A Deep Learning Approach for Drug-Target Affinity Prediction

by

Albina Li

B.S., Nazarbayev University (2019)

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

Apr 2021

Author .....
Department of Computer Science
Apr 27, 2021

Certified by .....
Siamac Fazli
Associate Professor
Thesis Supervisor

Certified by .....
Ferdinand Molnár
Associate Professor
Thesis Supervisor

Accepted by .....
Vassilios D. Tourassis
Dean, School of Science and Technology

A Deep Learning Approach for Drug-Target Affinity Prediction

by

Albina Li

Submitted to the Department of Computer Science
on Apr 27, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

The identification of drug-target interaction (DTI) is a crucial part of the drug discovery and development process. *In vitro* and *in vivo* experiments for drug target validation and screening are, however, very expensive and take a lot of time to complete. There experiment on large scale are unfeasible, thus there is a huge demand for the development of computational *in silico* alternatives for DTI prediction. Several statistical and machine learning-based methods have been developed over time that focused on the binary classification of DTI. However, these interactions are very complex, as there is a dynamic fluctuation present between the protein and the bound compound and a continuous mutually flexible adjustment, which needs to be simplified by reaching an equilibrium state characterised by well established binding affinity descriptor. The exact estimation of the binding affinity in the DTI still remains a challenge to this day. Various machine and deep learning methodologies have been developed that utilize different feature representation approaches for both compounds and proteins. These algorithms generally utilize as input limited chemical information, which may not be meaningful and intuitive enough to be used as an effective descriptor.

In this work I am addressing the limitation of current methods by introducing a deep learning-based model that makes use of chemical representations of the molecules. Results of experiments on two benchmark datasets demonstrate that the proposed model outperforms the baseline model, which is one of the state-of-the-art methods in the drug-target affinity (DTA) prediction field.

Thesis Supervisor: Siamac Fazli
Title: Associate Professor

Thesis Supervisor: Ferdinand Molnár
Title: Associate Professor

Acknowledgments

I would like to express my sincere gratitude to my thesis supervisors, Professor Siamac Fazli and Professor Ferdinand Molnár, for their invaluable guidance and continuous encouragement at every stage of the project. I am also thankful to my family and friends for their unconditional support in this challenging academic year.

Contents

Chapter 1

Introduction

The impaired activity of proteins in living organisms that may result in development of various diseases can be modulated by drugs resulting in an alteration of protein function which may lead to desirable therapeutic effects [27]. The discovery of drugs for protein targets is a highly complex process that requires a vast amount of temporal and financial resources. The development of a *de novo* drugs can cost up to 2.6 billion dollars [23], and it takes about 10-17 years for it to develop a marketable drug that is approved by the Food and Drug Administration (FDA) [2, 33]. Drug re-purposing, where approved drugs with established safety and efficacy are used for purposes they were not originally developed for, is therefore becoming a great alternative.

The crucial part of drug re-purposing is to identify how already established drugs may work on the target of interest. Traditional screening methods, either *in vivo* or *in vitro*, are conducted to learn the selectivity and efficacy of the interaction for the drug-target pairs [26]. This process is very expensive and time-consuming, thus it is not possible to screen extra large chemical libraries in chemical space with multiple targets. Therefore there is a necessity in the development of computational methods that use statistical and machine learning approaches to estimate the interaction strength between drug-target pairs and systematically identify promising candidate molecules as hit compounds. The development of such *in silico* methods is inevitable as they significantly facilitate the process of drug development, while reducing the accompanying costs and invested screening time.

Most of the previous studies approach the problem of the estimation of drug-target interaction (DTI) as a binary classification problem [4, 5, 6, 13, 21, 24, 29, 41, 46]. The models are trained to predict whether or not the compound would interact with the target, thus neglecting the important information of how strong this interaction is. The descriptor called binding affinity reflects this missing information, which can be expressed by various means such as a dissociation constant (K_d), inhibition constant (K_i) or half maximal inhibitory concentration (IC50). Only recently researchers started addressing this problem by developing machine learning and deep learning based regression models for the prediction of drug-target affinity (DTA) [1, 10, 16, 17, 25, 28, 30, 31, 36, 47].

1.1 Related Work

At first similarity-based methods that utilize conventional machine learning techniques called KronRLS [31] and SimBoost [16] were introduced. KronRLS is a Regularized Least Squares based algorithm that uses 2D similarity matrices for compounds and Smith-Waterman similarity [38] representation for targets. SimBoost uses a gradient boosting machine learning method to predict binding affinities and is trained on features engineered from similarity matrices of drug-target pairs. The major issue when using these methods is that the feature representation is limited by the similarity space. A novel molecule having low similarity with the molecules used in the training, will lead to inaccurate predictions when provided to a model. One possible way to overcome this problem is to use a wide variety of molecules that would cover the whole chemical space, but this is rather unrealistic, because the resources necessary to calculate similarity matrices limit the number of molecules that can be used in the training.

To mitigate the downside of similarity-based methods, a deep learning-based DTA prediction method called DeepDTA was developed [28]. The model works on the 1D representation of compounds and proteins. For representing compounds the simplified molecular input line entry system, otherwise known as SMILES is used. SMILES is a

line notation for describing the chemical structures of molecules using ASCII characters. It was developed by David Weininger [43] to represent molecules in computer-readable format. For target representation the amino acid sequence of proteins is used, which is the string of sequential amino acids from the N-terminal to C-terminal end of the protein molecule also known as protein primary structure. The underlying architecture of DeepDTA learns the abstract feature representation of drugs and targets by using convolutional neural networks (CNNs). Concatenated feature vectors are fed to a set of fully connected layers that are dedicated to predicting continuous values of binding affinity. DeepDTA outperforms both KronRLS and SimBoost on two benchmark datasets (Davis [7] and KIBA [39]), so the success of this deep learning-based method has sparked the interest of the scientific community and more variants have been developed since.

Some of the examples include CNN based feature representation methods, such as WideDTA [30], MT-DTI [36] and PADME [10] models. WideDTA represents the compounds and drugs as words, and it uses four different information sources which are drugs given in SMILES format, protein sequences, protein motifs and domains, and ligand maximum common substructures. MT-DTI uses the same input as DeepDTA, but introduces the alternative representation of molecules based on the self-attention mechanism. While WideDTA and MT-DTI let the CNN learn the protein representation, PADME uses fixed-rule descriptors to represent proteins. For compound representation SMILES are used.

Alternatives to CNN based models include GANsDTA [47] and DeepCDA [1]. Instead of using CNNs to learn feature representations, GANsDTA utilizes generative adversarial networks (GAN) whereas DeepCDA applies the integrated CNN and long-short-term-memory (LSTM) model to obtain representations for compounds and proteins. For the regression task of binding affinity prediction, the models utilize a multi-layer perceptron (MLP) similar to all previous studies.

State-of-the-art approaches in the field of DTA prediction are GraphDTA [25] and DGraphDTA [17]. These two methods have an underlying architecture that combines the graph neural networks with conventional CNNs. Unlike learning compound and

protein features from 1D representations, these models utilize structural information of molecules that are available in their graph representation. Experiments on benchmark Davis and KIBA datasets, both representing datasets used in the protein kinase drug-discovery field, showed significant predictive improvement over other methods on several performance measure metrics.

