TRIBHUWAN UNIVERSITY
INSTITUTE OF ENGINEERING, PULCHOWK CAMPUS
LALITPUR, NEPAL

073 MSCS 652

PROTEIN DRUG INTERACTION FROM THEIR SEQUENCE:USING SMILES AND FASTA

BY
ANUP ADHIKARI

SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE AND KNOWLEDGE ENGINEERING

NOVEMBER, 2019

RECOMMENDATION

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering for acceptance, a thesis entitled "*Protein Drug Interaction from their sequence:Using SMILES and FASTA* ", submitted by Anup Adhikari in partial fulfillment of the requirement for the award of the degree of "Master of Science in Computer System and Knowledge Engineering.

.....

Supervisor

Dr. Surendra Shrestha (PhD)

.....

External Examiner

Name:

DEPARTMENTAL ACCEPTANCE

(Use letter department's letter pad for this page)

The thesis entitled "Protein Drug Interaction from their sequence:Using SMILES and FASTA ", submitted by Anup Adhikari in partial fulfillment of the requirement for the award of the degree of "Master of Science in Computer System and Knowledge Engineering" has been accepted as a bonafide record of work independently carried out by him in the department.

. . . . . .

Dr. Surendra Shrestha (PhD)

Head of the Department

Department of Electronics and Computer Engineering,

Pulchowk Campus,

Institute of Engineering,

Tribhuvan University,

Nepal.

COPYRIGHT

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%.

Keywords: Convolutional Neural Network, Protein, Drug, No Free Lunch Algorithm

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

LIST OF TABLES

ACRONYMS

Adam Adaptive Moment Estimation. 27

CADD Computer-aided drug discovery. 2, 3

CASP Critical Assessment of Structure Prediction. 9

CNN Convolutional Neural Network. 5, 23–25, 30, 39

EDT Evolutionary Distance Transformation. 4, 6

ET Evolutionary Transformation. 14

FeatDTI Feature Based Drug Target Interaction. 6, 30

HTS High Throughput Screening. 2

KDE Kernel Density Estimation. 18

KIBA Kinase Inhibitor Bioactivity. 3, 14, 31

NFL No Free Lunch Algorithm. 7, 40

PSSM Position Specific Scoring Matrix. 6, 15, 16, 19, 29, 39

PSSM-DT Position Specific Scoring Matrix Distance Transformation. 4, 6, 14, 30, 31

QSAR Quantitative Structure Activity Relation. 3

R2RSRV Residue Residue Statistical Residual Vector. 3, 20, 39

RPT Residue Probing Transformation. 4, 6, 14, 30

SGD Stochastic Gradient Descent. 27

CHAPTER ONE : 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 interest 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 bio-informatics ever since the "Next Industrial Revolution" possibilities started grow (Leelananda and Lindert, 2016; Brown et al., 2017). 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 structural 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 2111 drugs from CHEMBL and 229 proteins from UniProt to get their structural information. The interaction of 180244 is obtained from the research work produced by (Tang et al., 2014), and by removing the unrecognized interactions. The interactions are 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. While 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 (Wong et al., 2018). Therefore, modeling a machine

[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.

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 were used for the prediction of drug and protein interaction problem. The main reasons are: STITCH scores don't fully explore the primary thermodynamic dissociation constants used for drug-target interaction profile. The information of other scores are present in KIBA (Tang et al., 2014), as shown in section 3.3.1. Again, KIBA scores dataset is sourced from multiple databases (Kanehisa, 2000), (Wishart et al., 2018), (Hecker et al., 2012) and (Sharma et al., 2010) a 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 their essential roles in cellular signaling transduction for many cancers and inflammatory diseases (Tang et al., 2014; Kanehisa, 2000). We concentrate on proteins dataset, specifically because their interaction is quite tricky when considered among chemical, atomic, structural and electrical nature of protein residues (Mathai et al., 2019). Our basis for forming the matrices and vectors related to protein sequence came 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: labeled encodings, PSSM-DT, EDT and RPT.

1.3 Objectives

The objectives of the research are:

- To determine the effective feature matrices related to protein.
- To modify DeepDTA - a deep learning architecture using Convolutional Neural Network (CNN) for predicting the protein-drug interactions.

1.4 Scope of Work

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.

1.5 Organization of Thesis

CHAPTER 1 is the introductory chapter that includes background of the study, problem statement, objectives of thesis and scope of the work.

CHAPTER 2 is the literature review about the work. It describes No Free Lunch Algorithm and Generalized Optimization; their concepts and implications to this work. The build up describes how the various literatures assist to provide motivation to this work.

CHAPTER 3 describes details of Methodology used for Protein and Drug Interaction prediction problem. The system block of Feature Based Drug Target Interaction (FeatDTI) is described. The properties of data used for training the model are described. It underlines the principles on Position Specific Scoring Matrix (PSSM), Position Specific Scoring Matrix Distance Transformation (PSSM-DT), Evolutionary Distance Transformation (EDT), Residue Probing Transformation (RPT), and labeled encodings used for feeding the Deep Learning Network. The components of Feature Based Drug Target Interaction (FeatDTI) used as deep learning model in this work are described.

CHAPTER 4 describes the experimental settings and results produced in this work. The analysis of work produced is described in detail.

CHAPTER 5 AND CHAPTER 6 describe the conclusions from the work and recommendations for future work.

APPENDIX enlists the data parameters used in the work and describes their integration in this work.

REFERENCES enlists the references used for this thesis completion.

CHAPTER TWO : LITERATURE REVIEW

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. (Wolpert and Macready, 2005)

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.(Leelananda and Lindert, 2016)

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. (CAS, 2018)

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.(Åstrand et al., 2019) 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 (Fout et al., 2017).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(Finkelstein et al., 2017). 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 (Adhikari, 2017).

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 (Becker and Ortiz, 2008). 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. (Choudhuri, 2014)

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.(Gooch, 2011) 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.(Wong et al., 2018)

The problem formulation of protein-drug interaction categorizes the efficiency of protein-interaction prediction method. The methods based on binary classification has high ac-

curacy, but has a big bias problem. This is because the unidentified interactions are also regarded as non-interactive pairs. (Mahato, 2016; Tang et al., 2014). The dataset is quite resourceful in case of classification problem. However, due to disadvantages of misclassification and same classification value for low-interactive and highly-interactive protein-drug pairs, regression problem is the better modificatiion of problem situation. (Tang et al., 2014)

No Free Lunch Theorem (Wolpert and Macready, 2005) on the other hand works by basing the prediction guesses based on a number of domain representations. The domain representations of input data is solved using multiple functions to get the domain specific abilities of problem situation. 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 THREE : METHODOLOGY

3.1 System Overview

3.1.1 System Block

A protein-drug interaction problem goes through various stages of data processing for building a prediction system in a Deep Learning Network. The systematic diagram is shown by Figure 3.1. 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 as shown in 3.2 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.

