PROTEIN DRUG INTERACTION FROM THEIR SEQUENCE

Using the SMILES and SEQUENCE information

Anup Adhikari

Supervisor
Dr. Surendra Shrestha

This Thesis is carried out as a part of the education at the Tribhuwan University and is therefore approved as a part of this education. However, this does not imply that the University answers for the methods that are used or the conclusions that are drawn.

Tribhuwan University, 2019
Institute of Engineering
Pulchowk Campus
Department of Electronics and Computer Engineering

Abstract

Protein and Drugs are the major analysis subjects in computational bioinformatics to produce conclusions in treatment of diseases. While scientific methods are progressing with experiments and medical principles, they are still expensive means to discover the cure of new diseases. In principle, the high computing systems can be used to reduce the costs related to discovery. With the evolving nature of disease for instance, due to mutation, there has been extensive research work on exploring the fundamental properties of proteins and drugs to find the correct match in treatment. Finding the interaction between drugs and proteins based on their molecular fingerprints and protein sequences has been explored using statistical methods and rule-based methods. The representation of drugs in fingerprints and proteins in sequence are used to map them to different domains, which are then trained in a deep neural net to produce a regression solution. Instead of relying on binary classification, the more superior KIBA scores are used to quantify the interaction score between the drugs and proteins. The feature vectors, PSSMDT, Embedding and RPT, are combined to aid the deep learning state-of-art solution with convolution and dense layers, and aid to prediction ci-index score of 87%.

Acknowledgement

I would like to express the deepest appreciation to my supervisor and Head of Department of Electronics and Computer Engineering, Pulchowk Campus Dr. Surendra Shrestha for his guidance throughout the period of this work. His invaluable support, understanding and expertise have been very important in completing this work. It was a great honor for me to pursue my thesis under his supervision.

I pay my sincere gratitude to Dr. Aman Shakya, MSCSKE Coordinator for his supervision and help during this research work.

I am highly grateful to Prof. Dr. Shashidhar Ram Joshi, Prof. Dr. Subarna Shakya, Dr. Sanjeeb Prasad Pandey, Dr. Dibakar Raj Pant and Dr. Basanta Joshi for their encouragement and guidance.

I would like to express my heartily gratitude towards the Institute of Engineering, Pulchowk Campus along with all my respected teachers, my friends, my family for giving me continuous support for their invaluable help.

Anup Adhikari

073 MSCS 652

Institute of Engineering

Table of Contents

List of Figures

Page

List of Tables

Page

Chapter 1: Introduction

1.1 Background

Treatment of diseases are mostly associated with applying foreign medicinal components into human body. The rudimentary means of curing diseases has been growing with applied ayurbedas since the past. The chemical perspective of curing diseases slowly evolved into modern chemistry as drug facilities developed around the globe. The extensive research and documentation changed the world where people have come to trust fully in chemist's drugs to mitigate the ailments in the body. With the growing chemical intereset in the community, the need to develop better drugs and quick solutions increased even higher. With the identification of new diseases and dire need of understanding the mechanism to cure such diseases, the drug research started gaining its speed.

Computer-aided drug discovery (CADD) mechanism have been developing bioinformatics ever since the "Next Industrial Revolution" possibilities started grow [12, 4]. The interest started as Fortune magazine published the article "Designing Drugs by Computer at Merk". Experimenting with computational power and technical human resources in biomedicine, the concepts started to form scopes like High Throughput Screening (HTS) – A technique to screen desired drugs from other drugs. HTS was evolving eventually to find precedence over finding novel therapeutics. The desire to increase high hit rate did grow as the traditional HTS led to few probable leads. As research developed on computational drug design, CADD study broadened based on the computational resources required. CADD can be classified into two general categories: Structure-based CADD and Ligand-based CADD.

Structure based CADD relies on knowledge of structural analysis of protein structures in particular to identify the drug leads. It associates to phenomenons like Binding Site Analyses, Docking Simulations, and Scoring Algorithms. In brief, all the struc-

tural properties of proteins are exploited to identify the possible drug candidates – the molecules which fit in the protein structure description. This work borrows the representational feature sets of proteins and drugs from this discipline.

Ligand-based CADD exploits similarities of known active and inactive molecules. It further exploits the chemical, electrical and functional properties from drugs and proteins. This work borrows the feature representations of chemical-electrical properties in the form of Residue Residue Statistical Residual Vector (R2RSRV).

This work relates mostly to the Structure based CADD and partly to Ligand-based CADD. Target Structure and Ligand Structures are the major parameters of the research. The de-novo design has not been explored yet but the research method in this work can be used to test the drug designs for Structure Generator[1]. The other aspects of target identification – Molecular Dynamics, Pharmacophore modeling, Ligand Docking, Quantitative Structure Activity Relation (QSAR) etc. are beyond the scope of this work. So, the predictions from the model may not be sufficient to conclude the predicted interaction results. The pharmacophore models could take the results from this work and make decisive conclusions.

The Dataset contains scores of the interaction of proteins and drugs based on KIBA scores. We use 52498 drugs from CHEMBL and 254 proteins from UniProt to get their structural information. The interaction of 180244 is obtained from the research work produced by [17], and by removing the unrecognized interactions. The interactions is based on KIBA score – an integrated approach by combining the power of thermodynamic constants and activity percentage of drug-target interaction profile.

1.2 Statement of Problem

The simple technique of encoding the sequence information of drugs and proteins to identify if a drug will interact with the protein or not has a major issue in that while

[1]Structure Generator: The molecules which are highly active, readily synthesizable and devoid of undesirable properties are used to construct new possible drugs and can be tested with multiple targets.

drugs encoding information can be used to make drug related predictions, the protein encodings require additional feature vector input to properly form their representational vectors. For instance, the docking of drugs to protein structure doesn't only depend on surface area –a condition that structural representation can learn with proper algorithm –, but also with electric field and H-bond properties [20]. Therefore, modeling a machine learning algorithm sometimes overfit the situation or poorly classify the problem. In this work, we explore various features integration like R2RSRV and PSSM matrix along with sequence feature set and reproduce a regression problem for solving the prediction problem.

1.2.1 Selection of Prediction Score

Out of the many score functions; STITCH, Davis, Metz_Anastassiadis and KIBA scores, KIBA scores are used for the prediction of drug and protein interaction problem. The main reasons being: STITCH scores don't fully explore the primary thermodynamic dissociation constants used for drug-target interaction profile and other scores are used by KIBA [17]. Again, KIBA scores database consists of experimental data and secondary data (from literature) of drug-target interaction. Choosing the KIBA as the output score for two protein and drug sequences, we model our machine learning algorithm for prediction of interaction.

1.2.2 Selection of Features

For the protein family, the focus here is with the kinase target family because of its essential roles in cellular signaling transduction for many cancers and inflammatory diseases [17, 11]. We concentrate on proteins dataset, specifically because their interaction is quite tricky when considered among chemical, atomic, structural and electrical nature of protein residues [13]. Our basis for forming the matrices and vectors related to protein sequence comes from the fact that these features represent specific properties related to the protein and its residues. Also, the literatures describing the feature sets characteristics

and results motivates us towards the selection of these parameters: PSSM-DT, EDT, RPT and embedding vectors.

1.3 Objectives

The objectives of the research are:

- To determine the effective feature matrices related to protein.
- To determine the right machine learning algorithm for predicting the protein-drug interactions.

1.4 Organization of Report

1.4.1 Choosing Method of Interaction

Out of the two methods of contact prediction: Global Methods and Local Methods, where Global Method tries to predict the label of one residue pair considering the label of others while Local Method tries to predict the label of one residue pair without considering the label of others; we use Global Methods as a means of contact prediction. We try to run different variations in Residual Methods: Using Distance Prediction, Coevolutionary features, Sequence Representation.

1.4.2 Deep Learning Network Selection

Convnets, as they still are quite helpful in solving an image recognition problem, we used the stack of CNN with other keras layers to understand the performance of prediction of interaction with protein drug set. The image problem is in analogy as the different canonical dimension of drugs being mapped with canonical dimension of proteins. The value of pixel can be thought of as an interaction value of drug substituent with protein substituent.