1.2 Motivation

Although the field of DTA prediction is relatively new, it attracts a lot of attention, as a lot of methods have been developed based on various input representations. However, most of these methods do not exploit chemical properties much as their default is to use only the 1D SMILES representation of drugs and protein sequence information. Although they work quite well, models that use additional qualitative chemical descriptors may enhance the prediction performance. GraphDTA [25] and DGraphDTA [17] for instance utilize the graph representation for compounds that contain atom properties and/or information about the chemical bonds and atom inter-connectivity within a compounds. Interestingly, WideDTA [30] integrates the functional information found in the protein sequences into its model. In particular it uses sequence motifs and profiles to construct the predictor in combination with traditionally used 1D representations. These models performed the best on extensive experiments using benchmark Davis and KIBA datasets, which supports the idea that the integration of chemical as well as functional data is beneficial for developing DTA predictors.

Apart from what has been already studied, there are still additional directions that are to be explored, and this is why this study is conducted. In this work, I am adopting the best practices in the field of DTA prediction, while integrating and providing novel auxiliary information such as chemical or functional descriptors. In particular, a novel model proposed in this work is the combination of CNN and a graph neural network. Multiple input representations are experimented with, which include graph-based, Coulomb matrix and molecular property representations of the compounds,

and categorical encoding, amino acid scale and domains/motifs representation for proteins. Numerous experiments are conducted and their performance on multiple metrics is compared against the baseline GraphDTA model [25] on two benchmark datasets, namely Davis and KIBA. Results show that the proposed model outperforms the baseline on all of the metrics, so it serves as proof that the supplementation of chemical and functional properties does indeed facilitate the performance of DTA predictors, and that this is a viable direction for future more applied research.

Chapter 2

Methodology

2.1 Datasets

In this work two datasets, namely Davis [7] and KIBA [39] are used for the development of a DTA prediction model. Both datasets are based on large-scale biochemical selectivity assays of kinase inhibitors, and are considered to be benchmark datasets for the evaluation of the binding affinity prediction.

The Davis dataset consists of 442 sequence entries from the kinase protein family, and 68 compounds, which overall constitute 30,056 DTI expressed by dissociation constant (K_d). As suggested by [16] K_d values are transformed into log space, pK_d as shown in equation 2.1. This is done for ensuring the numerical stability.

$$pK_d = -log_{10} + (\frac{K_d}{1e9}) \tag{2.1}$$

The original KIBA dataset contains 467 kinase protein family targets and 52,498 inhibitors. In total there are 246,088 affinity values that combine different sources, such as K_i, K_d, and $IC50$. Tang et al. introduced a novel bioactivity score called KIBA, that statistically combines all known values to normalize the mutual consistency between them [39]. This dataset was later updated to only include compounds and targets that have at least 10 interactions, thus resulting in 229 proteins and 2,111 drugs with 118,254 interactions [16]. The latter updated version is used in this work.

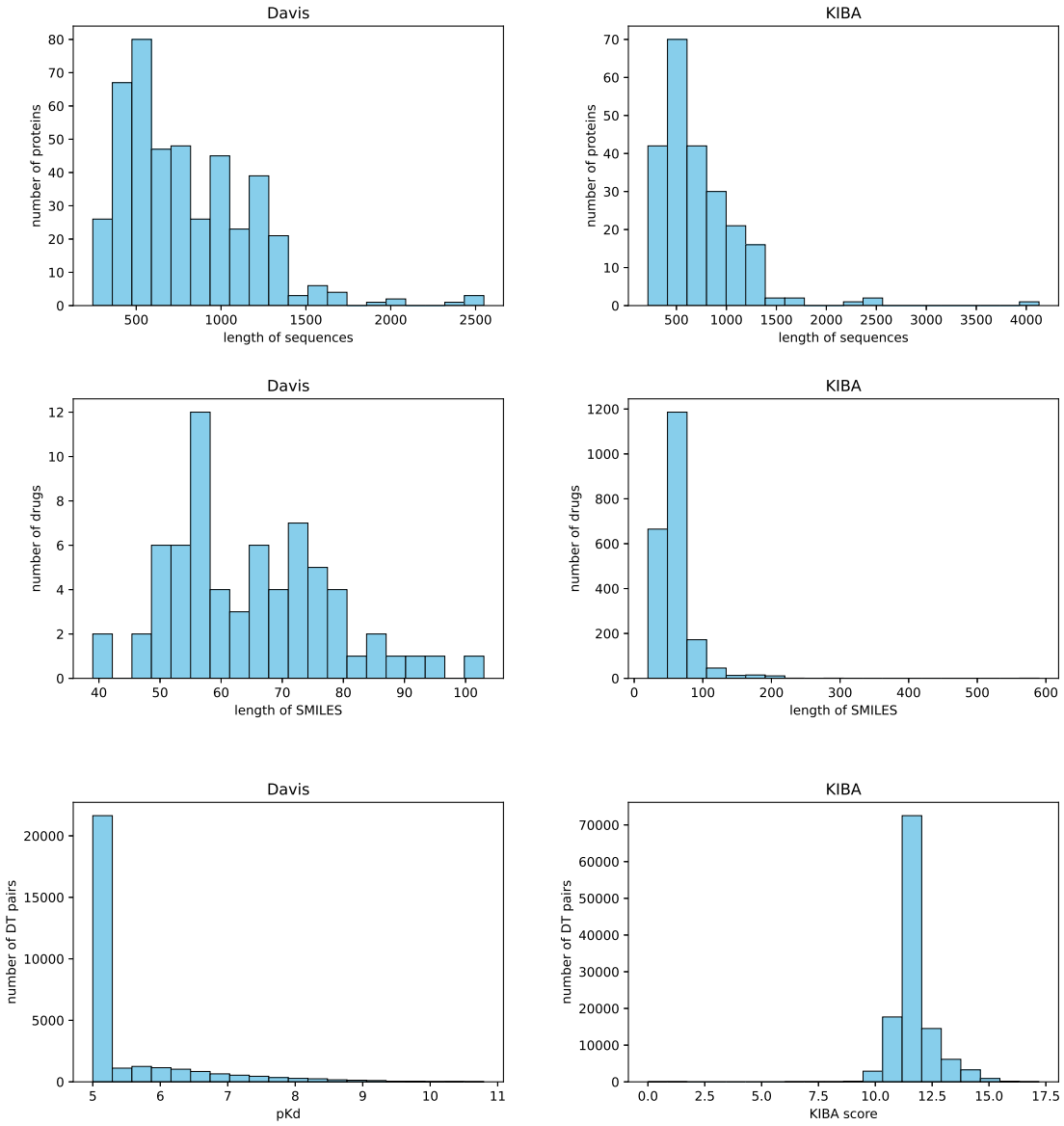

Figure 2-1: Summary of the Davis dataset (left panel) and KIBA (right panel) dataset. First row shows the distribution of the lengths of the protein sequences. Second row shows the distribution of the lengths of the SMILES strings. Third row represents the distribution of binding affinity values.