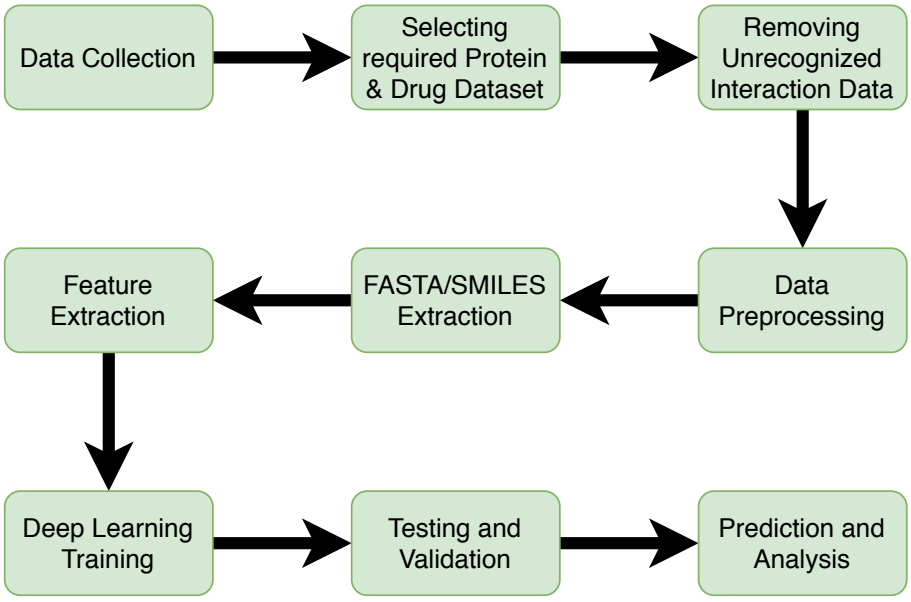


Figure 3.1: System Block Diagram for Protein-Drug Prediction

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

MEASLGIQMDEPMAFSPQRDRFQAEGSLKKNEQNFKLAGVKKDIEKLYEAVPQ
LSNVFKIEDKIGEGTFSSVYLATAQLQVGPEEKIALKHLIPTSHPIRIAAELQCLTV
AGGQDNVMGVKYCFRKNDHVVIAMPYLEHESFLDILNSLSFQEVREYMLNLFK
ALKRIHQFGIVHRDVKPSNFLYNRRLKKYALVDFGLAQGTHDTKIELLKFVQSE
AQQERCSQNKSHIITGNKIPLSGPVPKELDQQSTTKASVKRPYTNAQIQIKQGKD
GKEGSVGLSVQRSVFGERNFNIHSSISHESPAVKLMKQSKTVDVLSRKLATKKK
AISTKVMNSAVMRKTASSCPASLTCDCYATDKVCSICLSRRQQVAPRAGTPGFR
APEVLTKCPNQTTAIDMWSAGVIFLSLLSGRYPFYKASDDLTALAQIMTIRGSRE
TIQAAKTFGKSILCSKEVPAQDLRKLCERLRGMDSSTPKLTSDIQGHASHQPAIS
EKTDHKASCLVQTPPGQYSGNSFKKGDSNSCEHCFDEYNTNLEGWNEVPDEAY
DLLDKLLDLNPASRITAEEALLHPFFKDMSL

PubCHEM and CHEMBL are drug databases used for feature extraction of drug molecules. PubCHEM is a database containing 96,881,514 drug compounds and associates 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 (Tang et al., 2014) 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 (Tang et al., 2014; Öztürk et al., 2018). Higher score means that there is more strength of interaction between the two. We use 2111 ligands as drugs and 229 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

The prediction system is built upon the feature extraction of proteins and drugs. 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 (UniProt Consortium, 2018) 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.

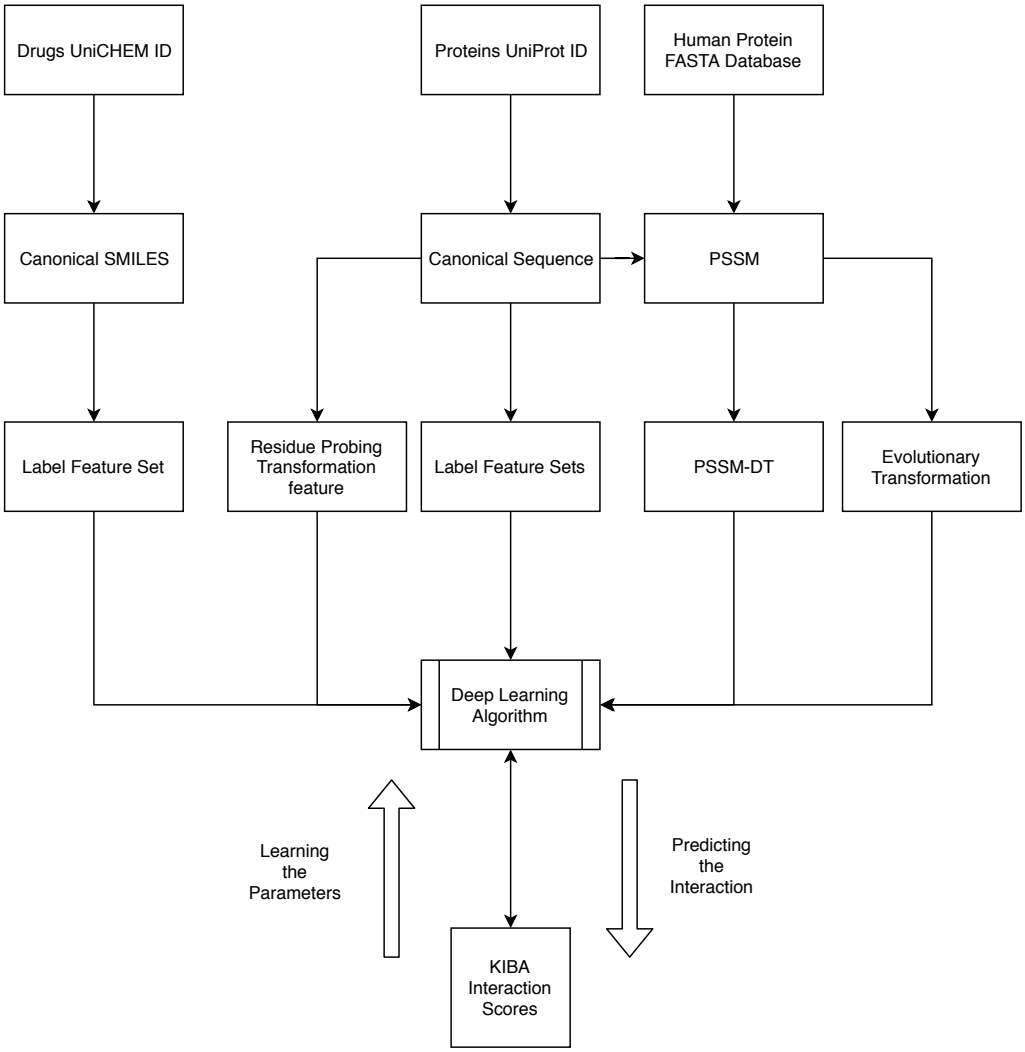


Figure 3.2: Schematic Block Diagram for Protein-Drug Prediction

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 (Tang et al., 2014). 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} (Tang et al., 2014) 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(Gaulton et al., 2017). Based on interaction data available, we remove the unknown values and obtained a total of 118254 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 229 proteins and 2111 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

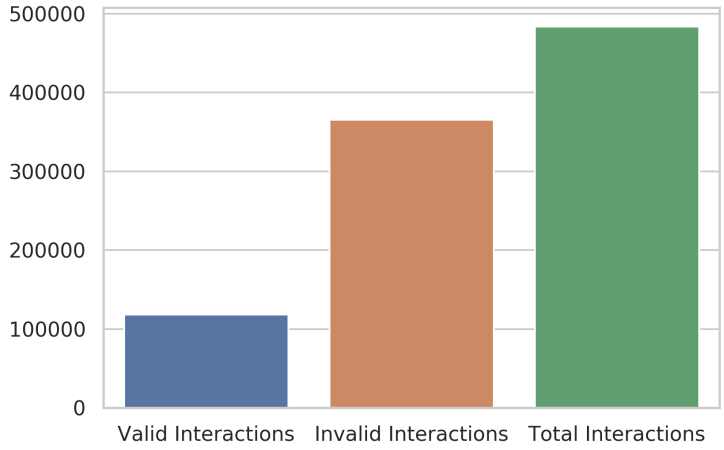


Figure 3.3: Bar Chart: Interactions of available dataset

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 conventional process of calculating PSSM score values. The sequence following shows the process of calculating the scores once the PSSM distribution of the whole protein sequence 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.

> ACTCagccccagcGGAGGTGAAGGACGTCCTTCCCCAGGAGCCGGTG
> AGAAGCGCAGTCGGGGGCACGGGGATGAGCTCAGGGGCCTCTAG
> AAAGATGTAGCTGGGACCTCGGGAAGCCCTGGCCTCCAGGTAGT
> CTCAGGAGAGCTACTCAGGGTCGGGCTTGGGGAGAGGAGGAGCG
> GGGGTGAGGCCAGCAGCA

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

3.3.3 PSI-BLAST

PSI-Blast tools relates with multiple sequence alignments from a family of protein sequences(Schaffer, 2001). This helps to create a PSSM - Equation (3.5) - matrix referred

Table 3.2: PSSM Scoring Matrix Calculation Steps: (a) shows 10 assumed protein fasta sequences. Their lengths is equal to 9 for this example. (b) shows the frequency of each amino acids (labelled in first column of the table with alphabetic symbols) present in positional index of fasta sequence from table (a). (c) shows the probability of amino acids appearing in different positions of fasta sequence for the protein in table (a). (d) shows the calculation of score for string'AGCCCCAGC' from example in 3.3.2. This matrix is used to tabulate PSSM matrix at each positional index of the example protein. (e) tabulates the scores obtained from (d) by sliding window of length 9 at first 10 positions for example in 3.3.2.