1.4.3 Training and Testing

A basic PC was used to create initial models. Google Colabs was used to train the deep convolutionary stack due to requirement of GPU. The models were saved on the runtime so that the next training could be resumed immediately after the cease of Colab's VM Session.

Chapter 2: Theorical Background

2.1 No Free Lunch Algorithm (NFL)

The no free lunch theorem for search and optimization applies to finite spaces and algorithms that do not resample points. It states "All algorithms that search for an extrema of a cost function perform exactly the same when averaged over all possible cost functions." To increase the scope of NFL-like analyses, we need to make two slight extensions: first, we must broaden the definition of performance measures to allow for dependence on f-the list of multiple functions, and second, we need to generalize fitness functions to allow for nondeterminism. [19]

The search problem in case of proteins can be thought to comprise of different sample spaces: Primary Structures, Secondary Structures, Evolutionary Structures, Chemical Parameters, Atomic Parameters etc. This work only tries to explore the primary, secondary and evolutionary nature of protein-residues.

Generalized Optimization

The major implication of NFL is useful when the sample spaces are operated by different algorithms. The theorem being that all algorithms in different sample spaces produce the highest optimum results for a given problem. Generalized Optimization follows that when these different algorithms that are optimized in different sample spaces are included in one algorithm, then the method provides us the best predictions.

2.2 Literature Review

Finding the interaction between drugs and proteins based simply on their primary structure information is one of the many challenges faced in drug-synthesis process. The

experimental methods are quite expensive in terms of time, money and resources. Still the mutation in cells are growing higher due to extensive use of chemical and electromagnetic radiations in our environment. In one hand, diseases are getting powerful and in the other hand, the experimental method can take months when finding a right cure is considered. One of the solutions to this is use of high computing ends that can automate some of its repetitive works. Therefore, computational methods help to lessen the amount of works required to find right drug partner for the evolving diseases.

Protein molecules are the workhorses of our body. For example: the blood protein hemoglobin is functional for O_2 /CO_2 transportation, antibodies defend against viruses and hormonal protein insulin regulates our blood sugar level. The protein has differences in structure, according to the desirable functional characteristics of our body. This structure is so important for our health, that understanding them can aid to cure diseases. For example, diseases like Parkinson's is unrelated to bacteria/virus but due to incorrect folding of proteins.

Our bodily functions are dependent on protein structure and their interdependent interactions play a vital role. Some of these proteins are of critical interest to biochemistry and biomedicine researchers.[2] For example, a protein known as amyloid beta, which forms plaques in the human brain, is a key to understanding Alzheimer's disease. Improving our understanding of correct protein structures can lead to the design of drug treatments that can target deactivation of proteins of interest. Also, the personalized treatment of any sick person by taking sample of protein structure may help design cure for specific cases (eg. due to mutation changes of protein structure), which otherwise is referred in for general case of differently related protein [7].Thus it will solve issues of wrong medication hazards, which are the general scenario for the developing and under-developed countries.

The rise of new machine learning methods and deep-learning techniques are closing the gaps to create better predictions. The cure of evolving diseases can be computationally researched by use of knowledge-based community. The community contributed databases in drugs and protein sectors are growing at the same pace. Clearly, both the data

resources and algorithmic techniques can help human community to counter-act against such circumstances.

In the field of bioinformatics, the long-standing problem of computationally predicting the structure of a protein remains unsolved[6]. The key to solving this problem is to accurately predict 'contacts', which requires measuring the physical distances between the amino acids of a folded protein. The current state-of-the-art methods like ProC_S3 and SVMcon are about 50% accurate [1].

Deep learning, which is a subfield of machine learning, has recently enabled accurate face recognition in Facebook, Google Photos etc. Google's self driving car already uses automatic driving [3]. It has also helped to accurately detect skin cancer. These demonstrated successes of deep learning algorithms clearly highlight its potential to greatly accelerate scientific problems such as protein contact prediction.

In the other hand chemical properties of drugs and the targets complicate the situation as they react differently with slight change in protein sequence. While computational techniques have helped to simulate the different conditions, the fundamental dataset is still long way to go. The reason being that the identification of proteins structure can take months. Again the isomeric states of proteins' structures can have different functional aspects to the body physiology. Moreover, the complexes tend to behave similarly even when the protein sequences are distantly related, one of the results of tertiary structures that the proteins are form of. [5]

The deep learning methods are quite good at predicting the molecular behaviour of the drug. However they present no good means when predicting the behaviour of proteins. This can be thought as protein-folding problem which when solved will help escalate the development of treatment facilites around the globe. The Critical Assessment of Structure Prediction (CASP) experiments is such community which holds competition to determine and advance the state of art in modeling the protein sequence from amino acids.[9] To find the computational measure to predict the drug for a given protein, the major fallback is that the simple encoding techniques don't incorporate the proteins behaviour related to

hydrophobicity, acidity, secondary and tertiary structures information.[20]

No Free Lunch Theorem [19] can be used for on the other hand works by basing the prediction guesses based on a number of prediction functions. Here, we use the sequence information of proteins to calculate the predictions on different feature transformation techniques and generalize those predictions using a stack of dense layers.

Chapter 3: Methodology

3.1 System Overview

3.1.1 System Block

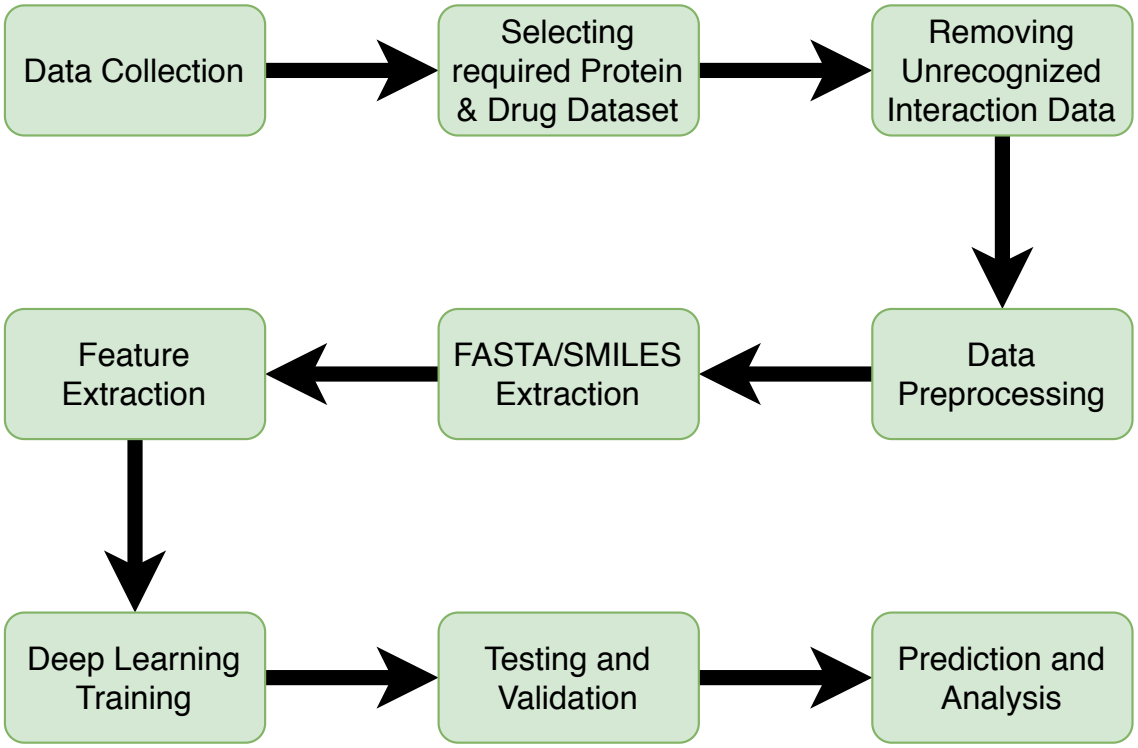


Figure 3.1: System Block Diagram for Protein-Drug Prediction

Figure 3.1 shows the different stages of research for building a protein drug prediction system. The data is collected by exploring the available internet sources and required dataset is downloaded for processing. From the data, the missing values are removed in data preprocessing to aid proper training. From the raw drugs and proteins profile, SMILES and FASTA sequences are extracted respectively. Now the features are extracted based on the sequence provided. Then, the features are fed into Deep Learning Algorithm where the training is performed to create the right prediction system. The training follows by testing and validation of data under different settings of protein and

drug combinations.