The number of compounds, proteins and interactions of both datasets are summarized in Table 2.1. Figure 2-1 gives more insight into all the components for Davis (left panel) and KIBA (right panel). First row represents the sequence lengths distribution for all proteins. The minimum length of protein sequences for Davis is 244, while the maximum and average are 2549 and 788, respectively. For KIBA the

minimum, maximum and average lengths of protein sequences are 215, 4128, and 728 characters. Second row shows the distribution for compound lengths in SMILES format. Minimum, maximum and average drug lengths for Davis are 103, 39, and 64, while for KIBA it is 590, 20, and 59 characters.

Dataset	Compounds	Proteins	Interactions
Davis	68	442	30 056
KIBA	2111	229	118 254

Table 2.1: Summary of datasets

Third row shows the distribution of binding affinity values in pK_d format for Davis and KIBA score format for KIBA. For Davis dataset lower pK_d values indicate lower binding affinity. A strong spike can be observed at pK_d value of 5, with more than half of the whole dataset (69%) belonging to it. These values correspond to "negative pairs" which either have a weak binding affinities or the interactions are not observed in the primary screen [31]. On the contrary for the KIBA dataset, the lower the KIBA score, the higher the binding affinity between drug-target pairs. The suggested threshold that separates positive and negative interaction values in KIBA is 12.1 in terms of KIBA score [39]. Distribution of scores reveals that most of the pairs (80%) are positive for this dataset.

2.2 Baseline Model

GraphDTA [25] is taken into consideration as state-of-the-art baseline model. The architecture introduced in this work is based on the combination of CNN and graph neural networks. The model takes the drug-protein pair as input and processes them in parallel to get representation vectors. These two vectors are then concatenated and forwarded to dense layers with the finishing regression layer that gives the prediction for their affinity value.

GraphDTA uses a sequence representation for the proteins, a string of ASCII

letters where each character corresponds to an amino acid. The sequence is first categorically encoded, then it undergoes the embedding layer, and several 1D CNNs that learn the feature representation. For the compounds the method utilizes the graph representation obtained by transforming the SMILES input. The graph neural network is then applied to obtain the latent feature vector. This approach captures the structural information of the drugs, which is lost if a conventional 1D SMILES representation would be used. The overview of the model architecture is illustrated in Figure 2-2.

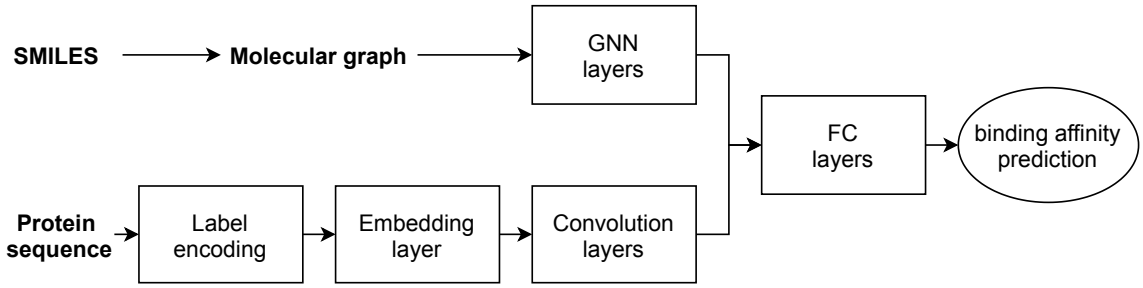

Figure 2-2: Baseline model architecture.

2.3 Input Representation

The purpose of this work is to explore how the integration of chemical and functional information into the development of DTA prediction models affects their performance. Multiple input representations are tested out, which include experiments on various compound and protein representations, as well as on their chemical and functional properties.

2.3.1 Molecule Representation

Molecular graph

Following the best practices in the field of DTA prediction, this work utilizes the graph representation for the compounds. SMILES strings are converted into a molec-

ular graph of interactions between atoms, which conserves the important structural information of the chemical structure. In particular, the compound preprocessing pipeline of [25] is adopted. Nguyen et al. are using the atom feature design by DeepChem [32]. Each node in the molecular graph is described by five atomic properties: atom symbol, atom degree - number of bonded neighbors plus number of Hydrogen atoms, total number of Hydrogen atoms, implicit value of atom, and aromaticity of atom [32]. Multi-dimensional feature vectors are constructed for each atom, and the edges are added between any pair of atoms if there exists a bond between them. This constitutes the final molecular graph. All of the computations are performed using RDKit, which is an open-source collection of cheminformatics and machine learning software tools [20].

Coulomb matrix

A different approach for compound representation is the Coulomb matrix descriptor introduced by Rupp et al. [35]. The Coulomb matrix features the approximation of the electrostatic interaction between nuclei inside the molecule [35]. It requires a set of nuclear charges Z_i and corresponding Cartesian coordinates R_i, which are used as follows to compute the matrix entries for any given molecule:

$$M_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \forall\ I\ =\ J \\ \frac{Z_i Z_j}{|R_i R_j|} & \forall\ I\ \neq\ J \end{cases} \tag{2.2}$$

The diagonal entries of the Coulomb matrix correspond to a polynomial fit of the potential energies of isolated atoms, while the off-diagonal entries encode the Coulomb repulsion between different pairs of nuclei in the molecule [9]. The matrix is therefore invariant to translations and rotations of the molecule. It is however not invariant under random atom permutations. There are several approaches that tackle this problem, one of which is the use of sorted Coulomb matrices. The idea is to order the rows of the matrix, such that $||M_i|| \leq ||M_{i+1}||$ for any given row i. This ensures that two different molecules necessarily have different Coulomb matrices [22].

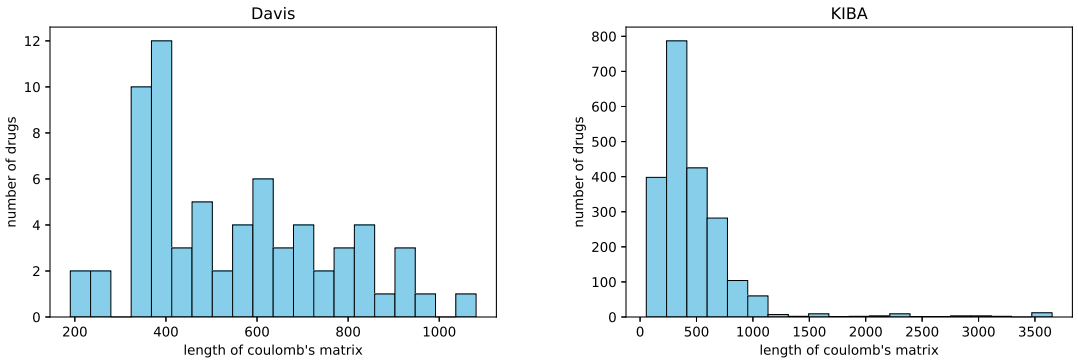

Figure 2-3: Length of Coulomb's matrix of compounds for Davis (left panel) and KIBA (right panel) datasets.