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

(a) Protein FASTA Sequence (Motifs)

	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

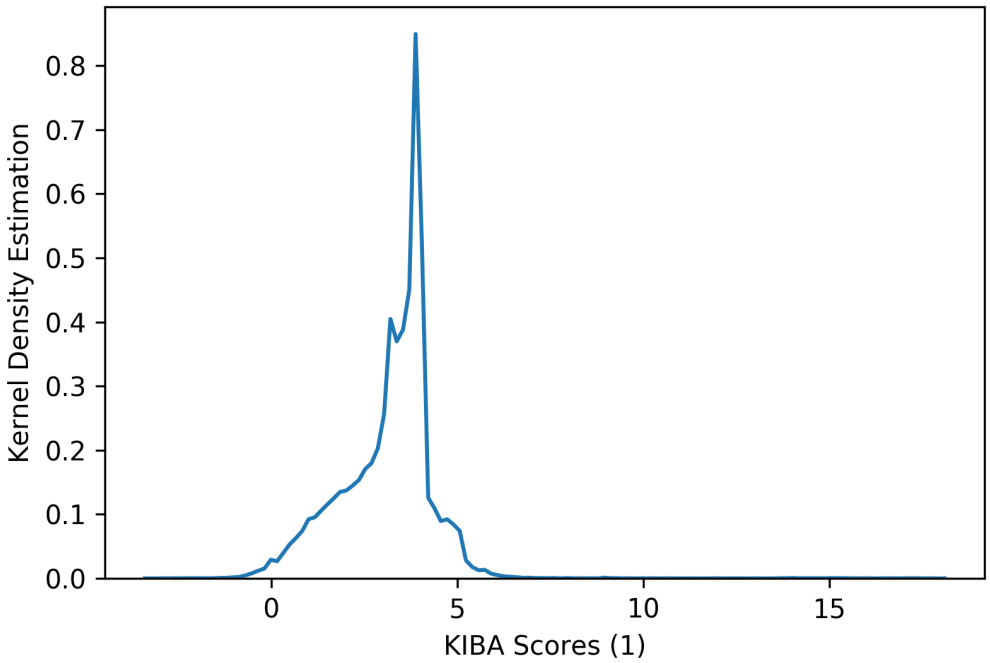
(c) Log-Likelihood Matrix

	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									

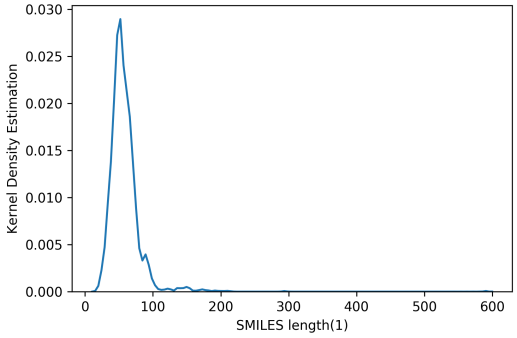
(d) Sliding Window Score Calculation

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

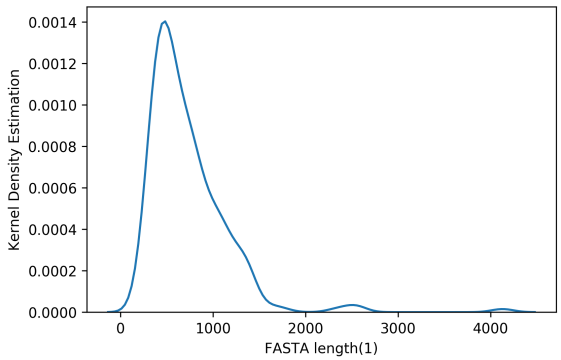
(e) Score of sliding window motifs



(a) Drug-Protein Interaction



(b) Drugs SMILES Sequence



(c) Protein FASTA Sequence

Figure 3.4: Kernel Density Estimation (KDE) Distribution of KIBA-interaction scores of Drug Sequences and Protein Sequences. (a) KDE Distribution of KIBA Scores in Protein-Drug Interaction Pair, (b) KDE Distribution of length in Labeled Encodings of Drug Sequence, (c) KDE Distribution of length in Labelled Encodings Protein Sequence

to as secondary protein structure. For this study, the PSSM profile of every protein sequence is obtained by executing iteration of PSI-BLAST against (Schaffer, 2001, KEGG) protein. PSSM profile 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.(Schaffer, 2001, 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 (Xu et al., 2015). 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(Zhang et al., 2014). 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 (Wong et al., 2018) 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 A.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., 2011), and implemented by (Mishra et al., 2019), emphasize domains with similar conservation rates by grouping domain families based on their conservation score in the PSSM profile.

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

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 3.5. The input layers are described in table 3.4.

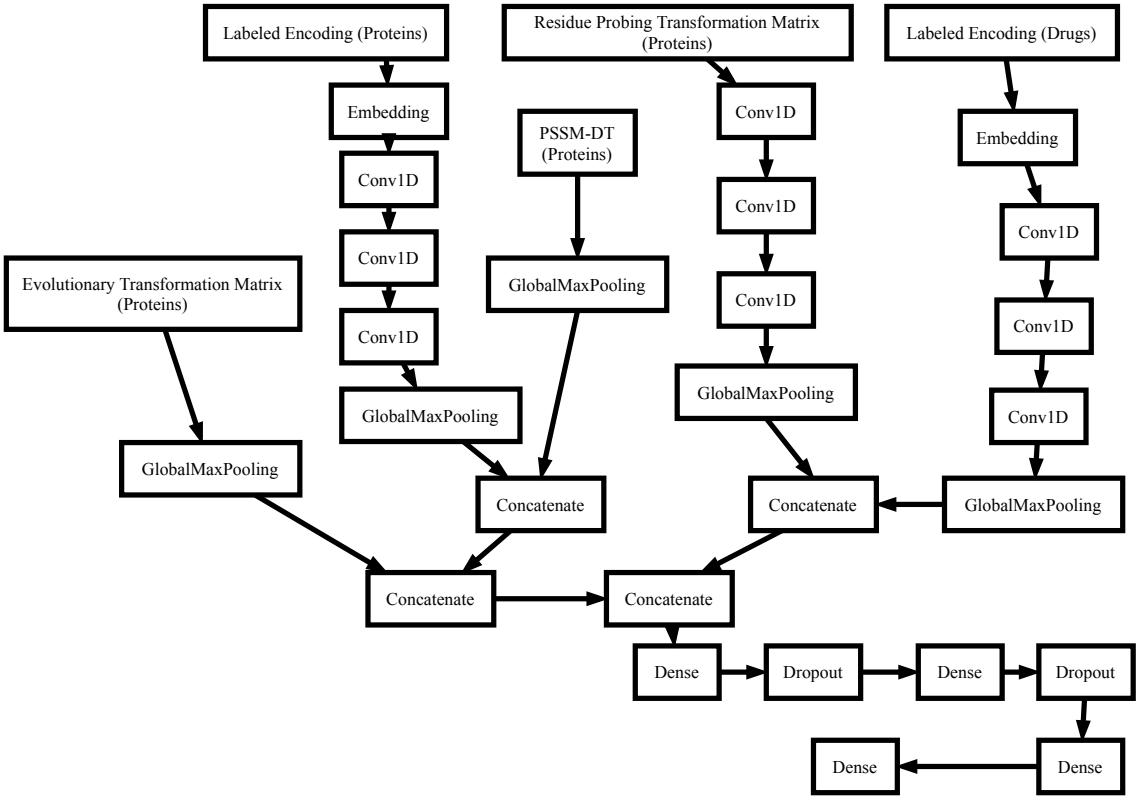


Figure 3.5: Deep Learning Model to predict Protein-Drug Interaction

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 3.5, the Embedding feature provided by keras for vector representation of both drug fingerprint and protein sequence are utilized. The label 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

Convolutional Neural Network (CNN) are used for feature generation in machine learning system. They improve a machine learning system in three perspectives: sparse interactions, parameter sharing and equivariant representations. While a conventional neural network is tight network for instance, a Dense Layer creates an interaction of every input unit to that of output units; CNN have sparse interactions. This differentiates the learning pattern of CNN from other networks by understanding the local patterns in the input vector. While traditional Neural Networks mostly involve learning the global parameters, CNN is used to learn local patterns. It does so by reducing the kernel (aka filter) smaller than the input. For example, when an image can have thousands of pixels, the kernel size can be tens or hundreds of pixels as shown in Figure 3.6.