3.1.2 Data Collection

The dataset is collected from open-internet database. Basically, there are three types of data required in this work: protein, drug and interaction sets. The UniProt Library has been used for extracting proteins features, PubChem for drug features, and NCBI for interaction scores. Additionally, PSI-BLAST is used to generate PSSM matrices for the protein features downloaded.

UniProt contains database of 173,281 proteins of human (Homo sapiens) (until 2019). The protein document consists of the taxonomic classification, identifiers to other databases for cross-linking, molecular properties, related specific bioactivity, functional property, canonical and isoforms of protein sequence. The protein fasta sequence in particular is of interest to this research. An API can be used to download the available information. `https://www.uniprot.org/help/programmatic_access`

>O00311

MEASLGIQMDEPMAFSPQRDRFQAEGSLKKNEQNFKLAGVKKDIEKLY
EAVPQLSNVFKIEDKIGEGTFSSVYLATAQLQVGPEEKIALKHLIPTSHPIRIAAEL
QCLTVAGGQDNVMGVKYCFRKNDHVVIAMPYLEHESFLDILNSLSFQEVREYM
LNLFKALKRIHQFGIVHRDVKPSNFLYNRRLKKYALVDFGLAQGTHDTKIELLK
FVQSEAQQERCSQNKSHIITGNKIPLSGPVPKELDQQSTTKASVKRPYTNAQIQIK
QGKDGKEGSVGLSVQRSVFGERNFNIHSSISHESPAVKLMKQSKTVDVLSRKLA
TKKKAISTKVMNSAVMRKTASSCPASLTCDCYATDKVCSICLSRRQQVAPRAG
TPGFRAPEVLTKCPNQTTAIDMWSAGVIFLSLLSGRYPFYKASDDLTALAQIMTI
RGSRETIQAAKTFGKSILCSKEVPAQDLRKLCERLRGMDSSTPKLTSDIQGHASH
QPAISEKTDHKASCLVQTPPGQYSGNSFKKGDSNSCEHCFDEYNTNLEGWNEVP
DEAYDLLDKLLDLNPASRITAEEALLHPFFKDMSL

PubCHEM and CHEMBL are drug databases used for feature extraction of drug molecules. PubCHEM is a database containing 96,881,514 drug compounds and asso-

ciates to each using CID identifier. It allows programmatic access and downloads of database text files. The SMILES structure provided by the PubChem library is used to generate features corresponding to each drug molecule. The properties associated with the molecule is explored using CHEMBL database using a programmatic api request provided.

`https://pubchemdocs.ncbi.nlm.nih.gov/programmatic-access`, `https://chembl.gitbook.io/chembl-interface-documentation/web-services`.

CHEMBL379218

PubCHEM CID 11314340

CC1=C2C=C(C=CC2=NN1)C3=CC(=CN=C3)OCC(CC4=CC=CC=C4)N

For the drug-target interaction (i.e. drug-protein interaction), KIBA scores were used [17] instead of binary classification. Thus, a regression model was used to predict the drug and protein interaction. The KIBA score regression has two major advantages over binary classification: interaction strength of similarly interacting ligands-target (drugs-protein) can be compared and the bias problem of unknown interactions is refrained [17, 15]. Higher score means that there is more strength of interaction between the two. We use 52498 ligands as drugs and 254 human proteins as target for the prediction problem.

Table 3.1: KIBA Score Table

	O00238	O00311	O00329	O00418
CHEMBL10	3.518514	3.100002	4.0	3.6
CHEMBL102000	NaN	NaN	NaN	NaN

Various components were used to form the prediction system. Protein interaction depends on its structural, chemical, molecular(related to H-bond) and electrostatic properties. The structural representation form basis for creating features in other properties. The primary canonical structure of protein-drug set are fed into interaction block. The interaction parameter is filtered accordingly to the filter type. Similarly, the drug feature set are created to be trained with the machine learning algorithm. Finally, after training the training dataset, cross validation of the model was done.

3.2 Building Components of Features Processing

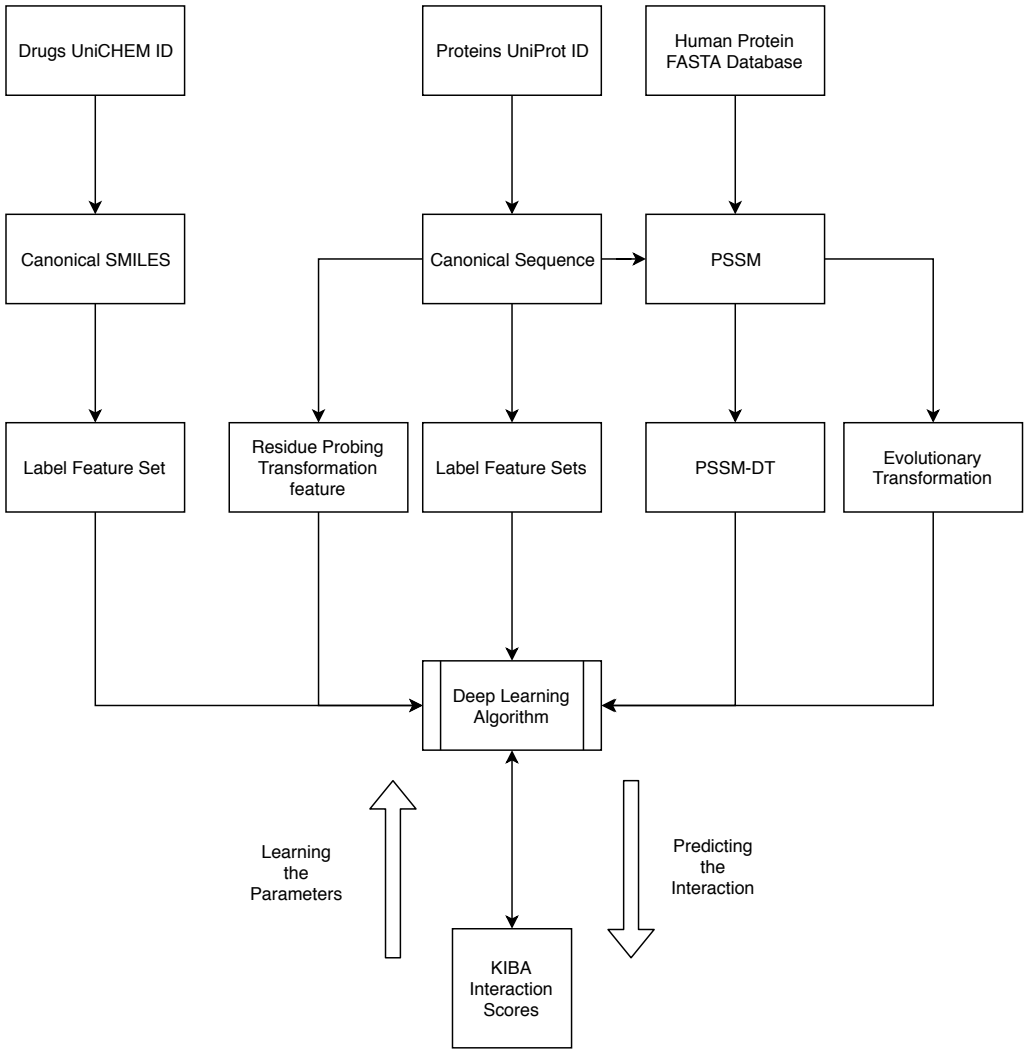


Figure 3.2: Schematic Block Diagram for Protein-Drug Prediction

Figure 3.2 shows the building components of the Input Vectors for feeding the deep learning network. The KIBA Prediction id done by feeding canonical protein fasta and drug smiles. The other features are constructed on their basis. PSSM matrix is constructed using PSSM matrices of human genome protein library from UniProt [18] and the protein's FASTA. Two features, Evolutionary Transformation (ET) and Position Specific Scoring Matrix Distance Transformation (PSSM-DT) are extracted from PSSM. From FASTA, Labelled encodings and Residue Probing Transformation (RPT) matrix are created. From the SMILES, only labelled encodings is extracted.