The sorted Coulomb matrix method is used in this work via the implementation by ChemML machine learning and cheminformatics program [15]. The resulting distribution of the lengths of Coulomb matrices can be observed in Figure 2-3 for both the Davis (left panel) and the KIBA (right panel) dataset. It can be clearly seen that for most of the compounds the length of the resulting Coulomb matrix representation falls into the range of up to 1000 molecules. More precisely 98% for Davis and 96% for KIBA datasets. It was therefore decided to create a fixed length Coulomb matrix representation of size 1000, with larger feature vectors truncated, whereas smaller feature vectors 0-padded.

Other molecular descriptors

A molecular descriptor is "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment" [40]. In this study additional molecular descriptors were utilized, calculated using RDKit, a toolkit widely used in the cheminformatics research community. All together, 33 various descriptors such as exact molecular weight, number of aromatic rings, number of rotatable bonds were calculated, thus constituting fixed-size feature vectors for each of the compounds. Table A.1 presents all utilized molecular descriptors.

2.3.2 Protein Representation

Integer/label encoding

One of the straightforward, machine learning friendly ways of representing a protein is to encode each amino acid in a sequence with a distinct integer. A thorough analysis was done in a study by Öztürk et al. [28], where authors scanned a large quantity of protein sequences and extracted 25 unique letters. Each of these letters is mapped into a corresponding integer, e.g. {'A':1, 'B':2, 'C':3, ... , 'Z':25} when encoding is implemented. Protein sequence 'MTVKTEA...' is, for instance, transformed into the following integer vector:

$$[M\ T\ V\ K\ T\ ...]\ =\ [12\ 19\ 21\ 10\ 19\ ...]$$

Protein sequences have varying length, but vectors of fixed length are preferred for optimal feature representation. Based on the distribution of protein sequence lengths, depicted in Figure 2-1 (first row), it was decided to use 1000 characters as a cutoff, since 73% and 80% of protein sequences have a length of less than 1000 characters for Davis and KIBA respectively. To obtain fixed vector length, longer sequences were truncated and shorter sequences were padded with 0.

Amino acid scale representation

A possible downside to the integer/label encoding is that the mapping is user-defined and it may not carry any meaning. There is, however, an alternative way of implementing the encoding with actual chemical properties, which is to use the amino acid scale representation. An amino acid scale is defined by a numerical value assigned to each type of amino acid. There are numerous amino acid scales which are based on different chemico-physical properties of the amino acids.

In this work I am using 11 various scales provided by the ProtScale web tool [12]: Average Flexibility, Bulkiness, Hydropathicity, Molecular Weight, Number Of Codons, Polarity, Recognition Factors, Refractivity, Relative Mutability, Retention Coefficient trifluoroacetic, and Transmembrane Tendency. As with the integer/label encoding

each amino acid in the sequence is replaced with the corresponding value defined in the scales, constituting a multi-dimensional array. Same as with integer/label encoding to keep things uniform, the length of each encoding is fit to the length of 1000, thus resulting in a representation of a size (11, 1000) for each protein.

Motifs and domains

Another piece of information that can be obtained from the protein is the protein sequence motifs and domains. These are the specific regions within the protein sequence, which may be important for folding, binding, catalytic activity and thermodynamics [30]. A protein motif is a continuous short amino-acid sequence pattern shared by similar proteins, which may be defined by a unique biological or chemical function [3]. A protein domain is a larger element of the protein's structure that often folds independently of the remaining protein chain and retains its functional properties independently from the full protein [3]. Both the motifs and domains are obtained and tested out in this work via the ScanProsite web tool [8], that scans the proteins against a large-scale database of motifs and profiles descriptors (PROSITE [37]).

For each of the proteins two lists of available motifs and domains are obtained using the aforementioned tool. All obtained motifs and domains are then used to extract three-residue "words" from the sequences. For instance, the given motif 'IGKGSFGKVLLARHKAEEVFYAVKVLQKKAILK' is represented as a set of three letter "words" 'IGK', 'GKG', 'KGS', ..., 'KAI', 'AIL', 'ILK'. Separate analysis on the datasets showed that for Davis there are 3449 unique three-residues for motifs, and 7851 unique three-residues for domains, whereas for KIBA there are 2859 and 7648 respectively. These residues are then categorically encoded and motifs and domains are represented as vectors of integers.

Figure 2-4 depicts the distribution of the number of three-residues per motif (first row) and per domain (second row). For both Davis (left panel) and KIBA (right panel) datasets most of the proteins (around 80-90%) have less than 60 residues per motif and less than 650 residues per domain. These numbers are therefore used as a cutoff to ensure the fixed length for the protein representation.

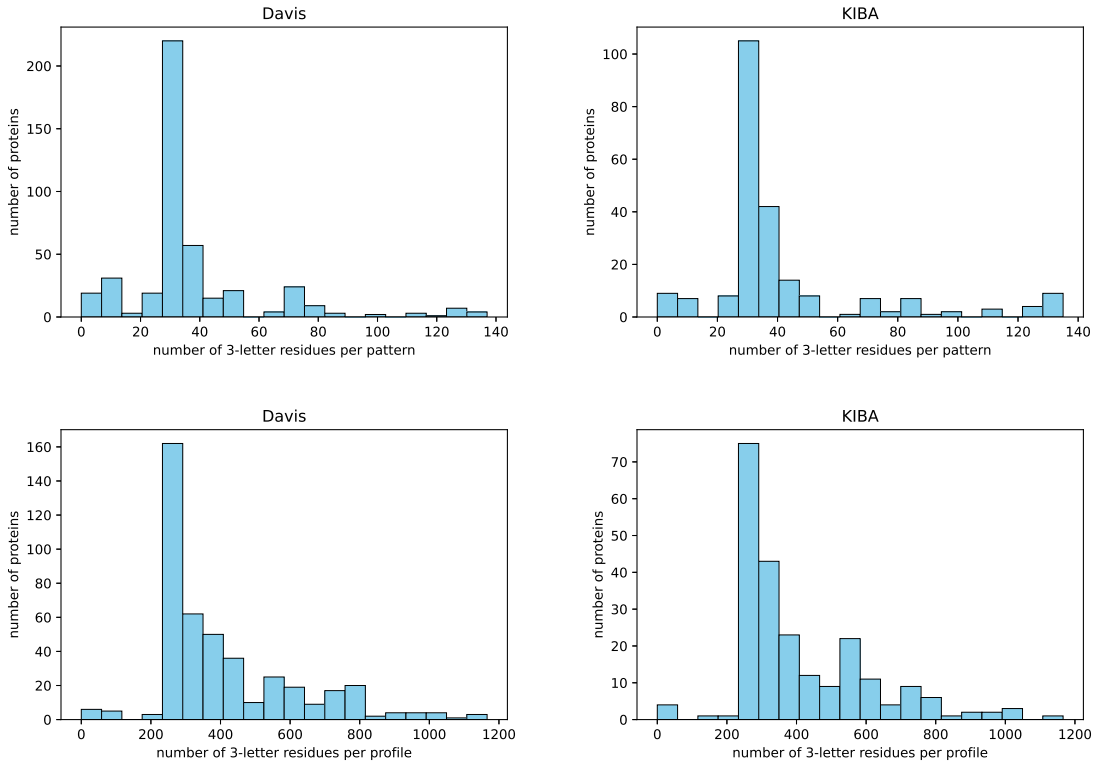

Figure 2-4: Number of 3-letter residues per motif (first row) and per domain (second row) for Davis (left panel) and KIBA (right panel) datasets.