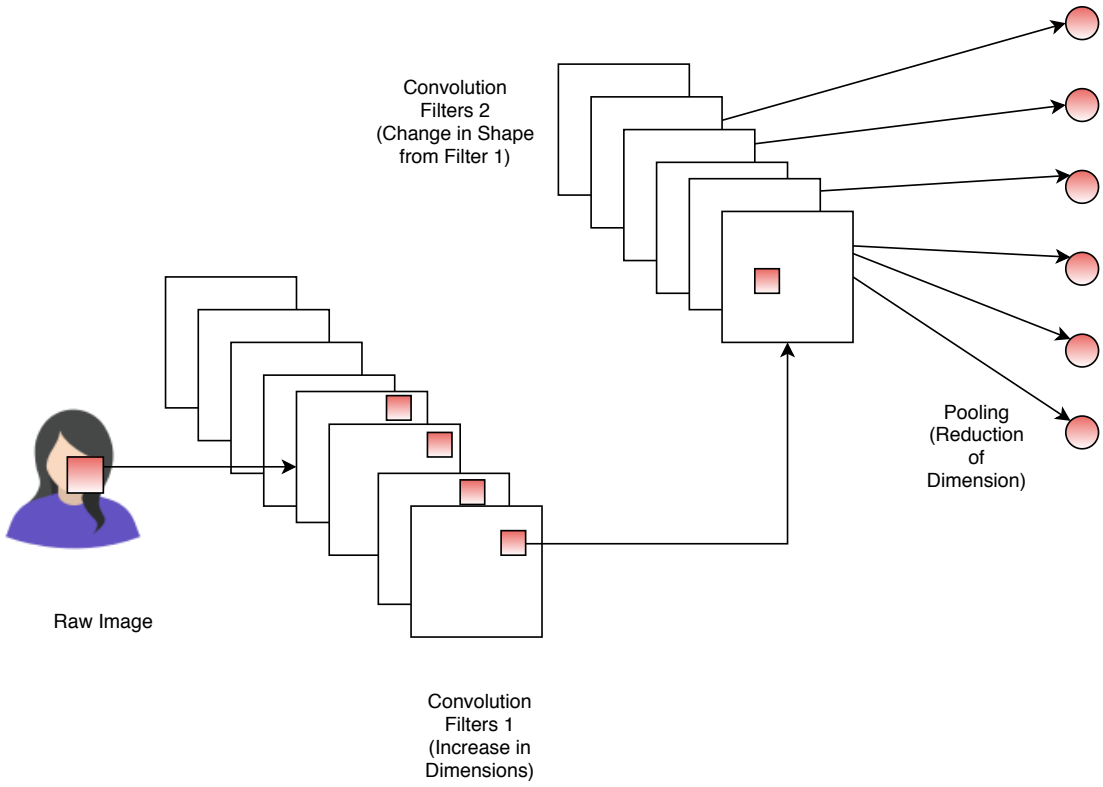


Figure 3.6: Working of CNN Block:The small window from the image grid forms a filter. In first layer there is use of 8 filters

Parameter Sharing refers when the same parameter is used for multiple times in model. In conventional neural net, each parameter is used only once when computing the output of a layer. In CNN, as the window sweeps over all the images, the same pixels create different representations with the position of sweeping window. So the parameter sharing property of CNN identifies one representation for the pixel at a time.

Equivariance property of CNN is caused due to parameter sharing. It is defined as a property defined as when the input is translated, so is the output. When processing some time-series data the convolution produces a timeline sequence of features that appear in the input. For example, the convolution operations in movie will produce the feature timeline for the picture pixels.

This research uses Convolutional Neural Network (CNN) to learn the sequential representation of drug and proteins. As the primary structure of proteins and drugs are in fact representations of their 2D representations, the convolution operation will help to learn such features for further analysis. Again, 3D representations are also learned by higher layers of convolution layers. (Adhikari, 2017)

Pooling Layer

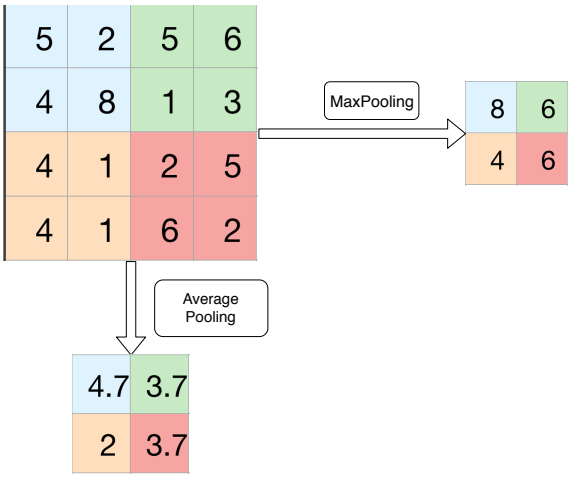


Figure 3.7: Pooling Layer: MaxPooling takes the maximum value from the pooling window and AveragePooling takes the mean from the pooling window.

The Pooling layer was used to modify the output of its preceding layer. For example the max pooling renders the maximum output from a rectangular neighborhood and aver-

age pooling renders average value from the the rectangular neighborhood. It was used to downsample the learned parameters from the grid of 2 dimensions returned by Convolution Layer. This work used Global Max Pooling to reduce the dimension and extract the extreme features learned from CNN. Thus, it got reduced to 1 dimension by taking the highest values from the window size(corresponding to shape of 1^{st} dimensional element). The Pooling operation has been described by Figure 3.7.

Dense Layer

Dense Layer is a neural layer which fully connects the input layer to output layer. It performs a linear operation on the layer's input vector. At every node in output of the dense layer generally follows an activation function that creates a generalization rule for the input vectors at the node. The research work used the dense layer to learn the global pattern from the feature data. The representation can be seen from Figure 3.8. As the output required is a regression value, it uses relu for activation.

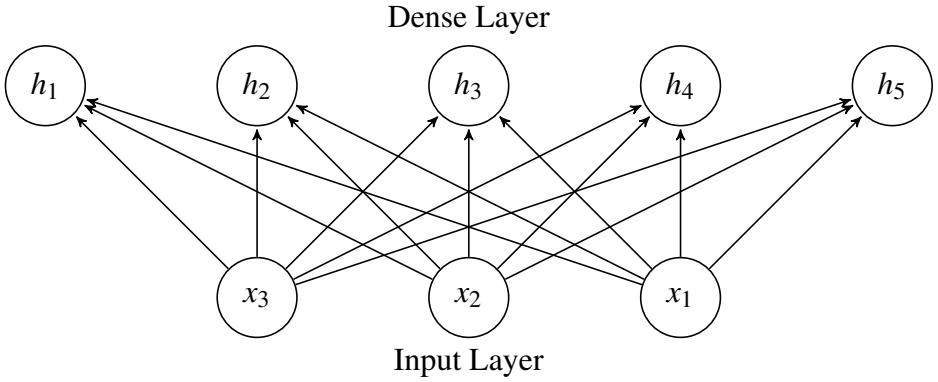


Figure 3.8: The input nodes are represented by x and the weights h are applied to each values of input nodes to generate an output

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 has been expressed diagrammatically in Figure 3.9.

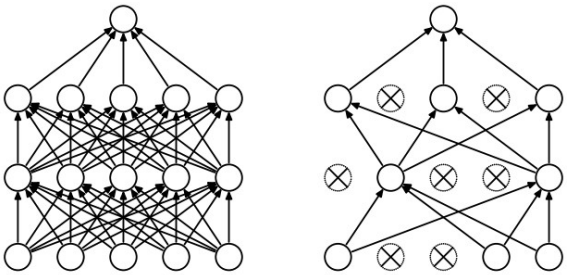


Figure 3.9: a) Standard neural network whose all the nodes have weights connected to higher nodes and lower nodes. b) Certain nodes belonging to same levels are disconnected. Some weights are also disconnected from other nodes depending on the percentage of dropout applied.

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. The first dimensional length of input matrices and their no. of dimension should be same to concatenate the matrices.

Activation Layer

Activation Layer is a function that takes an input and provides an output based on the value. There are various kinds of Activation functions like Sigmoid, ReLu, Leaky Relu, tanh, Gaussian, Sinusoid etc. The research uses ReLu function for the activation layers.