3.3 Dataset Description

3.3.1 Kinase Inhibitor Bioactivity (KIBA)

The Kinase Inhibitor Bioactivity (KIBA) Scores are collected from the publicly made available dataset [17]. The scores are based on thermodynamic constants K_i and K_d and, remaining enzyme activity(Activity % – IC_{50}).

$$KIBA = \begin{cases} K_i.adj & if\ IC_{50}\ and\ K_i\ are\ present \\ K_b.adj & if IC_{50}\ and\ K_d\ are\ present \\ \frac{K_i.adj\ K_b.adj}{2} & if\ IC_{50}\ ,K_i\ and\ K_d\ are\ present \end{cases} \tag{3.1}$$

where,

$$K_i.adj = \frac{IC_{50}}{1 + L_i(IC_{50}/K_i)} \tag{3.2}$$

$$K_d.adj = \frac{IC_{50}}{1 + L_d(IC_{50}/K_d)} \tag{3.3}$$

where L_d and L_i are parameters defining weights of IC_{50} in model adjustments for K_i and K_b

For a kinase inhibitor drug−target interaction, we consider the medians of three major bioactivity types IC_{50}, K_i, K_d where IC_{50} [17] is the concentration at which the inhibitor causes a 50% inhibition of enzymatic activity and K_i is defined by

$$Ki = \frac{IC_{50}}{1 + [S]K_m} \tag{3.4}$$

where, [S] is the experimental substrate concentration and K_m is the concentration of the substrate.

All the bioactivity types are available from CHEMBL[8]. Based on interaction data available, we remove the unknown values and obtained a total of 180244 interaction

KIBA score values in the range of -3.09 to 17.8. With the standard deviation of 1.22, it represents a total of 254 proteins and 52498 drugs.

3.3.2 Position Specific Score Matrix

Position Specific Scoring Matrix (PSSM) is a very useful protein feature. The protein feature represented by PSSM depends on the sequence of all the proteins in consideration. The HUMAN genome protein (a database of more than 100,000) is downloaded from UniProt Library. The PSSM matrix is constructed for each of the kinase proteins based on this HUMAN Genome Protein Database. With this, the PSSM matrix is characterized according to human proteins to anticipate the prediction of new identified kinase proteins.

Table 3.2 shows a rudimentary process of calculating PSSM score values. The sequence following shows the process of calculating the scores once the PSSM distribution of the whole family is calculated. Table 3.2d shows the score distribution of lowercase amino acid sequence (starting after 4th position) determined by the size of the sliding window.

ACTCagccccagcGGAGGTGAAGGACGTCCTTCCCCAGGAGCCGGTGAGA
AGCGCAGTCGGGGGCACGGGGATGAGCTCAGGGGCCTCTAGAAAGATGTAG
CTGGGACCTCGGGAAGCCCTGGCCTCCAGGTAGTCTCAGGAGAGCTACTCA
GGGTCGGGCTTGGGGAGAGGAGGAGCGGGGGTGAGGCCAGCAGCA

.3, .1, .1, 0, 0, .2, .7, .5, .2 == Sum(2.1) - posix(4) – See Table 3.2e

3.3.3 PSI-BLAST

PSI-Blast tools relates with multiple sequence alignments from a family of protein sequences[16]. This helps to create a PSSM - Equation (3.5) - matrix referred to as secondary protein structure. For this study, the PSSM profile of every protein sequence is obtained by executing iteration of PSI-BLAST against [16, KEGG] protein. PSSM pro-

Table 3.2: PSSM Analysis Design

1	GAGGTAAAC
2	TCCGTAAGT
3	CAGGTTGGA
4	ACAGTCAGT
5	TAGGTCATT
6	TAGGTACTG
7	ATGGTAACT
8	CAGGTATAC
9	TGTGTGAGT
10	AAGGTAAGT

(a) Protein FASTA Sequence

	1	2	3	4	5	6	7	8	9
A	3	6	1	0	0	6	7	2	1
C	2	2	1	0	0	2	1	1	2
G	1	1	7	10	0	1	1	5	1
T	4	1	1	0	10	1	1	2	6

(b) Frequency Table

	1	2	3	4	5	6	7	8	9
A	0.3	0.6	0.1	0.00	0.00	0.6	0.7	0.2	0.1
C	0.2	0.2	0.1	0.00	0.00	0.2	0.1	0.1	0.2
G	0.1	0.1	0.7	1.00	0.00	0.1	0.1	0.5	0.1
T	0.4	0.1	0.1	0.00	1.00	0.1	0.1	0.2	0.6

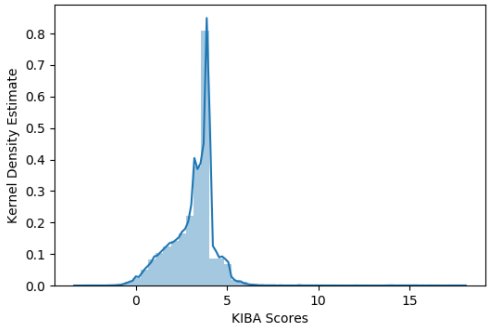
(c) Log-Likelihood Matrix

	1	2	3	4	5	6	7	8	9
A	0.3						0.7		
C			0.1	0.00	0.00	0.20			0.2
G		0.1						0.5	
T									

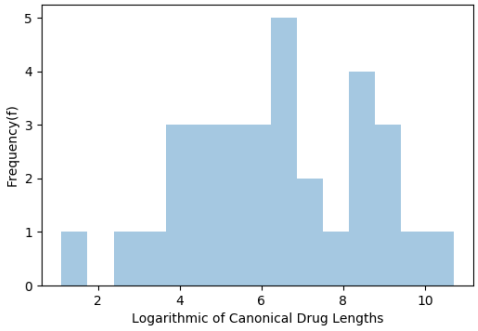
(d) Sliding Window Score Calculation

0	1	2	3	4	5	6	7	8	9
1.099	1	2.2	2.1	2.1	1.300	1.3	1.4	2	2.9

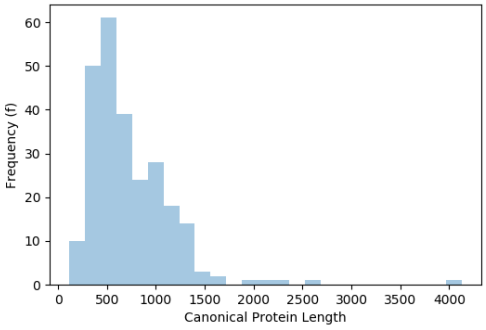
(e) Score of sliding window motifs



(a) KIBA Scores



(b) Logarithmic Label Encodings of Drug Sequence



(c) Logarithmic Labeled Encodings of Protein Sequence

Figure 3.3: Data Distribution of KIBA-interaction scores, Drug Sequences and Protein Sequences

file is a matrix of L*20 dimensions whereby 20 is the standard type of amino acids and L being the length of the protein. The larger positive scores represent conserved positions, which in turn implies critical functional residues that are required to perform various intermolecular interactions.[16, PSSM]

$$PSSM = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,20} \\ P_{2,1} & P_{2,2} & \dots & P_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ P_{L,1} & P_{L,2} & \dots & P_{L,20} \end{bmatrix} \tag{3.5}$$

PSSM-DT

Two forms of PSSM distance transformation techniques are used to transform the PSSM information into fixed dimensional vectors [21]. The PSSM-DT (PSSM-Distance Transformation) can transform the PSSM information into uniform numeric representation by approximately measuring the occurrence probabilities of any pairs of amino acid. It results in two types of feature matrices: PSSM-SDT and PSSM-DDT defined by:

$$[H]PSSM - SDT(i, lg) = \sum_{i=1}^{L-lg} S_{i,j} \times \frac{S_{i,j+lg}}{L - lg} \tag{3.6}$$

lg = distance of separation between same amino acid sequence

$$[H]PSSM - DDT(i_1, i_2, lg) = \sum_{j=1}^{L-lg} S_{i_1,j} \times \frac{S_{i_2,j+lg}}{L - lg} \tag{3.7}$$

i_1 and i_2 refer to tow different types of amino acids

Thus we have [380 (3.7)+20 (3.6) = 400] x lg matrix which will be used as protein-specific vector in this work.