2.4 Proposed Model

The proposed model adopts a simplified version of the framework used in the GraphDTA method [25], which is the combination of a conventional CNN and a graph neural network. The summary of this model is depicted in Figure 2-5. The left portion shows all drug representations and the right portion is dedicated to all target representations. Incoming dashed lines into 'combined representation' part mean that the choice of representation vectors is optional and depends on the conducted experiment.

While experimenting with several input representations, the model differs throughout the trials, but the underlying architecture stays the same. The model takes a compound-protein pair and feed-forwards them in parallel to learn feature representation vectors. All latent vectors are then concatenated and sent to two consecutive fully connected layers, which are completed by a regression layer that returns the binding affinity prediction.

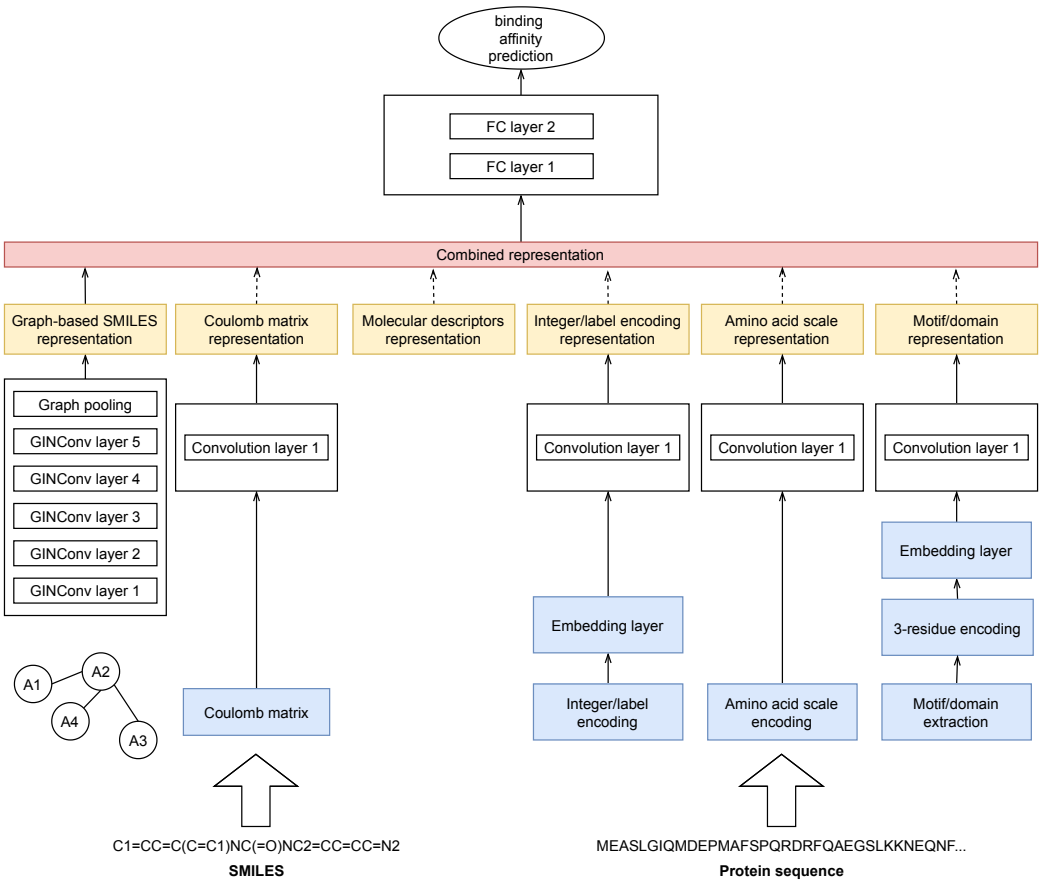

Figure 2-5: Proposed model architecture.

Three representation options for a compound are graph-based, Coulomb matrix, and molecular descriptors. To obtain a graph-based representation, the model converts a SMILES input into a molecular graph (as described in the previous section) and applies a graph algorithm to learn the feature vector. There are a number of graph neural network models, such as graph convolutional network (GCN) [19], graph attention network (GAT [42]), and graph isomorphism network (GIN) [45]. According to the work by Nguyen et al. [25] the best performance is exhibited by applying GIN that uses a multi-layer perceptron to update node features, so it was chosen for the proposed model. This network consists of five GIN layers with batch normalization in between, followed by global max pooling, and a fully connected layer.

A similar procedure is performed on the Coulomb matrix representation. Once the sorted Coulomb matrix is calculated, the model applies a simple 1D convolution layer to learn the features, and a fully connected layer to get the representation

vector. As for the last option, a vector of molecular descriptors is not processed by CNNs, as there is not enough data for it to be effective, so a raw molecular descriptor representation is used.

Protein sequence representations explored in this work are integer/label encoding, amino acid scale, and motif/domain representations. Given the protein sequence input, first the model does the encoding (described in detail in Section 2.3.2). For textual data (categorical amino acid and motif/domain three-residue encodings) the embedding layer is added that represents each token by a 128-dimensional vector. A 1D convolution layer is then applied to learn features, and the model processes the representation with a fully connected layer.

2.5 Performance Evaluation Metrics

For comparison of the performance between the proposed and the baseline model are four commonly used evaluation metrics for regression task, namely mean squared error (MSE), concordance index (CI) [14], Pearson correlation coefficient, and squared correlation coefficient r_m^2 [34].

Mean squared error

MSE is a metric that evaluates the difference between the true value and the predicted one. Given n samples, MSE is defined as the average of the sum of the squared difference between predicted p_i and true t_i values, where $i \in [1, n]$. Smaller MSE values indicate that the prediction is close to the ground truth values:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (p_i - t_i)^2 \tag{2.3}$$

Concordance index

Another common metric used for evaluation of DTA predictions is the concordance index (CI). It is a metric used to measure the probability of two predicted affinity values appearing in the same order as their real values. Drawing two random distinct

samples b_i and b_j, where $b_i > b_j$, the values are in correct order if their corresponding true affinity scores δ_i and δ_j satisfy the condition $\delta_i > \delta_j$:

$$CI = \frac{1}{Z} \sum_{\delta_i > \delta_j} \eta(b_i - b_j) \tag{2.4}$$

Z is a normalization constant, $\eta(x)$ is a step function:

$$\eta(x) = \begin{cases} 1, & x > 0 \\ 0.5, & x = 0 \\ 0, & else \end{cases} \tag{2.5}$$

CI values range between 0.5 and 1.0, where 0.5 is a result achieved by a random predictor, and 1.0 corresponds to the perfect predictive performance.