Rectified Linear Unit

The output of Rectified Linear Unit (ReLu) is from 0 to infinity. For a normalized ReLu, the output will be between 0 to 1. The parametric equation is shown in Eq 3.13.

$$f(x) = \begin{cases} 0 & ,for \quad x <= 0 \\ x & ,for \quad x > 0 \end{cases} \tag{3.13}$$

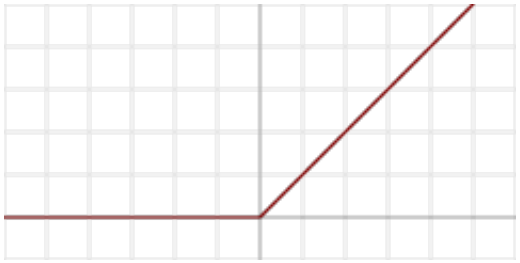


Figure 3.10: Relu Activation Function

3.4.2 Choice of Optimizers

Optimizers are the functions that estimate the global maximum/minimum for a problem domain. Their characteristics to find the global extremes or local extremes depend on the type of problem type and the optimization parameters used to train the algorithm. The choice of optimizer influences both the speed of convergence and whether it occurs. The optimizers aim to minimize the cost function $J(\theta;x^{(i)};y^{(i)})$ where θ is the optimization rule for cost function and $x(i), y(i)$ are the input,output parameters. The optimizers chosen to minimize the cost function associated to protein-drug prediction are:

Stochastic Gradient Descent(SGD)

The optimization formula for Stochastic Gradient Descent (SGD):

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta;x^{(i)};y^{(i)}) \tag{3.14}$$

Adaptive Moment Estimation(Adam)

The optimization formula for Adaptive Moment Estimation (Adam):

$$\begin{aligned} V_{d\theta} &= \beta_1 V_{d\theta} + (1-\beta_1)\, d\theta \\ S_{d\theta} &= \beta_2 S_{d\theta} + (1-\beta_2)\, d\theta \\ V\,\mathrm{corr}_{d\theta} &= \frac{V_{d\theta}}{(1-\beta_1)^t} \\ Scorr_{d\theta} &= \frac{S_{d\theta}}{(1-\beta_2)^t} \\ \theta &= \theta - \alpha \frac{d\theta}{\sqrt{S_{d\theta}} + \varepsilon} \end{aligned} \tag{3.15}$$

Root Mean Squared Propagation(RMSProp)

$$W_{dW} = \beta W_{dW} + (1-\beta)\nabla\theta^2 \tag{3.16}$$

$$\theta = \theta - \eta \cdot \frac{\nabla_\theta}{\sqrt{W_{dW}} + \varepsilon} J(\theta; x^{(i)}; y^{(i)}) \tag{3.17}$$

Adaptive Gradient Algorithm(AdaGrad)

$$W_{dW} = \beta W_{dW} + (1-\beta)\nabla\theta \tag{3.18}$$

$$\theta = \theta - \eta \cdot W_{dW} J(\theta; x^{(i)}; y^{(i)}) \tag{3.19}$$

CHAPTER FOUR : RESULTS

4.1 Experiments

The focus of this research is on the properties of protein owing to its' complex structures. The binding of protein and drug depend on various attributes of protein such as acidity, hydrophobicity, binding pockets and the structure of drug. The attributes are closely related to primary and secondary structure of protein themselves. Therefore, our model aims to relate these multiple components with matrix representation and to confirm to predictions made as per Figure 3.4 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(Schaffer, 2001). 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

The architecture as per Figure 3.5 was implemented for our model design via Python using the TensorFlow framework consisting of keras. The training contained of 200 epochs. The changing learning rates and early stopping were manipulated in the training process by using keras callback functions. The training and testing was performed using a 60:20:20 cross-validation set prepared in dataset He et al. (2017). To evaluate the performance of the model, we used concordance index(CI)(Xu et al., 2015) 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:

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

Table 4.1: Experimental Settings

Description	**Settings**
No. of filters	32
Filter Length(Compounds)	4,8
Filter Length(Proteins)	8,12
Epochs	200
Hidden Neurons	1024, 1024, 512
Batch-size	256
Dropout	0.1
Optimizer	Adam
Learning Rate	0.001
Training, Test, Validation splits	60,20,20

The method proposed by this research, Feature Based Drug Target Interaction (FeatDTI) contains 3 Convolutional Neural Network (CNN)-blocks. It evaluates Residue Probing Transformation (RPT), acrfulledt and Position Specific Scoring Matrix Distance Transfor-

mation (PSSM-DT) based on the inputs (Protein FASTA and Drugs SMILES) provided. The features are then calculated by the model to calculate the Kinase Inhibitor Bioactivity (KIBA) scores.

The deep learning method was implemented with various filter sizes of Convolutional Layer of Protein and Drugs. The different sizes were chosen as the lengths of drugs' canonical SMILES sequence and proteins' FASTA sequence differ. Only the comparable results are shown in table 4.2. The training and validation plots under different settings is clearly shown in Figure 4.3. The different settings described by Figure 4.1 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 are not 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 do not appear in training set.

Here we obtain four different models for prediction of drug and protein interaction. The correct model to choose will depend on type of drug and protein fed into the learning algorithm. To use the correct model, the research work following steps depending on check with input protein FASTA and drug SMILES sequence being fed:

- If Both drug and protein are found in the training dataset, use model S1 to predict the KIBA score.
- If drug is not found but protein is found in training dataset, use model S2 to predict the KIBA score.
- If protein is found but drug is not found in training dataset, use model S3 to predict the KIBA score.
- If none of the proteins or drugs is found in training dataset, use model S4 to predict the KIBA score.

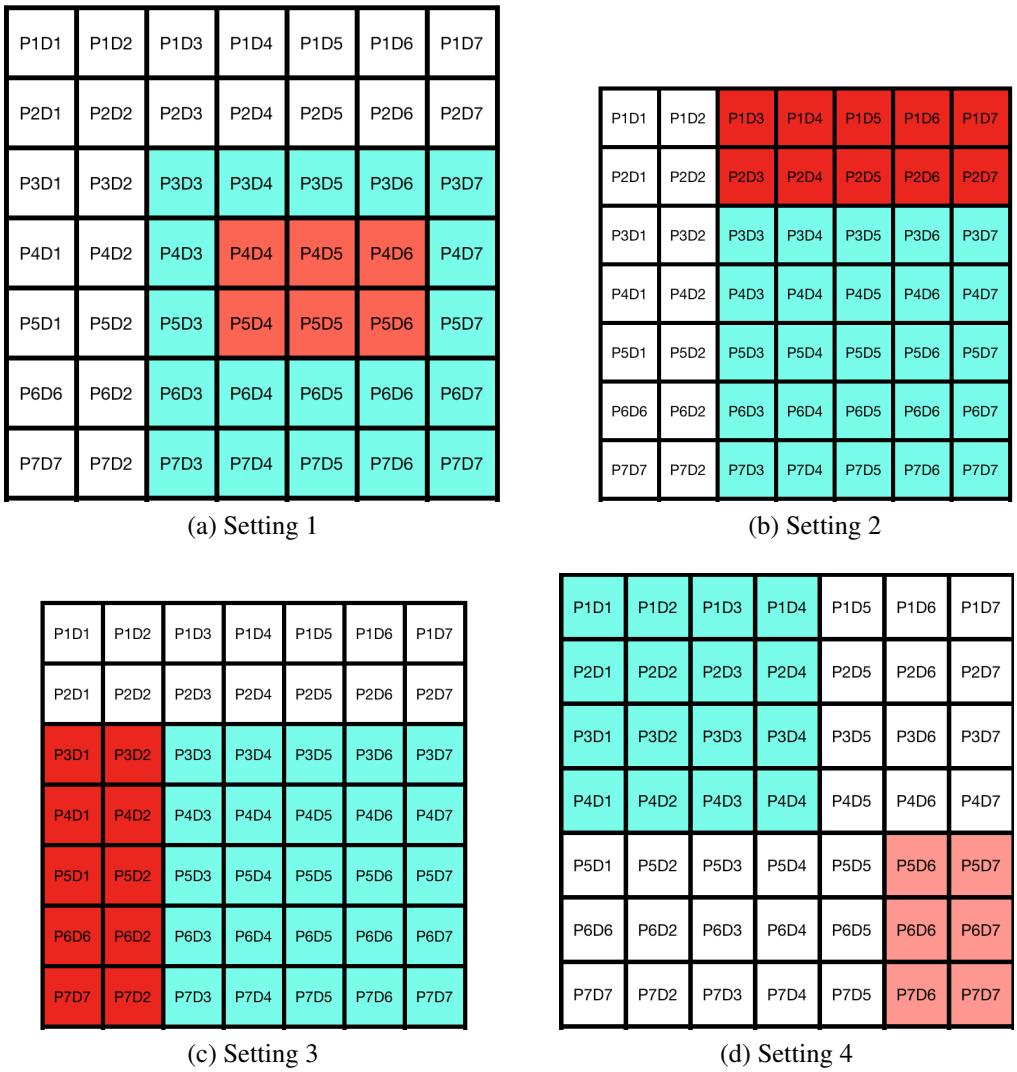


Figure 4.1: Different Settings Of Training and Validation Sets *(Brown-Red represents the Validation set and Cyan represents the Training Set)*, Description based on validation Set: (a) has repeated protein and drugs, (b) has repeated drugs but unique proteins, (c) has repeated proteins but unique drugs, and (d) has unique proteins and drugs.