Evolutionary Distance Transformation Matrix

The mutational information of protein can be more informative than the sequence information itself[22]. Evolutionary difference formula (EDF) is used to represent mutation difference between adjacent residues. Secondly, the PSSM is converted into 20 x 20 matrix (ED-PSSM). These extracts are the non co-occurrence probability for two amino acids separated by a certain distance d in the protein from the PSSM profile. For example, d=1 implies that the two amino acids are consecutive; d=2 implies that there is one amino acid between the two. Next, the EDT feature vector computed from ED-PSSM can be represented as (3.8):

$$P = [\partial_1, \partial_2, \ldots, \partial_\Omega] \tag{3.8}$$

where Ω is an integer that represents the dimension of the vector whose value is 400.. The non-co-occurrence probability of two amino acids separated by distance d can be computed as:

$$f(A_x, A_y) = \sum_{d=1}^{D} \frac{1}{L-d} \sum_{i=1}^{L-d} (P_{i,x} - P_{i+d,y})^2 \tag{3.9}$$

where $P_{i,x}$ and $P_{i+d,y}$ are the elements in the PSSM profile; A_x and A_y represent any of the the 20 different amino acids in the protein sequence. Finally we spread the $f(A_x, A_y)$ in equation 3.8 as: $\partial_1 = f(A_1, A_2)$, $\partial_{400} = f(A_{20}, A_{20})$

3.3.4 Residue feature

The Statistical Residue Vector Space R2RSRV [20] plays an important role in Residue Residue Interaction and creates a basis for structural stability of the protein sequence itself. It is related to the tertiary structure of the protein sequence. Nonetheless, another function is to create a correlated sequence of information whereby two proteins are distantly related by sequence. Simultaneously, it is highly related to the functional characteristic of protein. With this, Table 1 as attached in Appendix depicts a 20 x 20 matrix whose rows and columns represent 20 standard amino acids.

Residue Probing Transformation(RPT) feature

RPT as proposed by Jeong et al.[10], and implemented by Pujan et al.[14], emphasize domains with similar conservation rates by grouping domain families based on their conservation score in the PSSM profile.

$$RPT = \begin{bmatrix} S_{1,1} & S_{1,2} & \dots & S_{1,20} \\ S_{2,1} & S_{2,2} & \dots & S_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ S_{2,1} & S_{2,2} & \dots & S_{2,20} \end{bmatrix} \tag{3.10}$$

The RPT matrix (Equation 3.10) is then tranformed into feature vector of 400 dimensions, as shown in Equation 3.11.

$$V = [f_{s_{1,1}}, f_{s_{1,2}}, \dots, f_{s_{i,j}}, \dots, f_{s_{20,20}}] \tag{3.11}$$

where,

$$f_{s_{i,j}} = \frac{s_{i,j}}{L} (i, j = 1, 2, \dots, 20) \tag{3.12}$$

3.3.5 Labelled Encodings

The labeled encoding techniques is used to represent the canonical structure of drugs and proteins. The structural canonical information is preserved while sending the feature set to deep learning method. An array of integers are formed from particular sequence while representing the structural information.

The Labelled Encodings of protein and drugs can be defined by Table 3.3 :

Table 3.3: Labeled Encoding of Proteins and Drugs

A –> 1	C –> 2	B –> 3	E –> 4
D –> 5	G –> 6	F –> 7	I –> 8

(a) Label Encodings for Proteins

# –> 1	% –> 2	: –> 3	+ –> 5
4 –> 13	7 –> 14	F –> 25	I –> 26

(b) Label Encodings for Drugs

3.4 Deep Learning Model

The Features formed from data processing block are then subjected to deep learning model. The implementation is done using using keras library in python. The implemented model is represented by Figure 3.4. The input layers are described in Table 3.4.

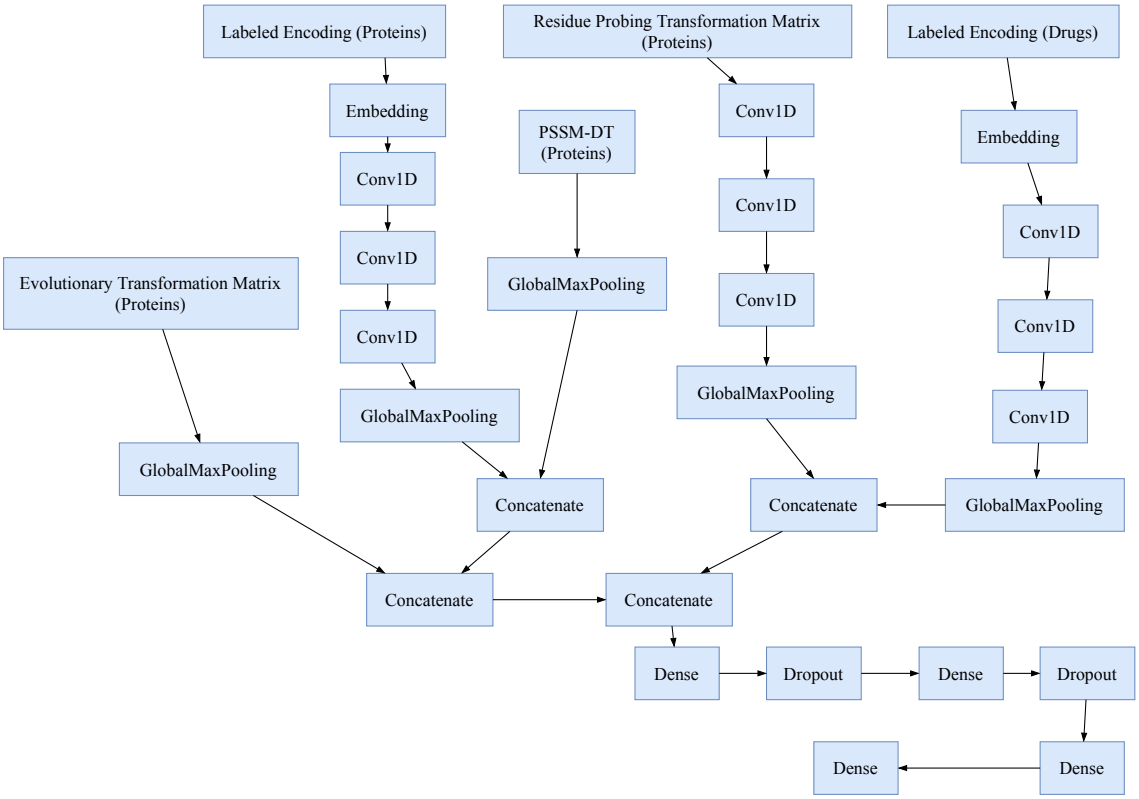


Figure 3.4: Deep Learning Model to predict Protein-Drug Interaction

Table 3.4: Inputs Used in the Deep Learning Network

S.No.	Input Layer Name	Used Feature Vector	Type
1	input_1	Label Encodings	Drug
2	input_2	Label Encodings	Protein
3	input_3	Evolutionary Distance Transformation Vector	Protein
4	input_4	PSSM-DT Vector	Protein
5	input_5	Residue Probing Transformation Vector	Protein

3.4.1 Components description used from Tensorflow (Keras)

Embedding Layer

From figure 3.4, the Embedding feature provided by keras for vector representation of both drug fingerprint and protein sequence are utilized. The one-hot encodings of the drugs and protein sequences are inputs to this layer. It turns positive integers (indexes) into dense vectors of fixed size. eg. [[4], [20]] -> [[0.25, 0.1], [0.6, -0.2]].

Convolution Neural Network

To learn the local patterns in the input vector, we use CNN. While Dense Layers can learn the global parameters, CNN is used to understand the local patterns. It does so by increasing the depth layer, which in turn is designed to learn different patterns as shown in Figure 3.5.

