Interdisciplinary Sciences: Computational Life Sciences (2023) 15:306–315
https://doi.org/10.1007/s12539-023-00557-z

ORIGINAL RESEARCH ARTICLE

Deep Learning-Based Modeling of Drug–Target Interaction Prediction Incorporating Binding Site Information of Proteins

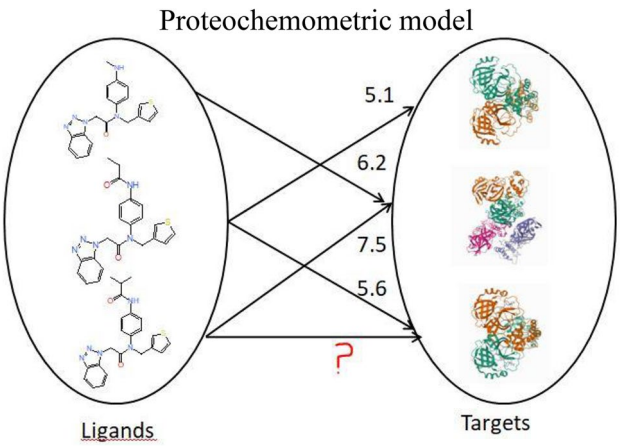
Sofia D'Souza[1] · K. V. Prema[2] · S. Balaji[3] · Ronak Shah[1]

Received: 24 February 2022 / Revised: 22 February 2023 / Accepted: 22 February 2023 / Published online: 26 March 2023

Abstract

Chemogenomics, also known as proteochemometrics, covers various computational methods for predicting interactions between related drugs and targets on large-scale data. Chemogenomics is used in the early stages of drug discovery to predict the off-target effects of proteins against therapeutic candidates. This study aims to predict unknown ligand–target interactions using one-dimensional SMILES as inputs for ligands and binding site residues for proteins in a computationally efficient manner. We first formulate a Deep learning CNN model using one-dimensional SMILES for drugs and motif-rich binding pocket subsequences of proteins as inputs. We evaluate and compare the proposed deep learning model trained on expert-based features against shallow feature-based machine learning methods. The proposed method achieved better or similar performance on the MSE and AUPR metrics than the shallow methods. Additionally, We show that our deep learning model, DeepPS is computationally more efficient than the deep learning model trained on full-length raw sequences of proteins. We conclude that a beneficial research approach would be to integrate structural information of proteins for modeling drug-target interaction prediction of large datasets for more interpretability, high throughput, and broad applicability.

Graphical abstract



Keywords Drug–target interaction · Machine learning · Deep learning · Protein–ligand interaction · Sequence alignment

1 Introduction

Rapid quantitative prediction of drug-target interaction is essential to drug discovery and development. For decades, the interactions between compounds and proteins were identified by carrying out expensive and time-consuming wet-lab experiments. Drug target interaction prediction is an important task in the drug discovery process, As the chemical space is large, of the order of 10^{60} molecules, it is arduous and almost impossible to identify interactions of all the compounds against different targets in the lab. On the other hand, computational screening of drug-target interaction aids in finding a smaller subset of

✉ S. Balaji
s.balaji@manipal.edu

Extended author information available on the last page of the article

probable candidates which could be taken up for further screening. Computational modeling can help identify side effects of compounds, as a single compound could have effects on multiple targets. Further, the models could be used to identify novel compounds interacting with known targets as well as find an alternate use for existing compounds based on novel interactions. Due to the emergence of high throughput screening, the amount of experimental data available in public databases has significantly increased. The availability of biological data relating to the protein sequence or structure in public databases has also grown tremendously. Chemogenomic models could utilize the available data to predict unknown interactions between proteins and compounds. These models predict the interactions between the ligands and targets by combining information from similar drugs and targets [1]. The models are constructed on the basis of the similarity principle which states that drugs with similar properties interact with similar targets [2].

The Proteochemometric (PCM) or a Chemogenomic model has the advantage of being able to predict interactions between ligands and proteins, even when there is no 3D structure available or when there are a few or no known ligands for the protein [3]. Ligand-based chemogenomic approaches are being pursued in drug discovery as they are computationally less expensive compared to structure-based approaches and can be trained on a large number of available bioactivity data. Consequently, the prediction of interactions greatly enhances the discovery of novel interacting targets and compounds that may find application in drug repurposing efforts [4–7].

Deep learning models such as CNN have shown excellent predictive capability in the field of computer vision. These methods have been used in bioinformatics in genomic studies as well as in models for drug discovery [8, 9]. These models are capable of identifying and learning complex patterns from molecular data [10]. The advantage of a deep learning CNN model is that the raw data can be represented better using non-linear transformations to effectively learn the hidden patterns in the data.