Table 4.2: Experiments results under different settings (S1, S2, S3, S4). The model S1 and S4 performed well with same filter size 8 in convolution layer drug-protein. The model S2 performed well with drug filter window of 4 and protein window of 8. The model S3 predicted well with drug window size of 8 and protein window size of 12.

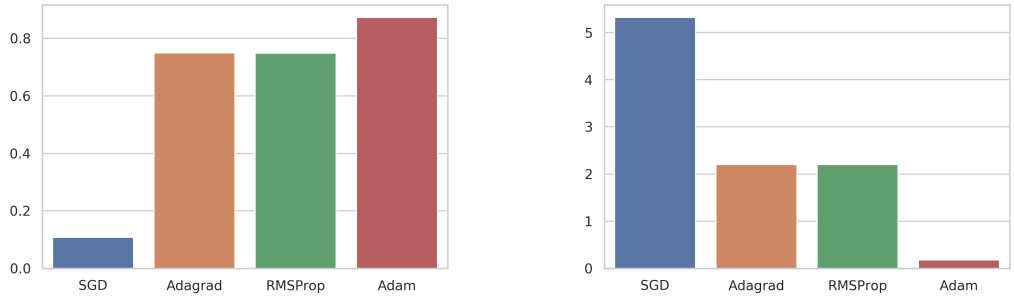
S.No.	Setting	Drug Smiles Window Size	FASTA Window Size	CI-val	MSE
1	S1	4	8	0.813936	0.893064
2	S1	4	12	0.810844	1.285271
3	S1	8	8	0.873094	0.177053
4	S1	8	12	0.807230	1.056755
1	S2	4	8	0.788051	0.678707
2	S2	4	12	0.802355	0.360900
3	S2	8	8	0.803720	0.792821
4	S2	8	12	0.803853	0.672790
1	S3	4	8	0.815181	1.562749
2	S3	4	12	0.805739	1.541746
3	S3	8	8	0.813063	1.259206
4	S3	8	12	0.824767	1.396765
1	S4	4	8	0.803982	0.412271
2	S4	4	12	0.805961	1.510426
3	S4	8	8	0.815464	1.614631
4	S4	8	12	0.787073	0.485298

Performance of optimizers in Setting 1

The values obtained by using different optimizers is shown in table 4.3 and Figure 4.2. It shows that Adam optimizer exhibits the best optimizer results in comparison to Stochastic Gradient Descent, Adagrad and RMSProp. In fact the other optimizers perform quite poorly compared to Adam optimizer. The learning factor was lowered by using callback ReduceLROnPlateau by a factor of 0.2 when the validation loss didn't increase in 5 consecutive epochs.

Table 4.3: Scores obtained using other optimizers

Selection of Optimizer	CI-Index	MSE
SGD	0.1087	5.320060
Adagrad	0.749753	2.202080
RMSProp	0.748766	2.202080
Adam	0.873094	0.177053



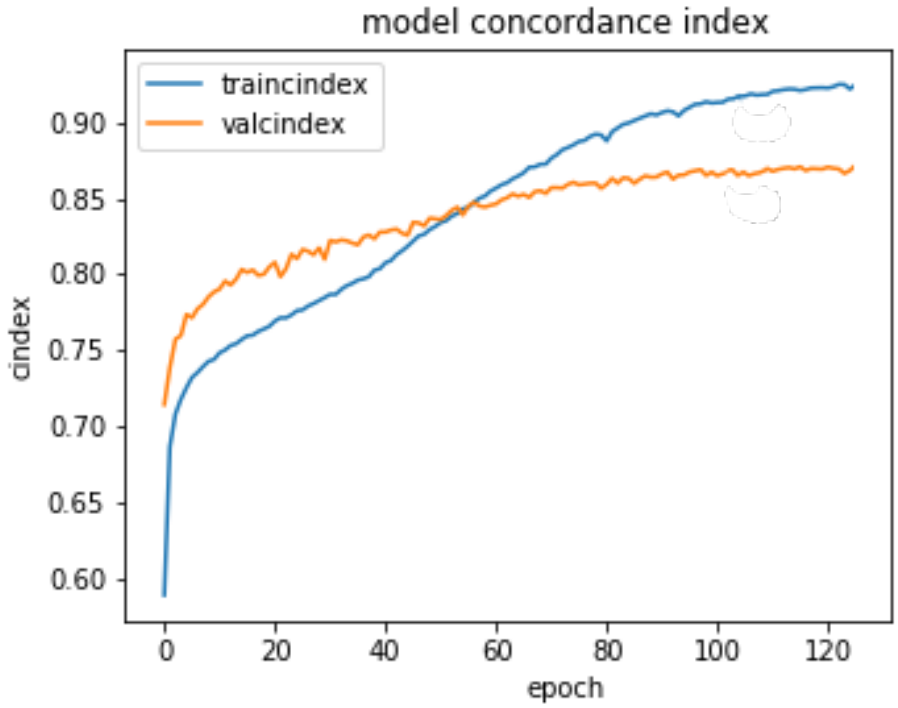
(a) Concordance Index

(b) Mean Square Error

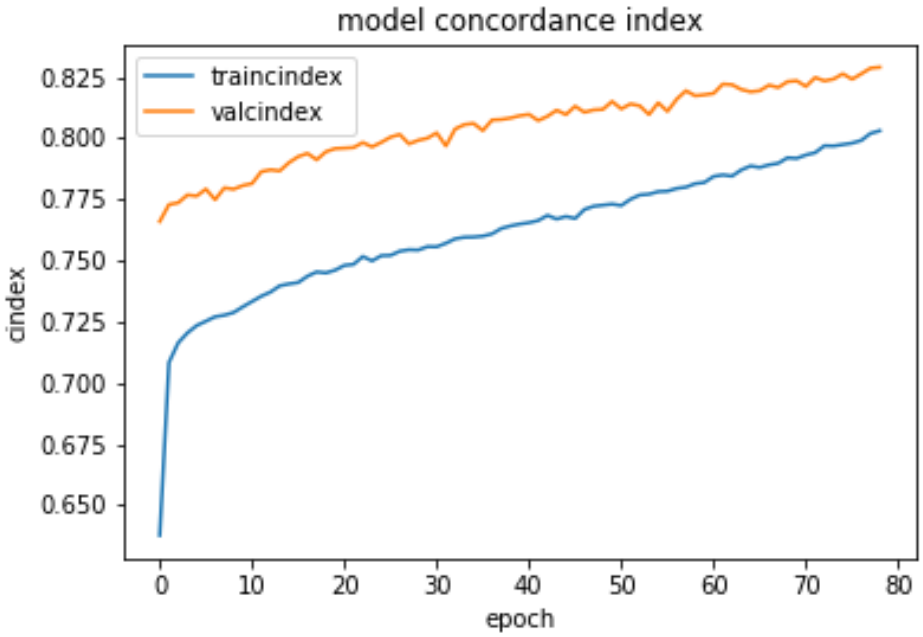
Figure 4.2: Bar Chart: Plot of concordance index (CI) and Mean Squared Error(MSE). Adam Optimizer (a) shows highest CI score and (b) shows lowest MSE among all other optimizers. The worst is statistic gradient descent optimizer for protein-drug interaction prediction.

4.2 Analysis

Table 4.2 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 4.3 and 4.4.

