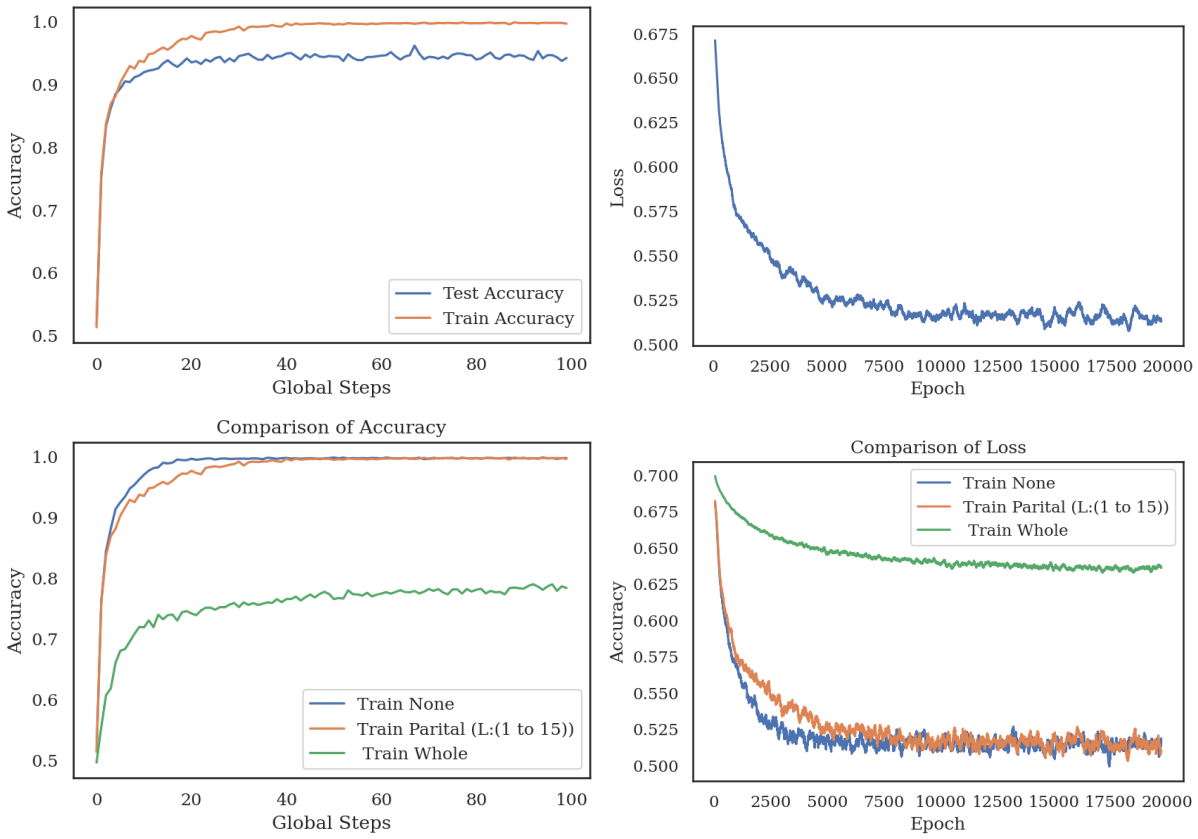

FIGURE 3.5: Effect of freezing and unfreezing layers

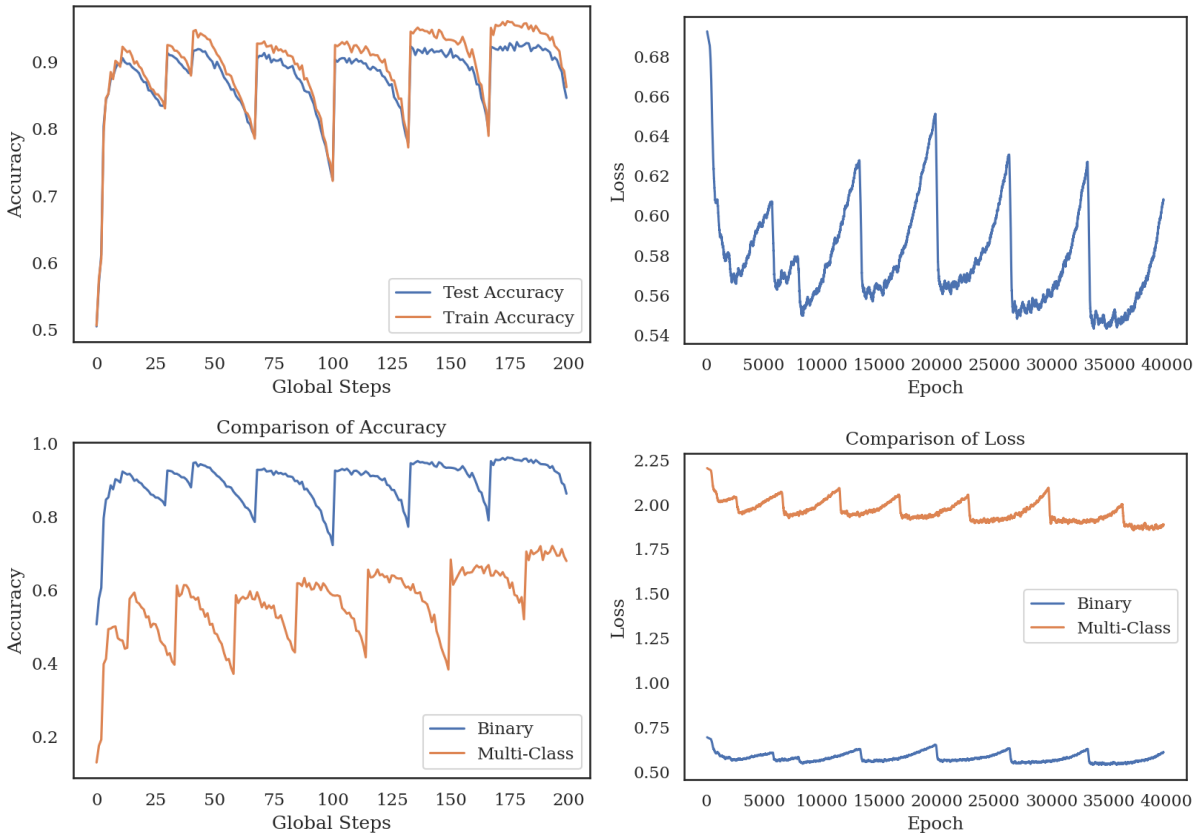

FIGURE 3.6: Dynamic Layer Freezing Effect

3.3 Fine-Tuning Strategy

To improve the performance of fine-tuning the pre-trained model, we propose using the strategy of linear probing and then fine-tuning. This strategy has been proposed by Kumar et al. [23] for fine-tuning computer vision models. This approach first follows training the model while unfreezing only the head, then re-training it with all the unfrozen layers. The Fig. 3.7 illustrates the approach.

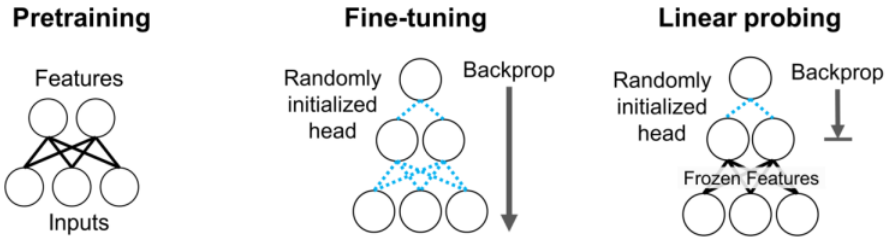

FIGURE 3.7: Linear Probing, Fine-Tuning Visualized

After following the linear probing followed by fine-tuning (LP-FT) strategy, it has been observed that the model generalizes better on unseen data, hence creating a better model. Fig. 3.8 illustrates the same. Implementing the same in our model, we observed similar results. Although over-fitting occurred, it was significantly lesser than in the earlier model, and the training and testing accuracy was within 2.7% of each other compared to the 8.4%. Moreover, the