Several authors have studied protein–ligand interaction prediction using machine learning and deep learning techniques. Deep learning models using 3D structures of protein–ligand complexes were developed to predict interactions [11–13]. However, these methods are confined to known protein–ligand complexes. 2D Similarity-based methods using similarities of ligands against similar targets have been employed in predicting interactions. In KronRLS, the authors constructed chemical structure similarity matrices and sequence similarity matrices to represent ligands and proteins. The prediction for each protein–ligand pair is based on the similarity score, which is defined as a Kronecker product of the two matrices [14]. As this method captures only the linear dependencies in the data, a non-linear method, SimBoost, using gradient boosting machine [15] was introduced to predict binding affinity with a prediction interval [16]. In this method, a large number of features were calculated for each protein–ligand pair other than the ligand and the protein features using similarity matrices and constructed features. A deep belief network (DBN) was trained by stacking restricted boltzmann machines (RBM's) to predict novel DTI's between approved FDA drugs and targets using Extended-connectivity fingerprints (ECFP) and protein sequence composition (PSC) descriptors [17]. In another study, similar PSC descriptors were used to characterize proteins, and compounds were represented using molecular graph convolution (MGC) to train a scalable neural network model which was compared to the baseline machine learning models, SimBoost and KronRLS [18]. In MDeePred, the proteins were represented using physical, chemical and biological features using CNN to predict compound–protein interactions to achieve significant improvement in prediction performance compared to the baseline methods [19]. A deep LSTM model was used to predict DTIs on four target classes using chemical fingerprints and evolutionary information of proteins [20]. In DeepDTA, the one-dimensional SMILES representation of ligands and raw sequences of proteins were encoded into vector representations using CNN blocks. Further, the combined representations of ligands and proteins were employed to predict interactions. However, the protein sequences were not effectively represented as the model was trained on lengthy sequences [21]. In DeepCDA, the model learned the compound and protein encodings using a combination of CNN and LSTM in the feature encoder layer, which feeds the output to the subsequent layers [22]. The RNN-based encoders, seq2seq were used to encode SMILES of compounds and protein sequences separately in Deepaffinity [23]. The CNN models appended to the RNNs were used to concatenate the outputs of compounds and proteins and fed into more connected layers to predict affinity. The above-discussed machine learning models calculated the similarity matrices of drugs and targets which is computationally expensive. On the other hand, the deep learning models computed a large number of drug and protein descriptors which makes the models less interpretable. The unequal and raw protein sequences were used to model drug-target interaction (DTI) prediction in all the above methods which significantly increased the training time.

The hypothesis in this work is that an interpretable drug–target interaction prediction model could be developed using one-dimensional SMILES as drug descriptors, and protein-binding site subsequences. The prediction could be achieved by incorporating the combined features using a Deep CNN, which has outperformed state-of-the-art machine learning models due to its ability to learn useful patterns from raw data using a hierarchical structure of the deep neural network. The extracted protein subsequences contain useful binding information for representing the contact residues and residues involved in medium-range interactions.

In this paper, we have modeled the compound–protein interaction prediction using a one-dimensional representation of proteins and ligands by training a deep CNN model using the extracted features of proteins as subsequences. The protein subsequences incorporating the binding pocket information of proteins were used as input instead of the raw sequences. The method uses the one-dimensional features of drugs and proteins and does not require the 3D structures as inputs to the model. It is possible to develop predictive models by using the amino acid residues of the binding site where structural information of proteins is available [24]. As many structures of proteins are available, the structural information of the binding domain of proteins was utilized to obtain motif-rich binding site residues lining the binding pocket. If the protein 3D structure is unavailable, ligand-binding sites could be predicted using different sequence-based tools. Unlike the above-discussed models, our models are trained on shorter protein sequences using a hybrid approach by incorporating structural information of protein binding sites.

The main contributions of the paper are as follows:-

- Proposing a better representation of proteins by considering residues of the binding pocket.
- Improving the training time of the prediction model due to shorter protein sequences.
- Compared our proposed model with the state-of-the-art deep learning model using training epochs as an additional metric.

Table 1 Dataset statistics

	Davis	KIBA
Datapoints	30,056	1,18,254
Sequences	442	2111
Ligands	68	229
Bioactivity	Pk_d	PIC_{50}

2 Materials and Methods

2.1 Dataset

The composition of ligands, proteins, and interactions of the benchmark datasets, Davis and KIBA, is shown in Fig. 1. The bioactivity values of Davis and KIBA datasets were converted to Pk_d and PIC_{50}, as described in the previous literature (Table 1). For a fair comparison with the earlier methods, we divided the datasets into six equal parts. One part was taken as an independent test set. The remaining five parts were used for tuning the hyper-parameters through five-fold cross-validation.

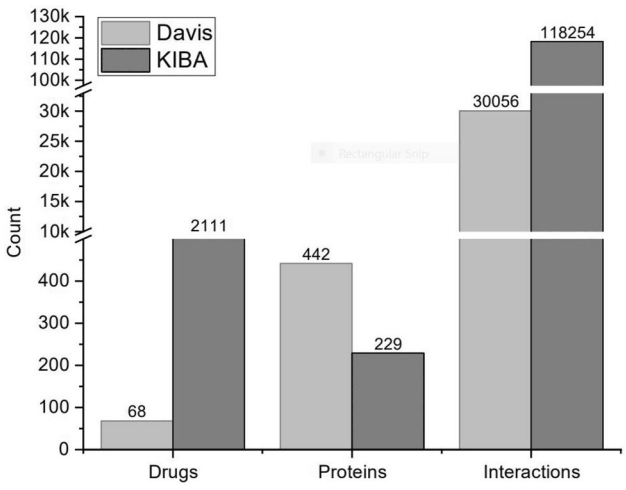


Fig. 1 Composition of DAVIS and KIBA Datasets

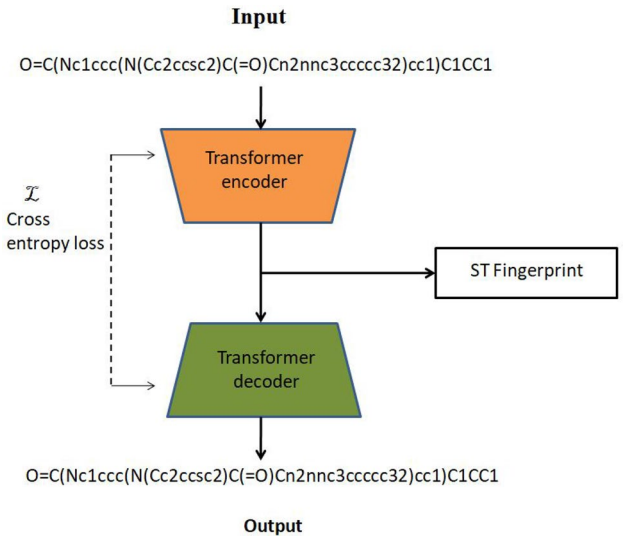


Fig. 2 Transformer encoder–decoder network using SMILES. The generated ST Fingerprint was used as input to the molecule encoder of the DeepPS (FP) model