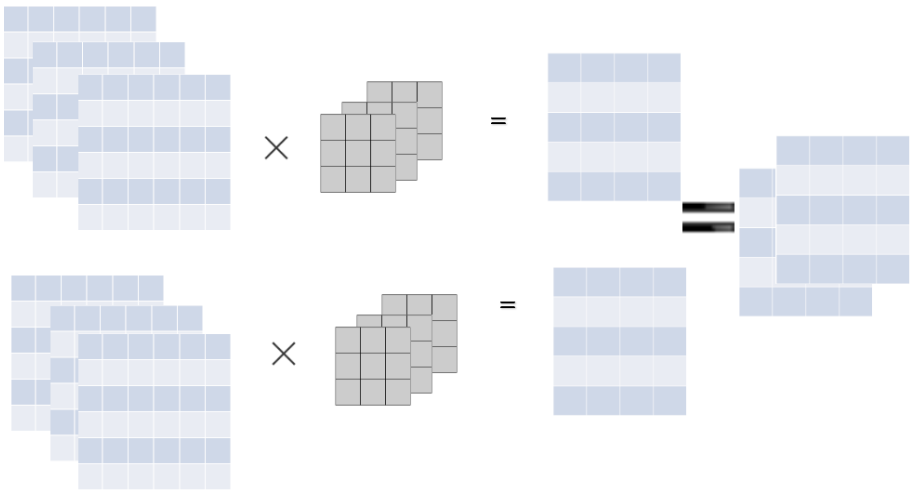


Figure 3.5: Convolutional Neural Network

Dense Layer

Dense Layer is a neural layer which fully connects the input layer to output layer. It can be used to learn the global pattern of the feature data. The representation can be seen from Figure 3.6.

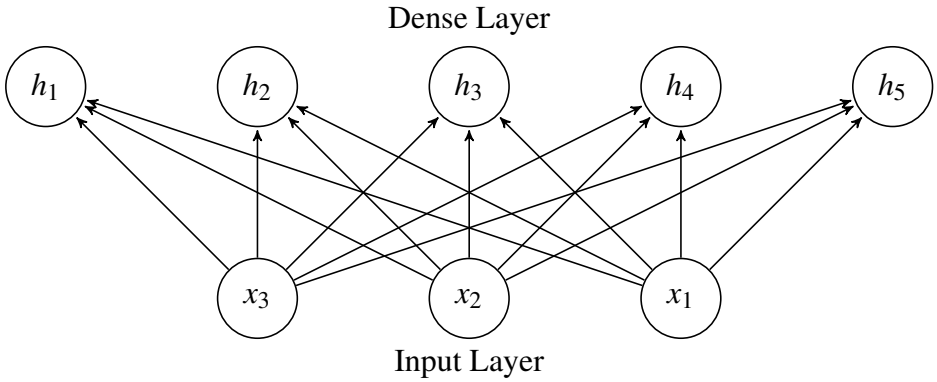


Figure 3.6: The input nodes are represented by x and the weights h are applied to each values of input nodes to generate an output

Dropout Layer

Our model becomes undesirable when every component of the input layer makes a significant change to the output layer. To reduce the effect of unimportant features the dropout layer was used. Thus the backpropagation network tries to ignore the noise features and minimizes the unrealizable prediction of the learning problem. This can be expressed diagrammatically in Figure 3.7.

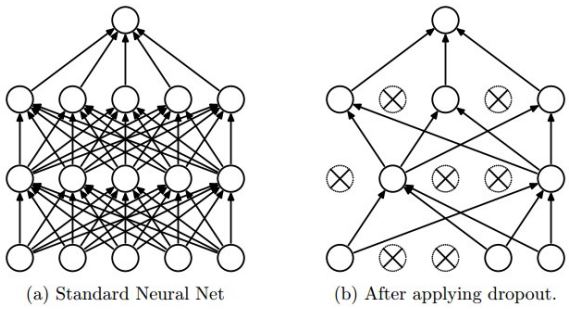


Figure 3.7: a) Standard neural network whose all the nodes have weights connected to higher nodes and lower nodes. b) Certain nodes belonging to same levels are disconnected. Some weights are also disconnected from other nodes depending on the percentage of dropout applied.

Pooling Layer

The Pooling layer is used to downsample the learned parameters from the grid of 3 dimensions returned by Convolution Layer. It gets reduced to 1 dimension by taking the highest values from the window size(corresponding to shape of 1st dimensional element).

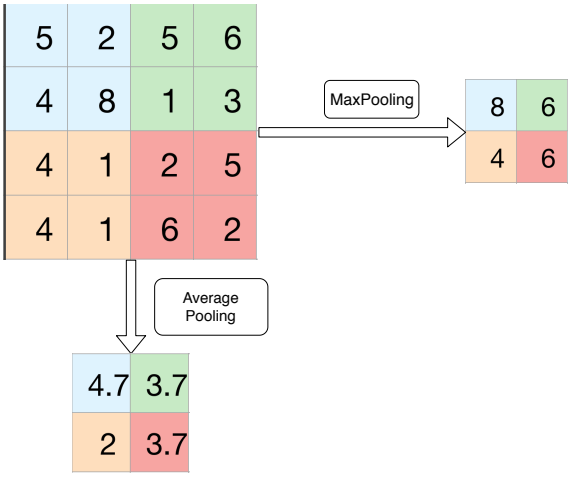


Figure 3.8: Pooling Layer

Concatenation Layer

Concatenation Layer as the name implies is used to simply join two vectors so that a feature set comprising of multiple features can be created. Their positional index indicates the feature set being manipulated.

Chapter 4: Experiments and Results

4.1 Experiments

The focus of the experiments are concentrated on the properties of protein as they have complex structures. The binding of protein and drug depend on various attributes of protein like acidity, hydrophobicity, binding pockets etc and the structure of drug. The attributes are quite closely related to primary and secondary structure of protein themselves. Therefore, our model aims to relate all these multiple components with matrix representation and confirming to Figure 3.3 prediction.

4.1.1 Features Selection

Primary Feature Selection

The sequence information of drugs and proteins live in their canonical form. Therefore we relied on Neural Net Sequence Embedding technique to form the primary representation. Both protein and drug were converted to Embedding vectors after creation of their labelled encodings.

Secondary Features Selection

These are the structural components of protein especially related to alpha and beta strands of Protein segments. All the protein Sequences are subjected to Equation (3.5) from the labelled encodings. The PSSM matrix is calculated using PSI-BLAST[16]. Then all the testing protein sets are evaluated with the resultant PSSM to form a new PSSM matrix specific to the testing protein. Thus, we expect to explore how proteins relate with the interaction experiments with the protein domain. From the PSSM, we evaluate the other evolutionary and distance vectors using equations 3.10, 3.7, 3.6 and 3.9.

4.1.2 Implementation

Stacked Features

Basically, we implement the architecture in Figure 3.4 for our model design. It is implemented in Python using the TensorFlow framework consisting of keras. The training contained of 200 epochs. The learning rates and early stopping were manipulated in the training process by using keras callback functions. The training and testing was done using a 5-fold cross-validation set prepared manually. To evaluate the performance of the model, we used concordance index(CI)[21] as defined by equation 4.1:

$$CI = \frac{1}{Z} \sum_{\delta_i > \delta_j} h(b_i - b_j) \tag{4.1}$$

where b_i is the prediction value for higher affinity δ_i and b_j is the prediction value for smallery affinity δ_j, Z is the normalization constant and $h(m)$ is the unit step function:

$$h(x) = \begin{cases} 1, & if x > 0 \\ 0.5, & if x = 0 \\ 0, & if x < 0 \end{cases} \tag{4.2}$$

4.2 Results

The deep learning method was implemented with various filter sizes of Convolutional Layer of Protein and Drugs. Only the comparable results have been shown in Table 4.1. The training and validation plots under different settings is clearly shown by Figure 4.1. The different settings are:

- Setting 1 (S1): The validation proteins and drugs appear in training set.
- Setting 2 (S2): The validation Protein is seen in training set but validation drugs aren't present while training.

- Setting 3 (S3): The validation Protein is absent in training set but validation drugs are present in training set.