Correlation coefficients

Some studies also calculate correlation coefficients between predicted (p) and true (t) affinity values, which measure the strength and direction of a relationship between them. Most commonly used Pearson correlation coefficient is defined as follows:

$$r_{p,t} = \frac{cov(p,\ t)}{\sigma(p)\ \sigma(t)} \tag{2.6}$$

The coefficient ranges between -1 and 1, where -1 indicates a strong negative relationship between values, 1 indicates a strong positive relationship, and 0 indicates that there is no linear relationship at all.

Additionally a squared correlation coefficient r_m^2 is used for evaluation, which is calculated using the following formula, where r is a correlation coefficient with intercept, and r_0 without intercept:

$$r_m^2 = r^2\ (1 - \sqrt{r^2\ -\ r_0^2}) \tag{2.7}$$

The performance of the predictor is considered acceptable, if it achieves an r_m^2 value of at least 0.5.

Chapter 3

Experiments and Results

3.1 Experimental Setup

The performance of the proposed model is evaluated against the performance of a baseline GraphDTA model on the Davis and KIBA datasets using all performance evaluation metrics described in Section 2.5, namely MSE, CI, Pearson correlation coefficient and r_m^2. The data for all experiments is split in the exact same way for both datasets, where 68% of data is dedicated for training, 16% for model evaluation, and 16% is kept as a holdout set for testing. Davis has 20,037/5,009/5,010 samples for train/validation/test split, whereas KIBA has 78,836/19,709/19,709 samples correspondingly.

The proposed model was implemented using Pytorch geometric [11]. The training was set for 100 epochs with a batch size 512 for the weight update. Adam optimization algorithm [18] was utilized with a constant learning rate of 0.0005. To ensure that the model is not overfitted, dropout was applied throughout the network. Another overfitting prevention was the implementation of early stopping with a patience of 15 epochs - if the MSE score did not improve on the validation set for more than 15 epochs the training stopped and the last best performing model state was saved.

Besides the experiment on the baseline model, overall 7 experiments with different input combinations for the proposed model were conducted for each dataset:

- **[AS]** Graph-based compound representation + amino acid scale encoding protein representation
- **[CM]** Graph-based and Coulomb matrix compound representation + integer/label encoding protein representation
- **[MD]** Graph-based and molecular descriptors compound representation + integer/label encoding protein representation
- **[CM-MD]** Graph-based, Coulomb matrix and molecular descriptors compound representation + integer/label encoding protein representation
- **[DM]** Graph-based compound representation + integer/label encoding and domains protein representation
- **[MT]** Graph-based compound representation + integer/label encoding and motifs protein representation
- **[CM-MT]** Graph-based and Coulomb matrix compound representation + integer/label encoding and motifs protein representation

3.2 Results

Table 3.1 and Table 3.2 show all resulting performances based on MSE, CI, Pearson correlation coefficient and r_m^2 metrics on test set of Davis and KIBA datasets. The first row in each table represents the performance of the baseline GraphDTA model. The rest of the table gives the results for all input combinations of the proposed model described in Section 2.3.

The data from the tables shows that the proposed model outperforms the baseline in each of the performance metrics. The largest improvement on prediction performance in terms of MSE and CI was achieved by applying the combination of the graph-based and Coulomb matrix compound representation, and label/integer encoding and motifs protein representation [CM-MT]. For the Davis dataset, the MSE

	Compound representation	**Protein representation**	**MSE**	**CI**	**Pearson**	**rm^2**
GraphDTA	graph-based	integer encoding	*0.3*	*0.793*	*0.859*	*0.596*
AS	graph-based	amino acid scale	0.533	0.592	0.796	0.317
CM	graph-based Coulomb matrix	integer encoding	0.283	0.81	0.875	0.604
MD	graph-based molecular descriptors	integer encoding	0.345	0.759	0.849	0.543
CM-MD	graph-based Coulomb matrix molecular descriptors	integer encoding	0.338	0.761	0.857	0.574
DM	graph-based	integer encoding domains	0.281	0.812	0.873	**0.658**
MT	graph-based	integer encoding motifs	0.269	0.821	**0.88**	0.629
CM-MT	graph-based Coulomb matrix	integer encoding motifs	**0.264**	**0.822**	0.869	0.637

Table 3.1: MSE, CI, Pearson correlation coefficient, and r_m^2 scores of the test set of Davis dataset

	Compound representation	**Protein representation**	**MSE**	**CI**	**Pearson**	**rm^2**
GraphDTA	graph-based	integer encoding	*0.265*	*0.803*	0.*781*	*0.598*
AS	graph-based	amino acid scale	0.325	0.775	0.726	0.503
CM	graph-based Coulomb matrix	integer encoding	0.241	0.819	0.807	0.603
MD	graph-based molecular descriptors	integer encoding	0.287	0.794	0.765	0.561
CM-MD	graph-based Coulomb matrix molecular descriptors	integer encoding	0.255	0.809	0.794	0.6
DM	graph-based	integer encoding domains	0.254	0.811	0.792	0.609
MT	graph-based	integer encoding motifs	0.243	0.812	0.802	0.624
CM-MT	graph-based Coulomb matrix	integer encoding motifs	**0.227**	**0.823**	**0.822**	**0.655**

Table 3.2: MSE, CI, Pearson correlation coefficient, and r_m^2 scores of the test set of KIBA dataset

score decreased from *0.3* (baseline model) to *0.264*, while the CI score increased from *0.793* to *0.822*. As for KIBA, MSE dropped from *0.265* to *0.227* and CI grew from *0.803* to *0.823*. Pearson and r_m^2 scores are also improved - from *0.859* to *0.869* and from *0.596* to *0.637* for Davis; from *0.781* to *0.0.822* and from *0.598* to *0.655* for KIBA.

Other combinations that performed comparably well are inputs including Coulomb matrix representation for compounds, and domains/motifs representation for proteins (CM, CM-MD, DM, and MT trials). Most of the scores superseded the baseline performance, which can be observed in Tables 3.1 and 3.2.

Figure 3-1 illustrates the scatter plot of the predicted binding affinity values for the best performing configuration against the ground truth values for Davis (left panel) and KIBA (right panel) datasets. A perfect model performance is exhibited if a plot results in a straight line $p = t$, where p is a prediction and t is a true value. It can be observed that the density around the perfect prediction line is quite high, especially for KIBA dataset, and that there are very few samples where the prediction significantly differs from the actual binding affinity value.