2.2 Representation of Drugs

The molecules, represented as a one-dimensional SMILES notation [25] were encoded using CNN. The integer encodings were used to represent characters of SMILES comprising 64 labels. The integer encoded SMILES strings were given as input to the molecule encoder in the DeepPS model. However, in the DeepPS (FP) model, the one-dimensional SMILES strings of the molecules were used to generate fingerprints using the SMILES transformer, as shown in Fig. 2. The SMILES transformer comprised the encoder–decoder network with four transformer blocks each. Each block has four-head attentions with 256 embedding dimensions, and two linear layers [26]. The pre-trained SMILES transformer [27], trained on unlabelled SMILES, was used to generate ST Fingerprints. The symbol-level representations from each of the four transformer blocks were pooled together to obtain ST fingerprints. The fingerprints generated were 1024 bits for each molecule. The ST Fingerprints were used as input to the molecule encoder in the DeepPS (FP) model.

2.3 Feature Selection of Proteins

The Davis [28], and KIBA [29] datasets of kinases consists of the bio-activity data of typical and atypical kinases. A structure-based approach can understand the similarity and dissimilarity of both types of kinases' conserved regions. The conservation and variation of residues of the ATP binding pocket and the region in the vicinity of this pocket were studied using active-conformation structures [30]. The structure-based binding site alignment of conserved regions of highly similar kinases reveals the presence of common structural elements such as secondary structures and functional motifs such as "DFG" and "HRD" [31]. Most of the conserved regions are aligned. The unaligned blocks contain specific insertions of varying lengths in between in some kinases. Besides, some kinases have shifted secondary structures. Various types of inhibitors bind to the proteins at different binding sites. In typical kinases, the binding site consists of secondary structural elements and functional motifs present in the protein kinase domain. The key regions which are associated with the binding of inhibitors are the HRD motif, DFG motif, G-rich loop, alphaC-helix, catalytic and activation loops [32–34] To identify the binding domain, the protein sequences of the kinases in the datasets were extracted from Uniprot [35].

The binding sites of protein kinases contain specific motifs which are rich in information attributing to kinase specificity. Identifying the conserved regions that contribute to the specificity of kinases and representing them to be amenable for modeling can provide better predictive capability and interpretation. The varying amino acids in the conserved regions contribute to the specificity as the binding region is highly conserved in kinases. The binding site residues obtained from the catalytic cleft of kinases enable the comparison of the interaction pattern of kinase inhibitors. The binding pocket residues of all protein kinases present in the datasets were extracted using the binding site positions from the sequences after performing sequence alignment of the structural elements implicated in the binding process [36]. All the protein subsequences of the binding pocket comprised the G-rich loop, alphaC-helix, catalytic loop, and motifs such as VAIK/VAVK motif, HRD motif, and the DFG motif and seem to be aligned to the respective motif positions, except for some atypical kinases which had missing or differing secondary structure elements. The fixed length of 85 binding pocket residues was obtained for each of the proteins (Fig. 3). The protein subsequences containing the binding pocket residues were used as input to the protein encoder in the DeepPS and DeepPS (FP) models.

NAME	POCKET RESIDUES
"ABL1p"	HKLGGQQYGEVYEVAVKTLEFLKEAAVMKEIKPNLVQLLGVYIITEFMTYGNLLDYLREYLEKKNFIHRDLAARNCLVVADFGLS
"ALK"	RGLGHGAFGEVYEVAVKTLDFLMEALIISKFNQNIVRCIGVFILLELMAGGDLKSFLREYLEENHFIHRDIAARNCLLIGDFGMA
"AMPK-alpha1"	DTLGVGTFGKVKVAVKILKIRREIQNLKLFRPHIIKLYQVFMVMEYVSGGELFDYICKYCHRHMVVHRDLKPENVLLIADFGLS
"ASK2"	LVLGKGTYGVVYAIAIKEIPLHEEIALHRRLRKNIVRYLGSKIFMEEVPGGSLSSLLRSYLHDNHIVHRDIKGDNVLIISDFGTS
"AURKA"	RPLGKGKFGNVYLLALKVLQLRREVEIQSHLRPNILRLYGYFLILEYAPLGTVYRELQKYCHSKRVIHRDIKPENLLLIADFGWS

Fig. 3 Examples of the binding pocket residues with highlighted motifs selected as protein features