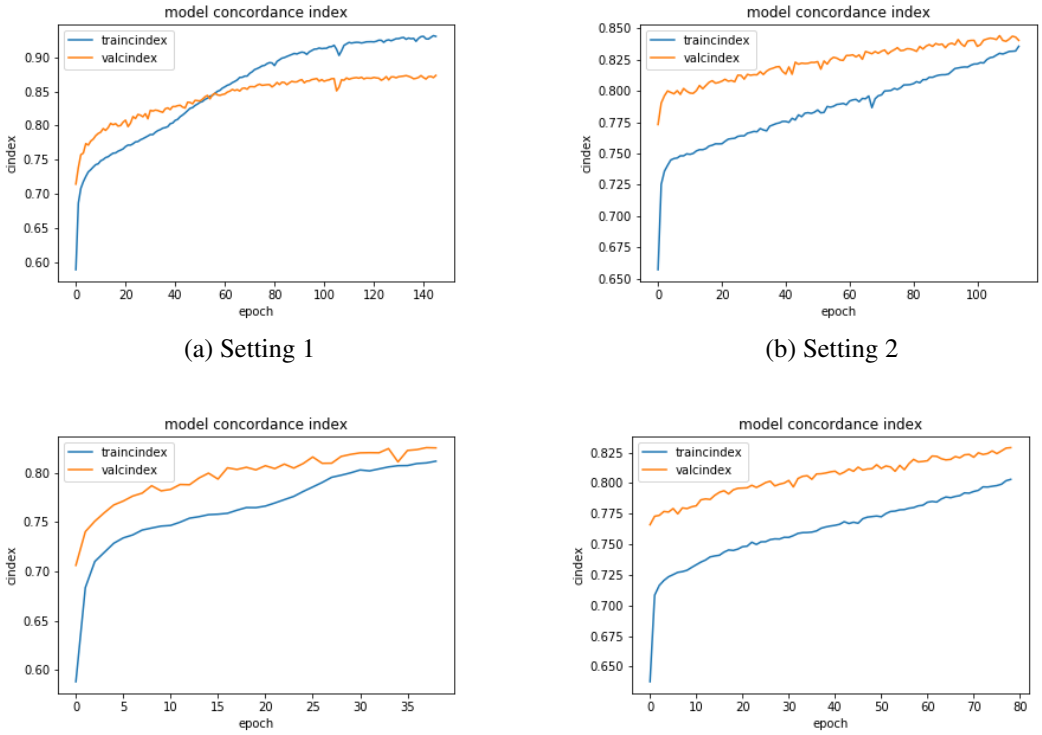
- Setting 4 (S4): The validation Protein and Drugs don't appear in training set

Table 4.1: Experiments results under different settings (S1, S2, S3, S4)

S.No.	Drug Smiles Window Size	FASTA Window Size	Setting	CI-val	MSE	CI-train
1	4	8	1	0.813936	0.811580	0.893064
2	4	12	1	0.810844	0.809919	1.285271
3	8	8	1	0.873094	0.873648	0.177053
4	8	12	1	0.807230	0.806509	1.056755
1	4	8	2	0.788051	0.789266	0.678707
2	4	12	2	0.802355	0.802892	0.360900
3	8	8	2	0.803720	0.806808	0.792821
4	8	12	2	0.803853	0.805306	0.672790
1	4	8	3	0.815181	0.815076	1.562749
2	4	12	3	0.805739	0.806455	1.541746
3	8	8	3	0.813063	0.813148	1.259206
4	8	12	3	0.824767	0.825603	1.396765
1	4	8	4	0.803982	0.804584	0.412271
2	4	12	4	0.805961	0.807584	1.510426
3	8	8	4	0.815464	0.816097	1.614631
4	8	12	4	0.787073	0.787795	0.485298

4.3 Analysis

The Table 4.1 shows that the optimal filter window size under all settings can be chosen equal to 8 for both drugs and proteins. The highest C-Index Score of (87.30%) was obtained under Setting 1 when both the drugs and proteins were present in the validation set and training set. This can also be better realized visually from Figure 4.1.



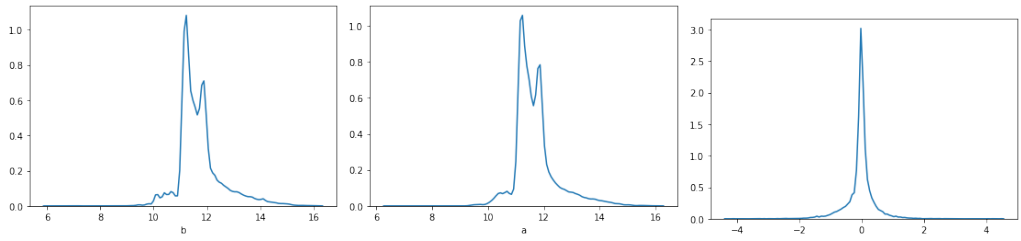
(a) Setting 1 (b) Setting 2

(c) Setting 3 (d) Setting 4

Figure 4.1: Plot of Training-Validation C-Index Scores over various Settings

4.4 Conclusion

The training of the model proved a better choice as we added the new feature dimension for proteins to get a generalization of all the feature sets extracted from the drug and protein sequence.



(a) Training Actual Scores (b) Training Predicted Scores (c) Training Scores Difference

Figure 4.2: Training Results based on KIBA Score Prediction

Similarly, the prediction scores and actual scores for validation sets is shown in Figure 4.4

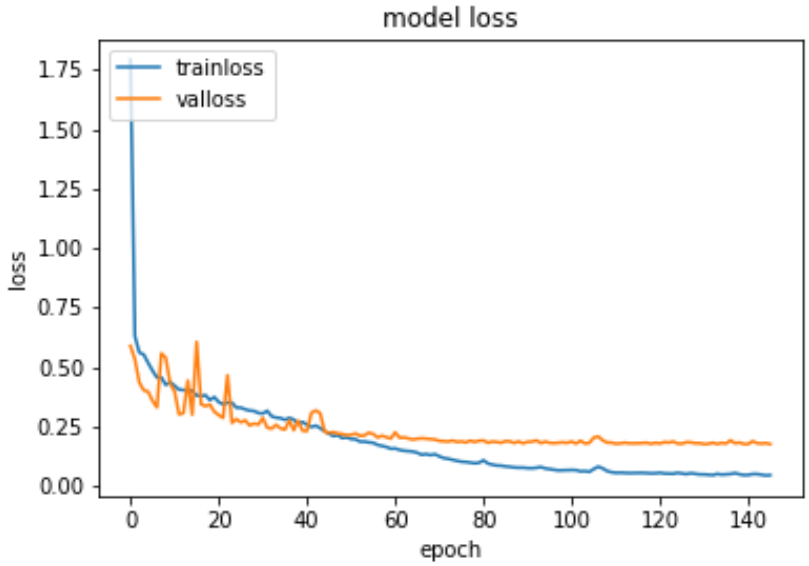
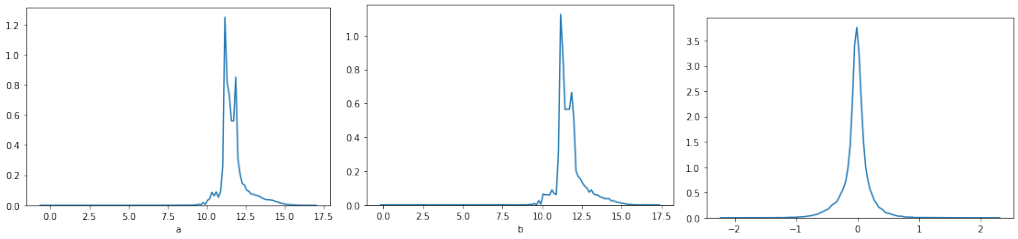


Figure 4.3: Training and Validation Loss Plot of Setting 1 (S1) with Filter Size = 8



(a) Testing Actual Scores (b) Testing Predicted Scores (c) Testing Scores Difference

Figure 4.4: Testing Results on KIBA score Prediction

Chapter 5: Conclusion

5.1 Limitations

This work only relates with the molecular properties of drugs and proteins to determine the best targets. The limitations to the deep learning method of protein-drug prediction in this manner are:

- Pharmacophore model of drugs and proteins have not been addressed. This is due to inavailability of proper datasets and highly complex nature of protein mechanism that the work become out of scope.