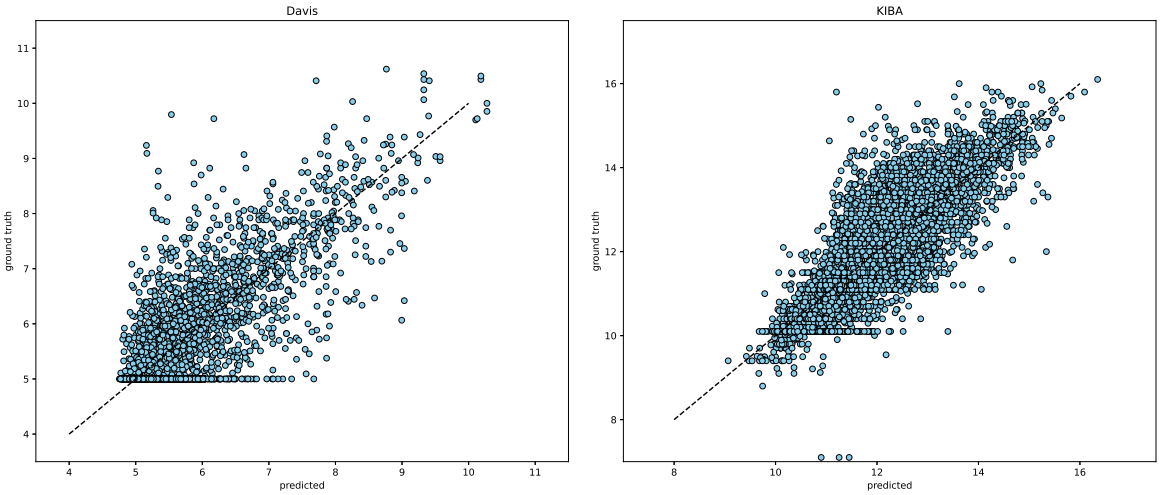

Figure 3-1: Predicted against ground truth values for Davis (left panel) and KIBA (right panel) datasets.

3.3 Discussion

The results of multiple experiments using the two large-scale kinase inhibitor datasets, Davis and KIBA, have proven that integration of chemical and functional information for compounds and proteins can improve DTA predictions. For instance, adding a Coulomb matrix to the compound representation, as well as motifs and domains to protein representation has shown to enhance the performance of the model (Tables 3.1 and 3.2). However, the use of the remaining representations did not fulfil the expectations.

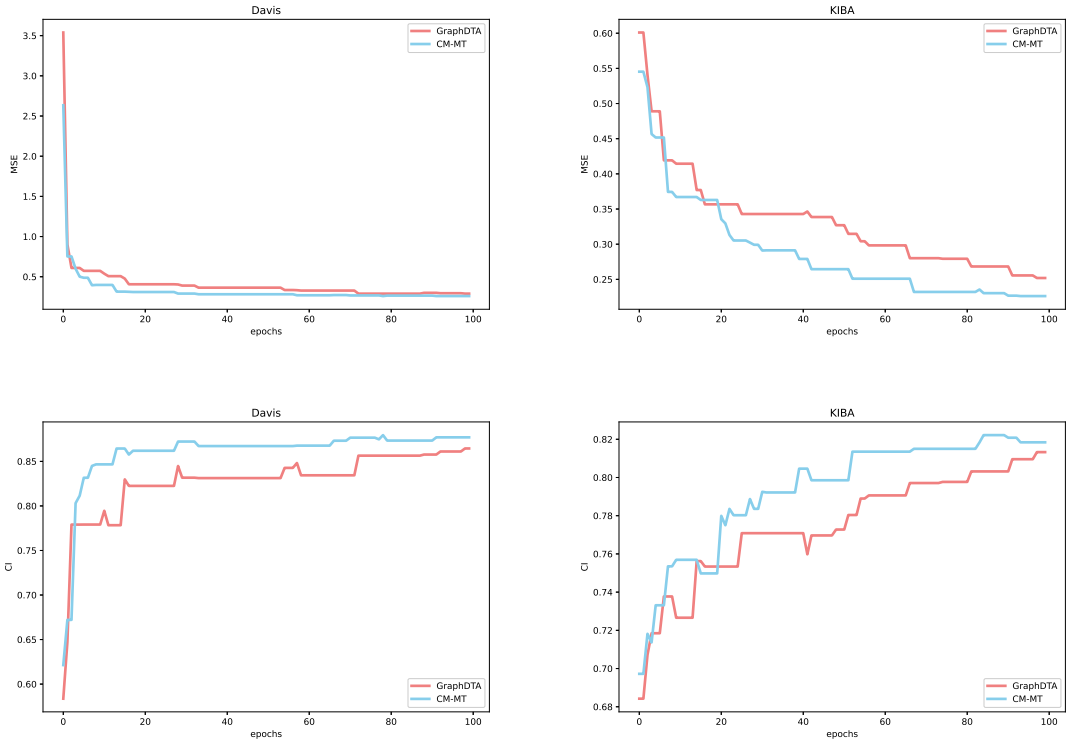

Figure 3-2: MSE (first row) and CI (second row) trends over 100 epochs for Davis (left panel) and KIBA (right panel) datasets.

Both molecular descriptors based compound representation and amino acid scale protein representation performed considerably worse, compared even to the baseline model. The main reason for such a failure might be the sub-optimal model used for input representation. In case of molecular descriptors for instance, no convolution layer based feature representation was introduced, so the feature vector is used in its

raw state. As for the amino acid scale representation, a 1D convolution layer was applied to (11,1000) shaped input - so the current model may not be complex enough to effectively learn feature representations.

A more thorough analysis of the performance evaluation over the whole period of training confirmed the limitation of the current model on input representation. Figure 3-2 illustrates the trend of change in MSE (first row) and CI (second row) of baseline and best performing CM-MT model on test set across 100 training epochs for Davis (left panel) and KIBA (right panel) datasets. Although the proposed model performs better than the baseline, towards the end of training the MSE and CI values for both models seem to converge. Such behavior indicates that after some time the proposed model starts to overfit. It is more noticeable for Davis, as there are significantly less training samples. The overfitting happens because the proposed model has to operate on twice the input data compared to the baseline model, while having the underlying architecture of the same complexity. Designing a more fitting architecture should solve the problem of overfitting, and this task will be addressed in the future work.

Chapter 4

Conclusion

In this study a machine learning model for DTA prediction was proposed and tested, which integrates graph neural networks with traditional CNNs in its underlying architecture. The model works on various input representations, which include conventional 1D string representations of compounds in SMILES format and protein sequences, as well as representations obtained from their chemical and functional properties. In particular, for compounds the sorted Coulomb matrix and molecular descriptors were applied, while for proteins amino acid scales and protein motifs and domains were tested out.

Performance of the proposed model was examined on two large-scale benchmark datasets based on biochemical selectivity assays for kinase inhibitors, namely Davis and KIBA, and compared to the baseline GraphDTA model. The proposed model superseded the predictive performance of the baseline on four evaluation metrics (MSE, CI, Pearson correlation coefficient and r_m^2). The best input configuration was a combination of graph-based and Coulomb matrix representations for compounds, and categorical encoding with motif representation for proteins. Compared to the baseline, MSE improved from *0.3* to *0.264* for Davis dataset and from *0.265* to *0.227* for KIBA. As for CI the values improved from *0.793* to *0.822* and from *0.803* to *0.823* for Davis and KIBA, respectively. Other successful input configurations included the combinations with a Coulomb matrix compound representation and motif/domain representation of proteins.