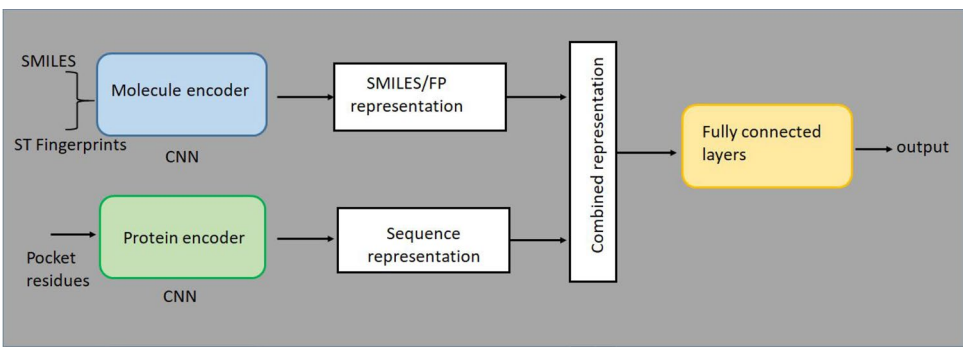


Fig. 4 CNN-based Chemogenomic model with SMILES and binding pocket residues as inputs for drugs and proteins (DeepPS). The inputs for drugs in DeepPS (FP) are ST Fingerprints and binding pocket residues for proteins)

2.4 Proposed Chemogenomic Model

The proposed CNN-based chemogenomic models with deep learning contain four main building blocks. The first block is the molecule encoder that encodes the SMILES strings of ligands in the DeepPS model and ST Fingerprints in the DeepPS (FP) model. The SMILES strings were represented using integer encodings to represent unique letters. The Transformer–encoder–decoder network was utilized to generate SMILES transformer fingerprints (ST fingerprints)in DeepPS (FP) model. Second, the protein encoder embeds the features from the pocket residues given as input using label encodings to represent 26 categories. The pocket residues are used as inputs for proteins in both DeepPS and DeepPS (FP) models. The SMILES and proteins were of different lengths for both datasets. The fixed length of 85 and 100 were chosen for SMILES of Davis and KIBA datasets, respectively, while the length of the protein subsequences was fixed at 85 for both datasets for effective representation. The outputs from the molecule encoder and the protein encoders were concatenated and given as input to three fully connected layers of the CNN with dropout layers in between them. The dropout layers are used to reduce the overfitting of the data. The final CNN output layer predicts the outputs. The architecture, along with the building blocks, is shown in Fig. 4.

3 Results

The baselines evaluated in our experiments are the KronRLS, SimBoost, and DeepDTA. The DeepPS model was trained on the SMILES and binding pocket residues. The second model, DeepPS (FP) was trained on Smiles transformer fingerprints and binding pocket residues.

We evaluated the performance of our method on benchmark datasets Davis and KIBA. The same settings for the train and test folds were used as given in the literature [21] for a fair comparison. The entire dataset was divided into six folds, out of which one fold was used as an independent test set. The remaining folds were used for training using nested cross-validation to obtain tuned hyper-parameters. The parameter settings used for the CNN model are as given in Table 2. The maximum sequence length of the proteins for the models was set to 85 for both datasets as only 85 residues are involved in binding. An early stopping strategy using validation mean squared error (MSE) as a performance measure was adopted to avoid overfitting of the model during training.

Table 2 Parameter settings for CNN

Parameters	Range
No. of filters	32*1; 32*2; 32*3
Length of filter (compounds)	[4, 6, 8]
Length of filter (proteins)	[4, 8, 12]
Hidden neurons	1024; 1024; 512
Batch size	256
Dropout	0.1
Optimizer	Adam
Learning rate	0.001

3.1 Evaluation Metrics

In this study, we used four evaluation metrics, MSE, CI, r_m^2, and Area under precision-recall (AUPR). The evaluation metrics other than AUPR were used to evaluate continuous regression outputs. AUPR was obtained by binarising the regression outputs based on the threshold value. A threshold value of 7 was chosen for the Davis dataset and 12.1 for the KIBA dataset according to the previous work [37].

Concordance index (CI) was utilized to measure the effectiveness of the model with continuous outputs [38]. It measures the probability of the similarity between the actual values and the predicted values of two random protein–ligand pairs.

$$CI = \frac{1}{Z} \sum_{\delta_i > \delta_j} h(m_i - m_j) \tag{1}$$

where m_i is the prediction value for the greater affinity δ_i, m_j is the predicted value for the smaller affinity δ_j, Z is the normalization constant that equals the number of data pairs with different label values and $h(x)$ is the step function defined as

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

Mean squared error is defined as

$$MSE = \sum_{i=1}^{\infty} h(y_i - y_j)^2 \tag{2}$$

The external predictive power of the model is given by r_m^2 metric, which is defined as follows.

$$r_m{}^2 = r^2 \times (1 - \sqrt{r^2 - r_o^2}) \tag{3}$$

where r_o^2 is the squared correlation coefficient without intercept, r^2 is the squared correlation coefficient with intercept.

Table 3 The average Concordance Index and Mean squared error scores of the test set of Davis dataset for the compared methods

	Proteins	Compounds	CI	MSE
KronRLS	S-W	Pubchem Sim	0.871 (0.0008)	0.379
SimBoost	S-W	Pubchem Sim	**0.872 (0.002)**	**0.282**
DeepDTA	CNN	CNN	0.851 (0.004)	0.379
DeepPS (FP)	CNN	Tran-CNN	0.861(0.007)	0.375
DeepPS	CNN	CNN	0.854(0.007)	0.353

For every metric, the value for the best performing method has been highlighted in bold font

Table 4 The average Concordance index and Mean squared error scores of the test set of KIBA dataset for the compared methods

	Proteins	Compounds	CI	MSE
KronRLS	S-W	Pubchem Sim	0.782 (0.0009)	0.411
SimBoost	S-W	Pubchem Sim	0.836 (0.001)	0.222
DeepDTA	CNN	CNN	0.765(0.002)	0.375
DeepPS (FP)	CNN	Tran-CNN	0.782(0.003)	0.310
DeepPS	CNN	CNN	**0.844(0.003)**	**0.218**