5.2 Future Works

We have only approached the sample space of No Free Lunch Algorithm (NFL) by forming the different feature-sets. However, the generalized optimization has not been achieved yet. The grid-search CV and Stacking Generalization can be used to create a better optimized machine learning method to solve the problem of protein-drug interaction from sequence information.

Appendix A: R2RSRV

Table 1: R2RSRV Matrix

	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	5.21	2.42	0.88	1.71	-1.59	1.13	0.95	0.48	-1.05	-3.20	0.65	1.44	-0.82	-1.54	-0.94	-0.62	-1.66	-3.14	-2.23	-2.14
V	2.42	9.46	1.33	0.49	-0.32	0.54	1.55	-2.12	-0.91	-1.80	-2.88	-1.05	-0.81	-1.32	-0.29	-0.58	-2.39	-3.69	0.66	-1.42
L	0.88	1.33	9.90	1.08	-0.42	2.17	2.41	-2.29	-3.40	-2.32	0.48	-0.77	-2.28	1.67	-0.77	-0.08	-3.49	-2.16	-2.10	0.19
F	1.71	0.49	1.05	6.11	0.55	0.89	0.52	-2.00	-1.10	-2.09	-0.11	1.14	0.83	-1.33	-1.79	0.42	-3.62	-0.96	-1.71	-1.33
C	-1.59	-32	-0.42	0.55	15.35	-1.35	-0.21	0.59	-1.52	1.53	-1.07	-1.16	0.28	0.95	-0.52	-1.47	-1.95	-2.23	-1.80	-0.84
M	1.13	0.54	2.17	0.89	-1.35	5.40	-0.28	0.44	-2.15	-1.50	-0.71	-0.33	-0.31	0.19	0.01	0.27	-3.38	-1.74	-0.72	-1.51
A	0.95	1.55	2.41	0.52	-0.21	-0.28	7.08	-2.04	-1.04	-0.61	-1.15	-1.22	-1.58	0.11	-0.53	-0.82	-1.06	0.17	-1.11	-2.74
G	0.48	-2.12	-2.29	-2.00	0.59	0.44	-2.04	5.65	1.67	-1.32	-0.82	0.27	-0.60	0.75	-2.24	1.68	0.70	-1.01	1.72	1.22
T	-1.05	-0.91	-3.40	-1.10	-1.52	-2.15	-1.04	1.67	4.42	1.23	0.59	-1.36	-0.04	-1.48	-0.06	-2.61	4.66	0.02	0.29	-0.74
S	-3.20	-1.80	-2.32	-2.09	1.53	-1.50	-0.61	-1.32	1.23	6.22	-1.10	-1.40	-0.79	-2.66	2.14	-0.08	4.57	0.95	0.11	-0.38
W	0.65	-2.88	0.48	-0.11	-1.07	-0.71	-1.15	-0.82	0.59	-1.10	1.08	-0.45	5.88	0.15	-2.84	-2.84	-1.98	-1.35	-0.27	4.08
Y	1.44	-1.05	-0.77	1.14	-1.16	-0.33	-1.22	0.27	-1.36	-1.40	-0.45	6.40	0.21	1.11	0.75	-2.73	-3.07	-0.45	0.87	-0.33
P	-0.82	-0.81	-2.28	0.83	0.28	-0.31	-1.58	-0.60	-0.04	-0.79	5.88	0.21	1.73	-1.13	0.66	0.82	-2.51	1.37	0.14	-0.40
H	-1.54	-1.32	1.67	-1.33	0.95	0.19	0.11	0.75	-1.48	-2.66	0.15	1.11	-1.13	5.03	-2.22	0.32	3.11	-1.46	-1.90	-0.06
E	-0.94	-0.29	-0.77	-1.79	-0.52	0.01	-0.53	-2.24	-0.06	2.14	-2.84	0.75	0.66	-2.22	2.59	-1.98	-4.29	0.07	3.52	3.45
Q	-0.62	-0.58	-0.08	0.42	-1.47	0.27	-0.82	1.68	-2.61	-0.08	-2.84	-2.73	0.82	0.32	-1.98	3.44	0.79	0.92	-0.67	0.24
D	-1.66	-2.39	-3.49	-3.62	-1.95	-3.38	-1.06	0.70	4.66	4.57	-1.98	-3.07	-2.51	3.11	-4.29	0.79	1.69	3.85	0.86	2.73
N	-3.14	-3.69	-2.16	-0.96	-2.23	-1.74	0.17	-1.01	0.02	0.95	-1.35	-0.45	1.37	-1.46	0.07	0.92	3.85	7.91	-0.63	-0.43
K	-2.23	0.66	-2.10	-1.71	-1.80	-0.72	-1.11	1.72	0.29	0.11	-0.27	0.87	0.14	-1.90	3.52	-0.67	0.86	-0.63	2.61	-3.54
R	-2.14	-1.42	0.19	-1.33	-0.84	-1.51	-2.74	1.22	-0.74	-0.38	4.08	-0.33	-0.40	-0.06	3.45	0.24	2.73	-0.43	-3.54	0.73

References

[1] B. Adhikari. Resdiue-residue contact driven protein structure prediction using optimization and machine learning. (July), 2017.

[2] M. Åstrand, J. Cuellar, J. Hytönen, and T. A. Salminen. Predicting the ligand-binding properties of Borrelia burgdorferi s.s. Bmp proteins in light of the conserved features of related Borrelia proteins. *Journal of Theoretical Biology*, 462:97–108, 2019.

[3] B. C. Becker and E. G. Ortiz. Evaluation of face recognition techniques for application to facebook. In *2008 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, pages 1–6. IEEE, sep 2008.

[4] F. K. Brown, E. C. Sherer, S. A. Johnson, M. K. Holloway, and B. S. Sherborne. The evolution of drug design at Merck Research Laboratories. *J. Comput. Aided. Mol. Des.*, 31(3):255–266, mar 2017.

[5] S. Choudhuri. Sequence Alignment and Similarity Searching in Genomic Databases. In *Bioinforma. Beginners*, pages 133–155. Elsevier, 2014.

[6] A. V. Finkelstein, A. J. Badretdin, O. V. Galzitskaya, D. N. Ivankov, N. S. Bogatyreva, and S. O. Garbuzynskiy. There and back again: Two views on the protein folding puzzle. *Phys. Life Rev.*, 21:56–71, 2017.

[7] A. Fout, B. Shariat, J. Byrd, and A. Ben-Hur. Protein Interface Prediction using Graph Convolutional Networks. *Nips*, (Nips):6512–6521, 2017.

[8] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, and A. R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, jan 2017.

[9] J. W. Gooch. Primary Structure. In *Encycl. Dict. Polym.*, volume 17, pages 917–917. Springer New York, New York, NY, 2011.

[10] J. C. Jeong, X. Lin, and X. W. Chen. On position-specific scoring matrix for protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):308–315, 2011.

[11] M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, jan 2000.

[12] S. P. Leelananda and S. Lindert. Computational methods in drug discovery. *Beilstein J. Org. Chem.*, 12(January):2694–2718, 2016.

[13] N. Mathai, Y. Chen, and J. Kirchmair. Validation strategies for target prediction methods. *Briefings in Bioinformatics*, 00(April):1–12, 2019.

[14] A. Mishra, P. Pokhrel, and M. T. Hoque. Thesis – StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics*, 35(3):433–441, feb 2019.

[15] H. Öztürk, A. Özgür, and E. Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[16] A. A. Schaffer. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14):2994–3005, jul 2001.

[17] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, and T. Aittokallio. Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, mar 2014.

[18] T. UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 46(5):2699–2699, mar 2018.

[19] D. H. Wolpert and W. G. Macready. Coevolutionary free lunches. *IEEE Trans. Evol. Comput.*, 9(6):721–735, 2005.

[20] A. K. Wong, H. Y. Sze-To, and G. L. Johanning. Pattern to Knowledge: Deep Knowledge-Directed Machine Learning for Residue-Residue Interaction Prediction. *Scientific Reports*, 8(1):1–14, 2018.

[21] R. Xu, J. Zhou, H. Wang, Y. He, X. Wang, and B. Liu. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Systems Biology*, 9(1):1–12, 2015.

[22] L. Zhang, X. Zhao, and L. Kong. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. *Journal of Theoretical Biology*, 355:105–110, aug 2014.