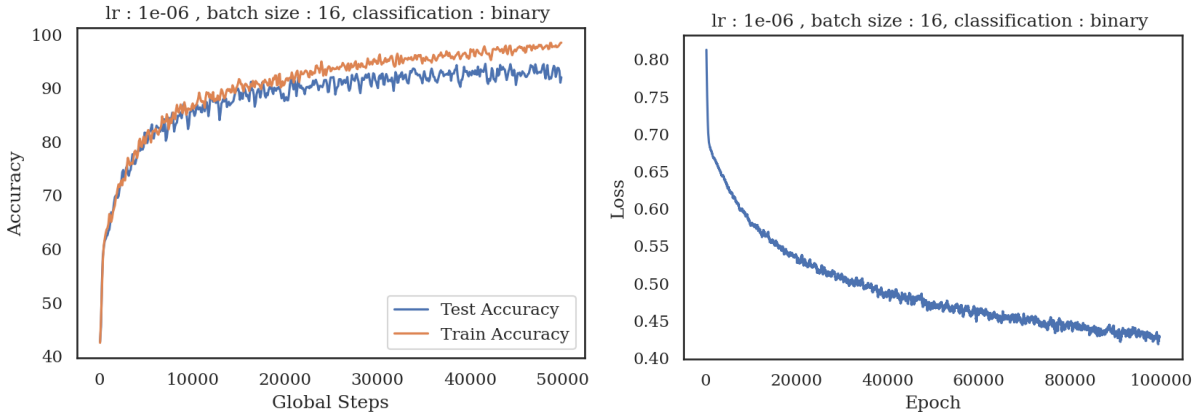

FIGURE 3.8: Effect of fine-tuning strategies

3.4 Future Work

The future work involves incorporation of the residue-residue model to the PortalCG pipeline to improve its performance. Moreover, also explore predicting the binding affinity value instead of predicting whether the protein and chemicals interact or not. This will allow for better screening

of potential drug molecules because the binding affinity values between the protein and chemical would be known.

3.5 Conclusion

The classification model for residue-residue contact prediction on pre-trained DISAE model achieves an accuracy of 65% and with the help of hyper-parameter tuning and different training strategies it is able to achieve 98.8% of accuracy. The high performance of the deep learning neural network indicates its effectiveness and efficiency. The ability of deep neural networks to identify and learn the complicated patterns present in large amounts of data makes them a powerful tool. Moreover, deep neural networks can learn abstract representations of compounds, proteins, and many more objects. This thesis validates the effectiveness of deep learning in the field of drug discovery.

Bibliography

[1] Badri Adhikari and Jianlin Cheng. "Protein residue contacts and prediction methods". In: *Data Mining Techniques for the Life Sciences*. Springer, 2016, pp. 463–476.

[2] Helen M Berman et al. "The protein data bank". In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[3] Stephen K Burley et al. "RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy". In: *Nucleic acids research* 47.D1 (2019), pp. D464–D474.

[4] Tian Cai et al. "Binding Site-enhanced Sequence Pretraining and Out-of-cluster Meta-learning Predict Genome-Wide Chemical-Protein Interactions for Dark Proteins". In: *bioRxiv* (2022).

[5] Tian Cai et al. "MSA-regularized protein sequence transformer toward predicting genome-wide chemical-protein interactions: application to GPCRome deorphanization". In: *Journal of chemical information and modeling* 61.4 (2021), pp. 1570–1582.

[6] Nigel Chaffey. *Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell. 4th edn.* 2003.

[7] Wallace KB Chan et al. "GLASS: a comprehensive database for experimentally validated GPCR-ligand associations". In: *Bioinformatics* 31.18 (2015), pp. 3035–3042.

[8] Deliang Chen et al. "Regulation of protein-ligand binding affinity by hydrogen bond pairing". In: *Science advances* 2.3 (2016), e1501240.

[9] Lifan Chen et al. "TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments". In: *Bioinformatics* 36.16 (2020), pp. 4406–4414.

[10] Mindy I Davis et al. "Comprehensive analysis of kinase inhibitor selectivity". In: *Nature biotechnology* 29.11 (2011), pp. 1046–1051.

[11] Ahmed Elnaggar et al. "ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing". In: *arXiv preprint arXiv:2007.06225* (2020).

[12] Evan N Feinberg et al. "PotentialNet for molecular property prediction". In: *ACS central science* 4.11 (2018), pp. 1520–1530.

[13] Renato Ferreira de Freitas and Matthieu Schapira. "A systematic analysis of atomic protein–ligand interactions in the PDB". In: *Medchemcomm* 8.10 (2017), pp. 1970–1981.

[14] Ran Friedman. "Computational studies of protein–drug binding affinity changes upon mutations in the drug target". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 12.1 (2022), e1563.

[15] Carmen Gaidau et al. "Wool keratin hydrolysates for bioactive additives preparation". In: *Materials* 14.16 (2021), p. 4696.

[16] Anna Gaulton et al. "The ChEMBL database in 2017". In: *Nucleic acids research* 45.D1 (2017), pp. D945–D954.

[17] Michael K Gilson et al. "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology". In: *Nucleic acids research* 44.D1 (2016), pp. D1045–D1053.

[18] Rohan Gupta et al. "Artificial intelligence to deep learning: machine intelligence approach for drug discovery". In: *Molecular Diversity* 25.3 (2021), pp. 1315–1360.

[19] James P Hughes et al. "Principles of early drug discovery". In: *British journal of pharmacology* 162.6 (2011), pp. 1239–1249.

[20] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[21] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[22] Andriy Kryshtafovych et al. "Critical assessment of methods of protein structure prediction (CASP)—Round XIV". In: *Proteins: Structure, Function, and Bioinformatics* 89.12 (2021), pp. 1607–1617.