For every metric, the value for the best performing method has been highlighted in bold font

Table 5 The average r_m^2 and AUPR scores of the test set for the Davis dataset

	Proteins	Compounds	r_m^2(std)	AUPR(std)
KronRLS	S-W	Pubchem Sim	0.407 (0.005)	0.661 (0.010)
SimBoost	S-W	Pubchem Sim	**0.644 (0.006)**	0.709 (0.008)
DeepDTA	CNN	CNN	0.526 (0.017)	0.567 (0.010)
DeepPS (FP)	CNN	Tran-CNN	0.573(0.005)	0.681(0.005)
DeepPS	CNN	CNN	0.546(0.003)	**0.710(0.003)**

For every metric, the value for the best performing method has been highlighted in bold font

The area under the precision-recall (AUPR) curve assesses a binary model by taking the average of the precision values across all recall values. The AUPR method is suitable for estimating the accuracy of datasets having imbalanced classes with skewed distribution [39]. The thresholds for binarising the outputs were chosen as proposed by He et al. [16].

In addition to these metrics, our models were evaluated on training time as an additional metric to gain insights on the model training in order to avoid overfitting of the model.

3.2 Comparison of Shallow and Deep Learning Models

The results of our chemogenomic models were compared with the baseline machine learning shallow methods, KronRLS and SimBoost, and the deep learning method, DeepDTA. The models were compared against shallow methods as these methods were trained on computed features. The results obtained by applying our method on Davis and KIBA datasets were evaluated on average mean squared error (MSE) and average Concordance index (CI) over the independent test set. The results on Davis and KIBA datasets are presented in Tables 3 and 4. The results obtained by the shallow methods have been taken from the literature. The code for the deep learning method DeepDTA was downloaded and run in our setting for comparison.

In the Davis dataset, our models have achieved comparable performances on the MSE and CI values against the other methods. Even though SimBoost has slightly better performance than DeepPS, our model is scalable and performs better than SimBoost on time and space complexity metrics as SimBoost requires computationally expensive matrix factorization as it relies on similarity matrices. The DeepPS model has achieved better performance than DeepPS (FP) as some information could have been lost in the generation of fingerprints.

On the KIBA dataset, the performance of the DeepPS model is better than all the models on the CI and MSE metrics. The KIBA dataset consists of more proteins and interaction data as compared to the Davis dataset resulting in better generalization.

The external predictivity of the model on an independent test set was analyzed using the r_m^2 metric [40]. The values obtained for our models were greater than 0.5 for both Davis and KIBA datasets indicating that the models were acceptable. The standard deviations are given in parenthesis. The AUPR values were calculated by binarising the outputs. A threshold ($Pk_d \geq 7$) was set for the Davis dataset to classify as binding. For the KIBA dataset, the value was set to 12.1. The results of the r_m^2 and AUPR metrics for both datasets are summarized in Tables 5 and 6.

The excellent correlation of the predictions obtained by different input representations and methods employed removes the chance correlation and emphasizes the predictive power of the models developed. The predicted versus

Table 6 The average r_m^2 and AUPR scores of the test set for the KIBA dataset

	Proteins	Compounds	r_m^2 (std)	AUPR (std)
KronRLS	S-W	Pubchem Sim	0.342 (0.001)	0.635 (0.004)
SimBoost	S-W	Pubchem Sim	**0.629 (0.007)**	0.760 (0.003)
DeepDTA	CNN	CNN	0.458 (0.009)	0.582 (0.004)
DeepPS (FP)	CNN	Tran-CNN	0.517(0.005)	0.691(0.002)
DeepPS	CNN	CNN	0.604(0.003)	**0.762(0.004)**

For every metric, the value for the best performing method has been highlighted in bold font

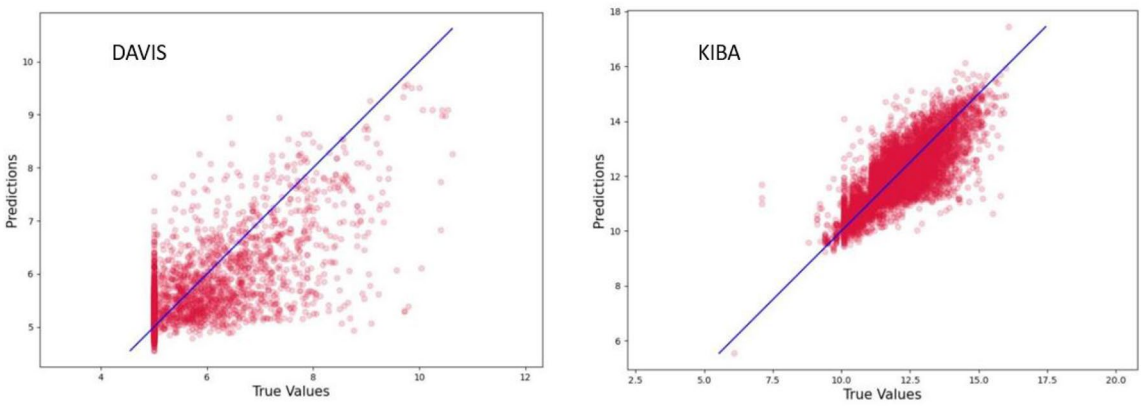


Fig. 5 Binding affinity prediction of DeepPS model. **a** experimental values versus predicted values for Davis dataset. **b** experimental values versus predicted values for KIBA dataset

Table 7 Comparison of metrics with baseline Deep learning model

Dataset	Method	Specificity	Sensitivity	AUROC
Davis	DeepDTA	0.98	0.36	0.67
	DeepPS	0.99	0.24	0.62
KIBA	DeepDTA	0.93	0.64	0.78
	DeepPS	0.93	0.65	0.79

actual plots obtained from the DeepPS model on Davis and KIBA datasets are shown in Fig. 5. The Davis dataset has less diverse ligands compared to the KIBA dataset. The data points are aggregated around the regression line in KIBA dataset compared to the Davis dataset.