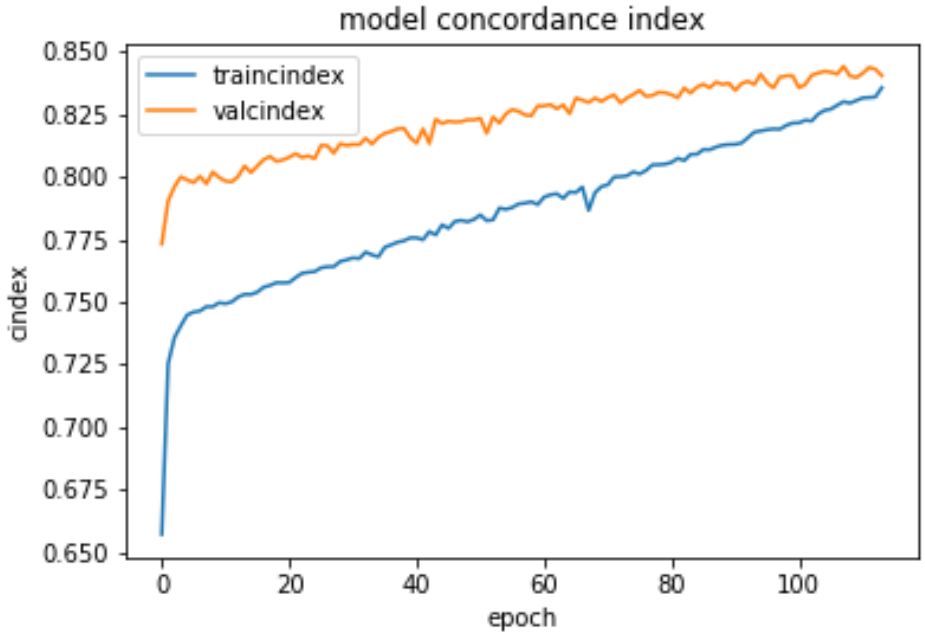


(a) Setting 1

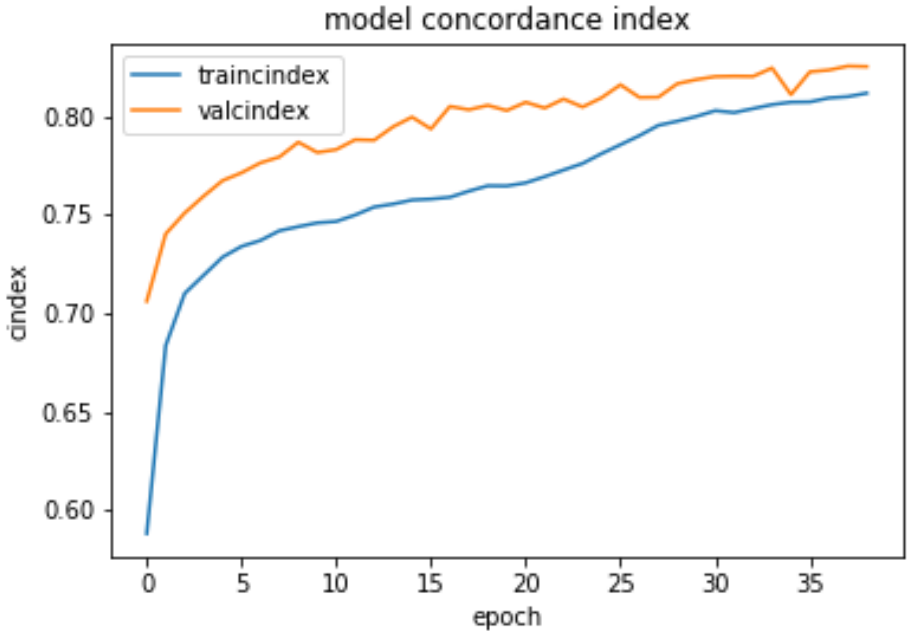


(b) Setting 2

Figure 4.3: Plot of Training-Validation C-Index Scores over various Settings: (a) has highest score and (b) shows lowest score. Under S1 Setting, the epoch reaches the optimal value after 110 epochs after which the call back automatically stops training the model further. Under S2 Setting, the model validation loss does not improve after 35 epochs.



(a) Setting 3



(b) Setting 4

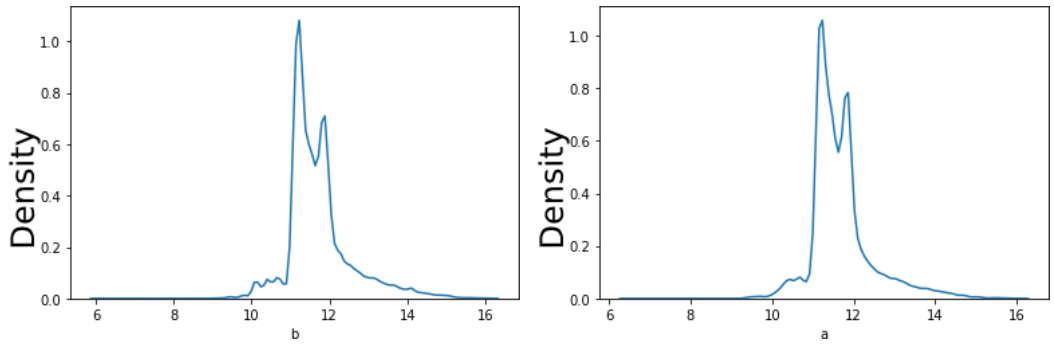
Figure 4.4: Plot of Training-Validation C-Index Scores over various Settings: (a) ranks 2nd and (b) ranks the 4rd. Setting S2 in Figure (b) shows that the trainings being comparable to Setting S3. As expected both S3 and S2 missed one of their representative drug-protein pair during training.

The Figures 4.3, 4.4 show that only S1 is a valid model. The dropout used improves the training for setting S1 only, that's why other settings S2, S3 and S4 have validation c-index score always higher than training c-index scores.

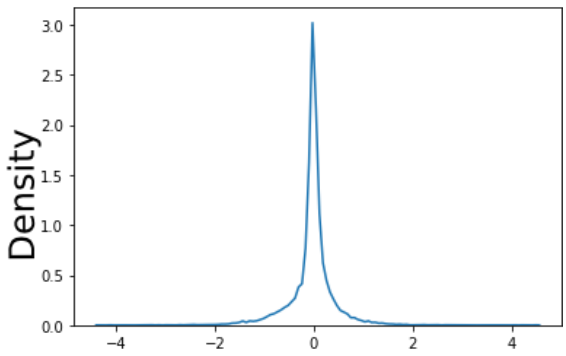
The comparison between the scores evaluated in same dataset has been compared and presented in table 4.4 (Öztürk et al., 2018). The benchmark dataset is publicly available in Journal of Chemical Information and Modeling (Tang et al., 2014; He et al., 2017).

Table 4.4: Comparison of performance: the performance of other algorithms is retrieved from literature (Öztürk et al., 2018). The method from the research is FeatDTI. It performs the best amongst all the predictors, with the closest predictor being the DeepDTA.

Methods	Proteins	Compounds	CI-Score	MSE
KronRLS	Smith-Waterman	Pubchem Sim	0.782	0.411
SimBoost	Smith-Waterman	Pubchem Simboost	0.836	0.222
DeepDTA	CNN	CNN	0.863	0.194
FeatDTI	CNN	CNN	**0.8730**	**0.177**



(a) Actual Scores (b) Predicted Scores



(c) Difference of Actual and Predicted Scores

Figure 4.5: Kernel Density Estimation(KDE) plot of Training Results based on KIBA Score Prediction: The model evaluated the data set with actual KIBA scores (a). We obtain the predicted KDE of KIBA scores in (b). The differences in KIBA scores prediction and actual scores corresponding to same drug-protein pair is plotted in Figure (c).