Results demonstrate the practical advantage of integrating additional chemical and functional information into the development of DTA predictions. However, a closer look into the performance evaluation metrics revealed that the underlying architecture for feature representation is sub-optimal, as it overfits the model once it is trained for a longer period. Optimizing the network would likely solve this problem, so this will be addressed in a future work alongside with other possible feature representation improvements.

Appendix A

Molecular descriptors

RDKit descriptor	Description
ExactMolWt	the exact molecular weight of the molecule
MolWt	the average molecular weight of the molecule
HeavyAtomMolWt	the average molecular weight of the molecule without the hydrogens
FpDensityMorgan1-3	topological fingerprints for molecular characterization and extended connectivity
MaxAbsPartialCharge MaxPartialCharge MinAbsPartialCharge MinPartialCharge	partial charges for molecules descriptors
NumRadicalElectrons NumValenceElectrons	the number of radical/valence electrons the molecule has
FractionCSP3	the fraction of C atoms that are SP3 hybridized
HeavyAtomCount NHOHCount NOCount	the number of heavy atoms/NHs or OHs/Nitrogens and Oxygens in a molecule
NumAliphaticCarbocycles NumAliphaticHeterocycles NumAliphaticRings	the number of aliphatic (containing at least one non-aromatic bond) carbocycles/heterocycles/rings for a molecule
NumAromaticCarbocycles NumAromaticHeterocycles NumAromaticRings	the number of aromatic carbocycles/heterocycles/rings for a molecule
NumSaturatedCarbocycles NumSaturatedHeterocycles NumSaturatedRings	the number of saturated carbocycles/heterocycles/rings for a molecule
NumHAcceptors NumHDonors	the number of Hydrogen bond acceptors/donors
NumHeteroatoms	the number of heteroatoms
NumRotatableBonds	the number of rotatable bonds
RingCount	the number of rings for a molecule
MolLogP MolMR	the Wildman-Crippen LogP/MR value [44]
TPSA	the topological polar surface area

Table A.1: RDKit molecular descriptors used in the proposed model

Bibliography

[1] Karim Abbasi, Parvin Razzaghi, Antti Poso, Massoud Amanlou, Jahan B Ghasemi, and Ali Masoudi-Nejad. DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics*, 36(17):4633–4642, 2020.

[2] Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673–683, 2004.

[3] Gerald Bergtrom. Protein Domains, Motifs, and Folds in Protein Structure.

[4] Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, 2009.

[5] Dong-Sheng Cao, Liu-Xia Zhang, Gui-Shan Tan, Zheng Xiang, Wen-Bin Zeng, Qing-Song Xu, and Alex F Chen. Computational prediction of drug target interactions using chemical, biological, and network features. *Molecular informatics*, 33(10):669–681, 2014.

[6] Murat Can Cobanoglu, Chang Liu, Feizhuo Hu, Zoltán N Oltvai, and Ivet Bahar. Predicting drug–target interactions using probabilistic matrix factorization. *Journal of chemical information and modeling*, 53(12):3399–3409, 2013.

[7] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

[8] Edouard De Castro, Christian JA Sigrist, Alexandre Gattiker, Virginie Bulliard, Petra S Langendijk-Genevaux, Elisabeth Gasteiger, Amos Bairoch, and Nicolas Hulo. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic acids research*, 34(suppl_2):W362–W365, 2006.

[9] Daniel C Elton, Zois Boukouvalas, Mark S Butrico, Mark D Fuge, and Peter W Chung. Applying machine learning techniques to predict the properties of energetic materials. *Scientific reports*, 8(1):1–12, 2018.

[10] Qingyuan Feng, Evgenia Dueva, Artem Cherkasov, and Martin Ester. Padme: A deep learning-based framework for drug-target interaction prediction. *arXiv preprint arXiv:1807.09741*, 2018.

[11] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[12] Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, Marc R Wilkins, Ron D Appel, Amos Bairoch, et al. Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*, pages 571–607, 2005.

[13] Mehmet Gönen. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, 28(18):2304–2310, 2012.

[14] Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

[15] Mojtaba Haghighatlari, Gaurav Vishwakarma, Doaa Altarawy, Ramachandran Subramanian, Bhargava U Kota, Aditya Sonpal, Srirangaraj Setlur, and Johannes Hachmann. ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(4):e1458, 2020.

[16] Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, 9(1):1–14, 2017.

[17] Mingjian Jiang, Zhen Li, Shugang Zhang, Shuang Wang, Xiaofeng Wang, Qing Yuan, and Zhiqiang Wei. Drug–target affinity prediction using graph neural network and contact maps. *RSC Advances*, 10(35):20701–20712, 2020.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[20] Greg Landrum. RDKit: Open-source cheminformatics.

[21] Yong Liu, Min Wu, Chunyan Miao, Peilin Zhao, and Xiao-Li Li. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS computational biology*, 12(2):e1004760, 2016.

[22] Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. *Advances in neural information processing systems*, 25:440–448, 2012.

[23] Asher Mullard. New drugs cost US $2.6 billion to develop. *Nature Reviews Drug Discovery*, 13(12):877, 2014.

[24] André CA Nascimento, Ricardo BC Prudêncio, and Ivan G Costa. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics*, 17(1):1–16, 2016.

[25] Thin Nguyen, Hang Le, and Svetha Venkatesh. GraphDTA: prediction of drug–target binding affinity using graph convolutional networks. *BioRxiv*, page 684662, 2019.

[26] Martin EM Noble, Jane A Endicott, and Louise N Johnson. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303(5665):1800–1805, 2004.

[27] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature reviews Drug discovery*, 5(12):993–996, 2006.

[28] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[29] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. *BMC bioinformatics*, 17(1):1–11, 2016.

[30] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. WideDTA: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*, 2019.

[31] Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16(2):325–337, 2015.

[32] Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. " O'Reilly Media, Inc.", 2019.

[33] Allen D Roses. Pharmacogenetics in drug discovery and development: a translational perspective. *Nature reviews Drug discovery*, 7(10):807–817, 2008.

[34] Kunal Roy, Pratim Chakraborty, Indrani Mitra, Probir Kumar Ojha, Supratik Kar, and Rudra Narayan Das. Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. *Journal of computational chemistry*, 34(12):1071–1082, 2013.

[35] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[36] Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C Ho. Self-attention based molecule representation for predicting drug-target interaction. In *Machine Learning for Healthcare Conference*, pages 230–248. PMLR, 2019.

[37] Christian JA Sigrist, Edouard De Castro, Lorenzo Cerutti, Béatrice A Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. New and continuing developments at PROSITE. *Nucleic acids research*, 41(D1):D344–D347, 2012.

[38] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[39] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero 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, 2014.

[40] Roberto Todeschini and Viviana Consonni. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*, volume 41. John Wiley & Sons, 2009.

[41] Twan van Laarhoven, Sander B Nabuurs, and Elena Marchiori. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, 27(21):3036–3043, 2011.

[42] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[43] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[44] Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

[45] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[46] Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–i254, 2010.

[47] Lingling Zhao, Junjie Wang, Long Pang, Yang Liu, and Jun Zhang. GANs-DTA: predicting drug-target binding affinity using GANs. *Frontiers in genetics*, 10:1243, 2020.