Based on the obtained results, it could be inferred that the binding between proteins and ligands depends not only on the binding pocket residues but also on residues outside the binding pocket that can have long-range effects. The flexibility of the protein and conformational adjustment during the binding process also contributes to the binding as adjacent pockets may also be involved in binding [41]. Our model based on the binding pocket residues achieved better or comparable results than shallow methods on MSE and CI metrics on both datasets suggesting that the motif-rich features representing the binding pocket were able to capture the physicochemical properties of the pocket. The motif-rich subsequences are part of the secondary structural elements of the proteins interacting with the ligands that contain the necessary binding features. Predicting novel interactions between ligands and proteins in drug discovery is more important than missing them out. In other words, false negatives should be minimized as false positives do get checked during wet-lab experiments. To achieve this, our model is computationally efficient but slightly less accurate for large-scale binding affinity prediction compared to other deep models trained only on raw sequences and SMILES strings.

3.3 Performance Evaluation of the Deep Learning Models on Davis and KIBA Datasets

For evaluating our model's performance, various metrics such as specificity, sensitivity, and accuracy were also computed by taking the thresholds from the generated regression outputs. The DeepPS model achieved slightly better performance compared to DeepDTA on the KIBA dataset and slightly lower values on the Davis dataset (Table 7). The low values could be attributed to the protein features included in the model. As the Davis dataset consists of a lesser number of proteins and interactions compared to the KIBA dataset, the model may not have been able to completely capture the patterns in the data. Also, as the binding affinity between drugs and targets depends on the local and non-local interactions, including distant amino acid residues contributing to the non-local interactions may improve model performance [42].

The training time of a model is proportional to the size of the inputs. Our best performing model, DeepPS, was evaluated on training time with the DeepDTA method taken as a baseline for comparison. The DeepDTA method was chosen as a baseline deep learning model as our models employed CNN blocks for encoding drug and protein features similar to DeepDTA. The plots of average CI and MSE metrics on the training sets for the Davis and KIBA datasets of the DeepPS are displayed in Fig. 6. The DeepPS model shows considerable improvement in training time, as seen by the fewer epochs. The learning of DeepPS was completed in 25 epochs and 35 epochs for Davis and KIBA datasets, respectively. The concordance index curves and the loss curves for the training and validation set indicate that the model is not

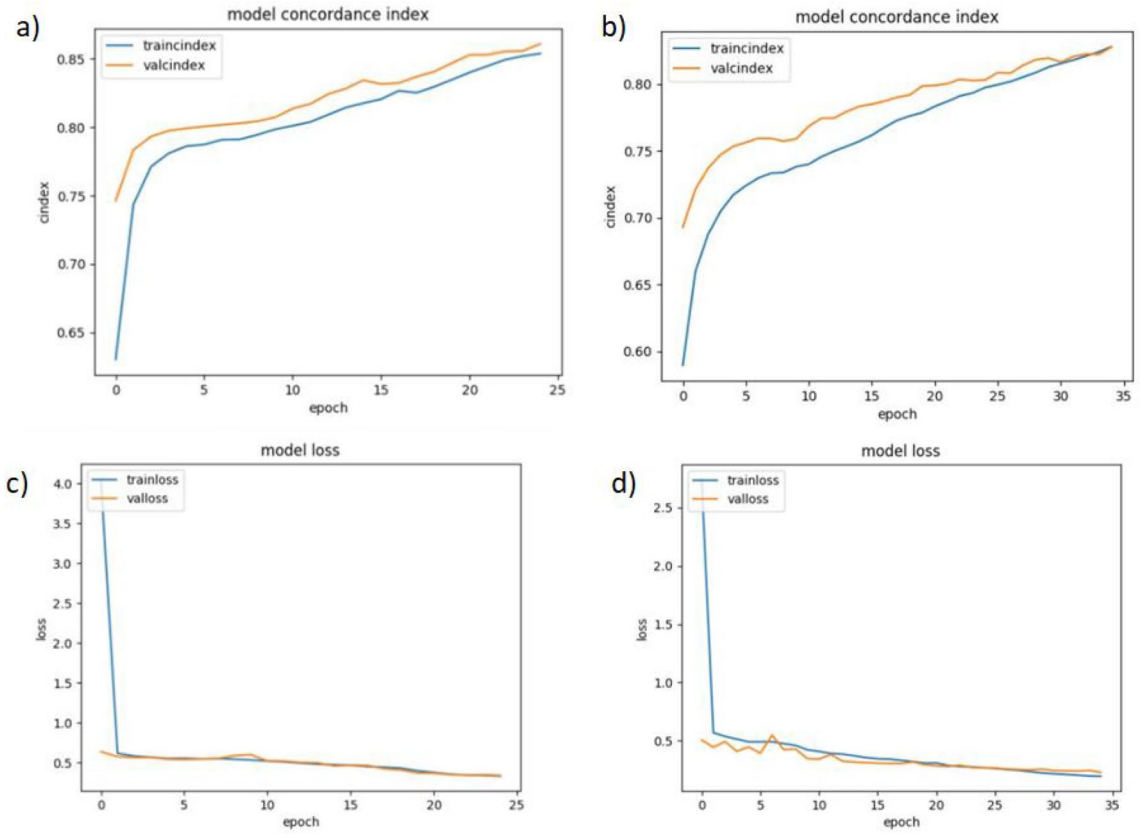


Fig. 6 Training results of DeepPS model. **a** CI plot for validation set and training sets of Davis dataset. **b** CI plot for validation set and training sets of KIBA dataset. **c** Plot of MSE for validation set and training sets of Davis dataset. **d** Plot of MSE for validation set and training sets of KIBA dataset