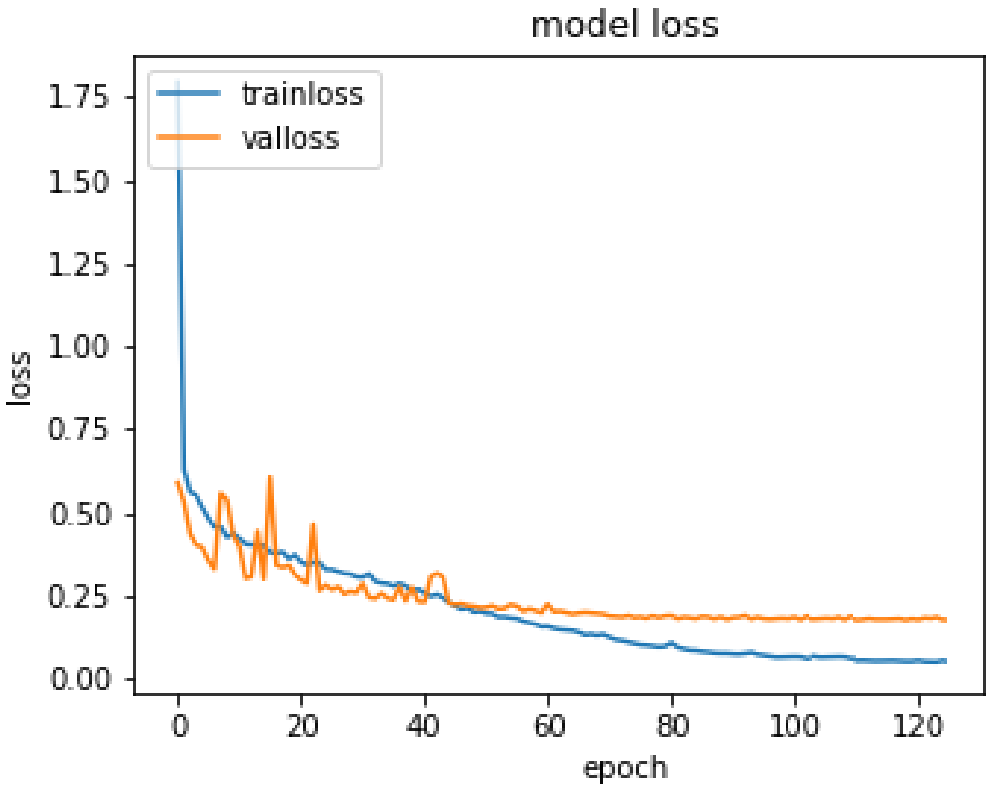


Figure 4.6: Training and Validation Loss Plot: Plot of Loss function of predicted values with training and validation set. The spikes were an unavoidable consequence of Mini-Batch Gradient Descent in Adam (batch size = 256). Some mini-batches have lossy data for optimization, inducing spikes during the initial phases of training. The loss plot was observed when trained with testing and validation drug-protein pair contained the drugs and proteins sequences used during the training phase. The filter size was 8 for both drugs and proteins. The loss function was stable with validation set and test data set around 115th epoch of training.

CHAPTER FIVE : CONCLUSIONS AND RECOMMENDATIONS

The research was conducted for series of experiments using different CNN architectures and validated using four different settings of drug and protein combinations. The experiments show that the network required past knowledge on proteins and drugs to yield the best results. The model would yield best scores when both the protein and drug were present in the training phase. We also conclude that if the domain has fewer parameters to learn from, we can represent them in new dimensions to achieve better results in machine learning problem.The research work concludes with two major statements.

5.1 Conclusions

Proteins need to be represented in higher dimensions relating to its properties of evolutionary components and physicochemical components. The transformation vectors obtained from Position Specific Scoring Matrix (PSSM) of a protein relate to other similar proteins. The similarity assessed by the motifs movements find the proteins from same family, the mutational properties of proteins. Thereby, it helps to identify homologous proteins in same domain and at the same time identifies the regions of differences in two protein sequences. The physicochemical properties of proteins represented by Residue Residue Statistical Residual Vector (R2RSRV) build properly with PSSM. The main reason for getting a good performance was mainly because they built the remaining information not learnt from the sequence data.

A deep learning network architecture consisting of three stacked Convolutional Neural Network (CNN) layers was designed. Other extreme features evaluated from PSSM were directly fed to the dense neurons. The optimized network has CNN with filter length of size 8 and three sets of stacked CNN layers. The generalization of learned features was obtained with stacked dense layers.

5.2 Recommendations

At present, the sample space of No Free Lunch Algorithm (NFL) has been explored by forming the different feature-sets. The different feature sets on chemical values of drugs can be used to represent a much broader spectrum of drug molecule representation. In case of protein, protein folds could be the incorporated information which could be supplemented for the available 3D protein structures and a good prediction algorithm chosen for CASP(CAS, 2018). For the features being evaluated, the Grid-Search CV could be applied over various algorithms and Stacking Generalization could be used to create a better optimized machine learning solution to predict protein-drug interaction score from sequence information. Successively, important is a correct modeling of Pharmacophore modeling of drug-proteins pairs so that it can be applied directly to medical supervisions.

CHAPTER ONE : APPENDIX

A.1 R2RSRV

Table A.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

A.2 Proteins Description

Table A.2: Description of proteins used in the thesis work. Source: https://www.ncbi.nlm.nih.gov/pubmed/

Protein	Name	Description
O00141	Serine/threonine-protein kinase Sgk1	Serine/threonine-protein kinase which is involved in the regulation of a wide variety of ion channels, membrane transporters, cellular enzymes, transcription factors, neuronal excitability, cell growth, proliferation, survival, migration and apoptosis. Plays an important role in cellular stress response. Contributes to regulation of renal Na+ retention, renal K+ elimination, salt appetite, gastric acid secretion, intestinal Na+/H+ exchange and nutrient transport, insulin-dependent salt sensitivity of blood pressure, salt sensitivity of peripheral glucose uptake, cardiac repolarization and memory consolidation. Phosphorylates SLC9A3/NHE3 in response to dexamethasone, resulting in its activation and increased localization at the cell membrane. Phosphorylates CREB1. Necessary for vascular remodeling during angiogenesis. Sustained high levels and activity may contribute to conditions such as hypertension and diabetic nephropathy.
O00443	Phosphatidylinositol 4-phosphate 3-kinase C2 domain-containing subunit alpha	Generates phosphatidylinositol 3-phosphate (PtdIns3P) and phosphatidylinositol 3,4-bisphosphate (PtdIns(3,4)P2) that act as second messengers. Has a role in several intracellular trafficking events. Functions in insulin signaling and secretion. Required for translocation of the glucose transporter SLC2A4/GLUT4 to the plasma membrane and glucose uptake in response to insulin-mediated RHOQ activation. Regulates insulin secretion through two different mechanisms: involved in glucose-induced insulin secretion downstream of insulin receptor in a pathway that involves AKT1 activation and TBC1D4/AS160 phosphorylation, and participates in the late step of insulin granule exocytosis probably in insulin granule fusion. Synthesizes PtdIns3P in response to insulin signaling. Functions in clathrin-coated endocytic vesicle formation and distribution. Regulates dynamin-independent endocytosis, probably by recruiting EEA1 to internalizing vesicles. In neurosecretory cells synthesizes PtdIns3P on large dense core vesicles. Participates in calcium induced contraction of vascular smooth muscle by regulating myosin light chain (MLC) phosphorylation through a mechanism involving Rho kinase-dependent phosphorylation of the MLCP-regulatory subunit MYPT1. May play a role in the EGF signaling cascade. May be involved in mitosis and UV-induced damage response. Required for maintenance of normal renal structure and function by supporting normal podocyte function. Involved in the regulation of ciliogenesis and trafficking of ciliary components.

A.3 Drugs Description

Table A.4: Description of Drug Compounds.

Drug	Chemical Formula	Description	Drug Bank ID
CHEMBL10[illegible]21H16FN3OS		This compound belongs to the class of organic compounds known as phenylimidazoles. These are polycyclic aromatic compounds containing a benzene ring linked to an imidazole ring through a CC or CN bond.	DB08521
CHEMBL10[illegible]19N8		This compound belongs to the class of organic compounds known as benzimidazoles. These are organic compounds containing a benzene ring fused to an imidazole ring (five member ring containing a nitrogen atom, 4 carbon atoms, and two double bonds).	DB01705

REFERENCES

1. (2018). CASP13. In *Thirteen. Meething*, pages 1–221, Riviera Maya, Mexico.

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

3. Åstrand, M., Cuellar, J., Hytönen, J., and Salminen, T. A. (2019). 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.

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

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

6. Choudhuri, S. (2014). Sequence Alignment and Similarity Searching in Genomic Databases. In *Bioinforma. Beginners*, pages 133–155. Elsevier.

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

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

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

10. Gooch, J. W. (2011). Primary Structure. In *Encycl. Dict. Polym.*, volume 17, pages 917–917. Springer New York, New York, NY.

11. He, T., Heidemeyer, M., Ban, F., Cherkasov, A., and Ester, M. (2017). SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, 9(1):1–14.

12. Hecker, N., Ahmed, J., von Eichborn, J., Dunkel, M., Macha, K., Eckert, A., Gilson, M. K., Bourne, P. E., and Preissner, R. (2012). SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.*, 40(D1):D1113–D1117.

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

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

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

16. Mahato, O. P. (2016). PDTI-DBN: PREDICTING DRUG AND TARGET INTERACTION USING DEEP BELIEF NETWORK. *Institute of Engineering/Pulchowk Campus*, 659:2016.

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

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

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

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

21. Sharma, V. K., Kumar, N., Prakash, T., and Taylor, T. D. (2010). MetaBioME: a database to explore commercially useful enzymes in metagenomic datasets. *Nucleic Acids Res.*, 38(suppl_1):D468–D472.

22. Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., and Aittokallio, T. (2014). 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.

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

24. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46(D1):D1074–D1082.

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

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

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

28. Zhang, L., Zhao, X., and Kong, L. (2014). 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.