[23] Ananya Kumar et al. "Fine-tuning can distort pretrained features and underperform out-of-distribution". In: *arXiv preprint arXiv:2202.10054* (2022).

[24] Zhenzhong Lan et al. "Albert: A lite bert for self-supervised learning of language representations". In: *arXiv preprint arXiv:1909.11942* (2019).

[25] Ingoo Lee, Jongsoo Keum, and Hojung Nam. "DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences". In: *PLoS computational biology* 15.6 (2019), e1007129.

[26] Guannan Liu et al. "GraphDTI: A robust deep learning predictor of drug-target interactions from multiple heterogeneous data". In: *Journal of Cheminformatics* 13.1 (2021), pp. 1–17.

[27] Alberto JM Martin et al. "RING: networking interacting residues, evolutionary information and energetics in protein structures". In: *Bioinformatics* 27.14 (2011), pp. 2003–2005.

[28] Vijay Mishra. "Artificial intelligence: the beginning of a new era in pharmacy profession". In: *Asian Journal of Pharmaceutics (AJP)* 12.02 (2018).

[29] Jaina Mistry et al. "Pfam: The protein families database in 2021". In: *Nucleic acids research* 49.D1 (2021), pp. D412–D419.

[30] Bohdan Monastyrskyy et al. "Evaluation of residue–residue contact prediction in CASP10". In: *Proteins: Structure, Function, and Bioinformatics* 82 (2014), pp. 138–153.

[31] Thin Nguyen et al. "GraphDTA: Predicting drug–target binding affinity with graph neural networks". In: *Bioinformatics* 37.8 (2021), pp. 1140–1147.

[32] Sergey Ovchinnikov et al. "Improved de novo structure prediction in CASP 11 by incorporating coevolution information into Rosetta". In: *Proteins: Structure, Function, and Bioinformatics* 84 (2016), pp. 67–75.

[33] Clare M O'Connor, Jill U Adams, and Jennifer Fairman. "Essentials of cell biology". In: *Cambridge, MA: NPG Education* 1 (2010), p. 54.

[34] C Nick Pace, J Martin Scholtz, and Gerald R Grimsley. "Forces stabilizing proteins". In: *FEBS letters* 588.14 (2014), pp. 2177–2184.

[35] Debleena Paul et al. "Artificial intelligence in drug discovery and development". In: *Drug discovery today* 26.1 (2021), p. 80.

[36] Roger Guy Pertwee et al. "International Union of Basic and Clinical Pharmacology. LXXIX. Cannabinoid receptors and their ligands: beyond CB1 and CB2". In: *Pharmacological reviews* 62.4 (2010), pp. 588–631.

[37] Bharath Ramsundar et al. *Deep Learning for the Life Sciences*. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837. O'Reilly Media, 2019.

[38] Roshan Rao et al. "Evaluating protein transfer learning with TAPE". In: *Advances in neural information processing systems* 32 (2019).

[39] Benjamin Sanchez-Lengeling et al. "Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC)". In: (2017).

[40] Michael Schneider and Oliver Brock. "Combining physicochemical and evolutionary information for protein contact prediction". In: *PloS one* 9.10 (2014), e108438.

[41] Adam Smith. "Screening for drug discovery: the leading question". In: *Nature* 418.6896 (2002), pp. 453–455.

[42] Jing Tang et al. "Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis". In: *Journal of Chemical Information and Modeling* 54.3 (2014), pp. 735–743.

[43] ScienceDaily University of Toronto. *A cell holds 42 million protein molecules, scientists reveal*. 2018. URL: https://www.sciencedaily.com/releases/2018/01/180117131202.htm.

[44] "UniProt: the universal protein knowledgebase in 2021". In: *Nucleic acids research* 49.D1 (2021), pp. D480–D489.

[45] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of chemical information and computer sciences* 28.1 (1988), pp. 31–36.

[46] David S Wishart et al. "DrugBank 5.0: a major update to the DrugBank database for 2018". In: *Nucleic acids research* 46.D1 (2018), pp. D1074–D1082.

[47] Lingwei Xie et al. "Deep learning-based transcriptome data classification for drug-target interaction prediction". In: *BMC genomics* 19.7 (2018), pp. 93–102.

[48] Jianyi Yang, Ambrish Roy, and Yang Zhang. "BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions". In: *Nucleic acids research* 41.D1 (2012), pp. D1096–D1103.

[49] Hong Zeng et al. "ComplexContact: a web server for inter-protein contact prediction using deep learning". In: *Nucleic acids research* 46.W1 (2018), W432–W437.

[50] Yang Zhang. "Template-based modeling and free modeling by I-TASSER in CASP7". In: *Proteins: Structure, Function, and Bioinformatics* 69.S8 (2007), pp. 108–117.