overfitting. The curves show a good fit as the validation gap is reduced to the point of stability (epochs) and has a small gap with the training curves. The training of the DeepDTA model was completed in 100 epochs for both datasets, which is longer. The improvement in training time of DeepPS could be attributed to the shorter sequence length of the proteins.

Further, we compared the execution time taken by our method against DeepDTA on the Davis and KIBA datasets. The results are summarized in Table 8.

The DeepPS and DeepDTA were executed on Google Colab cloud platform on a GPU machine. The starting and completion times were recorded. On comparing the training times, we can see that DeepPS is able to complete the job faster compared to DeepDTA mainly because of the pre-processing step. Further optimization could be achieved if the pre-processing step is incorporated in the DeepPS algorithm instead of a executing it as a separate script. The pre-processing involved aligning the protein structures to obtain the binding amino acids. The pre-processing step was carried out on a standalone machine.

Table 8 The approximate training time (in hours) of Davis and KIBA datasets

Method	Dataset	Pre-processing	Training time
DeepPS	Davis	2	4
	KIBA	3	7
DeepDTA	Davis	NIL	10
	KIBA	NIL	15

4 Conclusion

An understanding of the important features contributing to the predictive performance of the model is important for model optimization. However, as deep learning models are considered black boxes as it is not easy to understand the contributing features, we tried to optimize the neural network model by extracting the relevant protein features and combining them with the drug features. The proposed deep learning-based method predicts drug-target interactions using only one-dimensional SMILES strings of drugs and protein subsequences obtained from binding pocket information thereby proving our hypothesis. The CNN blocks were used for encoding one-dimensional descriptors of drugs and proteins. Further, our model trained on the binding site residues achieved comparable performance to the

baseline shallow methods and is computationally efficient than the baseline machine learning models as it does not require the construction of similarity matrices. This study also offers additional confidence to the previous works on these datasets to generalize using a hybrid chemogenomic approach for computationally efficient drug-target interaction prediction compared to other approaches while offering comparable performance values. Our findings could be used to model drug-target interactions to find side effects that could be used in drug repurposing efforts. Finally, this work provides a faster, rational, and straightforward predictive model that may be employed to guide future experiments in drug discovery.

Funding Open access funding provided by Manipal Academy of Higher Education, Manipal.

Data Availability The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of Interest All authors declare that they have no conflict of interest.

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. van Westen GJ, Wegner JK, IJzerman AP, van Vlijmen HW, Bender A, (2011) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. Med Chem Comm 2(1):16–30. https://doi.org/10.1039/C0MD00165A
2. Klabunde T (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. Br J Pharmacol 152(1):5–7. https://doi.org/10.1038/sj.bjp.0707307
3. Jacob L, Vert J-P (2008) Protein–ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 24(19):2149–2156. https://doi.org/10.1093/bioinformatics/btn409
4. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4(11):682–690. https://doi.org/10.1038/nchembio.118
5. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB et al (2009) Predicting new molecular targets for known drugs. Nature 462(7270):175–181. https://doi.org/10.1038/nature08506
6. Li YY, An J, Jones SJ (2011) A computational approach to finding novel targets for existing drugs. PLoS Comput Biol 7(9):1002139. https://doi.org/10.1371/journal.pcbi.1002139
7. Li Y, Jones S (2012) Drug repositioning for personalized medicine. Genome Med 4:27. https://doi.org/10.1186/gm326
8. Leung MK, Xiong HY, Lee LJ, Frey BJ (2014) Deep learning of the tissue-regulated splicing code. Bioinformatics 30(12):121–129. https://doi.org/10.1093/bioinformatics/btu277
9. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR et al (2015) The human splicing code reveals new insights into the genetic determinants of disease. Science. https://doi.org/10.1126/science.1254806
10. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
11. Wallach I, Dzamba M, Heifets A (2015) AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery, 1–11. arXiv:1510.02855. https://doi.org/10.1007/s10618-010-0175-9
12. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR (2017) Protein–ligand scoring with convolutional neural networks. J Chem Inf Model 57(4):942–957. https://doi.org/10.1021/acs.jcim.6b00740
13. Gomes J, Ramsundar B, Feinberg EN, Pande VS (2017) Atomic convolutional networks for predicting protein-ligand binding affinity. arXiv preprint arXiv:1703.10603. https://doi.org/10.48550/arXiv.1703.10603
14. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, Aittokallio T (2014) Toward more realistic drug–target interaction predictions. Brief Bioinform 16(2):325–337. https://doi.org/10.1093/bib/bbu010
15. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
16. He T, Heidemeyer M, Ban F, Cherkasov A, Ester M (2017) Simboost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J Cheminformatics 9(1):1–14. https://doi.org/10.1186/s13321-017-0209-z
17. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H (2017) Deep-learning-based drug–target interaction prediction. J Proteome Res 16(4):1401–1409. https://doi.org/10.1021/acs.jproteome.6b00618
18. Feng Q, Dueva E, Cherkasov A, Ester M (2018) Padme: a deep learning-based framework for drug-target interaction prediction. arXiv preprint arXiv:1807.09741. https://doi.org/10.48550/arXiv.1807.09741
19. Rifaioglu AS, Cetin Atalay R, Cansen Kahraman D, Doğan T, Martin M, Atalay V (2021) Mdeepred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery. Bioinformatics 37(5):693–704. https://doi.org/10.1093/bioinformatics/btaa858
20. Wang Y-B, You Z-H, Yang S, Yi H-C, Chen Z-H, Zheng K (2020) A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. BMC Med Inform Decis Mak 20(2):1–9. https://doi.org/10.1186/s12911-020-1052-0
21. Öztürk H, Özgür A, Ozkirimli E (2018) DeepDTA: deep drug-target binding affinity prediction. Bioinformatics 34(17):821–829. https://doi.org/10.1093/bioinformatics/bty593
22. Abbasi K, Razzaghi P, Poso A, Amanlou M, Ghasemi JB, Masoudi-Nejad A (2020) Deepcda: deep cross-domain compound-protein affinity prediction through LSTM and

convolutional neural networks. Bioinformatics 36(17):4633–4642. https://doi.org/10.1093/bioinformatics/btaa544
23. Karimi M, Wu D, Wang Z, Shen Y (2019) Deepaffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics 35(18):3329–3338. https://doi.org/10.1093/bioinformatics/btz111
24. Cortés-Ciriano I, Ain QU, Subramanian V, Lenselink EB, Méndez-Lucio O, IJzerman AP, Wohlfahrt G, Prusis P, Malliavin TE, van Westen GJ, et al (2015) Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. Med Chem Comm 6(1):24–50. https://doi.org/10.1039/C4MD00216D
25. Weininger D (1988) Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36. https://doi.org/10.1021/ci00057a005
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, p 5998–6008. https://doi.org/10.48550/arXiv.1706.03762
27. Honda S, Shi S, Ueda HR (2019) Smiles transformer: pre-trained molecular fingerprint for low data drug discovery. arXiv preprint arXiv:1911.04738. https://doi.org/10.48550/arXiv.1911.04738
28. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 29(11):1046–1051. https://doi.org/10.1038/nbt.1990
29. Tang J, Szwajda A, Shakyawar S, Xu T, Hintsanen P, Wennerberg K, Aittokallio T (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model 54(3):735–743. https://doi.org/10.1021/ci400709d
30. Knight JD, Qian B, Baker D, Kothary R (2007) Conservation, variability and the modeling of active protein kinases. PLoS One 2(10):982. https://doi.org/10.1371/journal.pone.0000982
31. Modi V, Dunbrack RL (2019) A structurally-validated multiple sequence alignment of 497 human protein kinase domains. Sci Rep 9(1):1–16. https://doi.org/10.1038/s41598-019-56499-4
32. Hemmer W, McGlone M, Tsigelny I, Taylor SS (1997) Role of the glycine triad in the atp-binding site of camp-dependent protein kinase. J Biol Chem 272(27):16946–16954. https://doi.org/10.1074/jbc.272.27.16946
33. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C et al (2007) Patterns of somatic mutation in human cancer genomes. Nature 446(7132):153–158. https://doi.org/10.1038/nature05610
34. Kanev GK, de Graaf C, de Esch IJ, Leurs R, Würdinger T, Westerman BA, Kooistra AJ (2019) The landscape of atypical and eukaryotic protein kinases. Trends Pharmacol Sci 40(11):818–832. https://doi.org/10.1016/j.tips.2019.09.002
35. Uniprot (2021) The universal protein knowledgebase in 2021. Nucleic Acids Res 49(D1):480–489. https://doi.org/10.1093/nar/gkaa1100
36. Kanev GK, de Graaf C, Westerman BA, de Esch IJ, Kooistra AJ (2021) Klifs: an overhaul after the first 5 years of supporting kinase research. Nucleic Acids Res 49(D1):562–569. https://doi.org/10.1093/nar/gkaa895
37. Öztürk H, Özgür A, Ozkirimli E (2018) Deepdta: deep drug-target binding affinity prediction. Bioinformatics 34(17):821–829. https://doi.org/10.1093/bioinformatics/bty593
38. Gönen M, Heller G (2005) Concordance probability and discriminatory power in proportional hazards regression. Biometrika 92(4):965–970. https://doi.org/10.1093/biomet/92.4.965
39. Raghavan V, Bollmann P, Jung GS (1989) A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans Inf Syst (TOIS) 7(3):205–229. https://doi.org/10.1145/65943.65945
40. Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN (2013) Some case studies on application of "rm2'' metrics for judging quality of quantitative structure-activity relationship predictions: emphasis on scaling of response data. J Comput Chem 34(12):1071–1082. https://doi.org/10.1002/jcc.23231
41. Stank A, Kokh DB, Fuller JC, Wade RC (2016) Protein binding pocket dynamics. Acc Chem Res 49(5):809–815. https://doi.org/10.1021/acs.accounts.5b00516
42. Zheng L, Fan J, Mu Y (2019) Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. ACS Omega 4(14):15956–15965. https://doi.org/10.1021/acsomega.9b01997

Authors and Affiliations

Sofia D'Souza[1] · K. V. Prema[2] · S. Balaji[3] · Ronak Shah[1]

Sofia D'Souza
sofiaregomlr@gmail.com

K. V. Prema
prema.kv@manipal.edu

Ronak Shah
ronakshah2040@gmail.com

[1] Department of Computer Science and Engineering, Manipal Academy of Higher Education, Manipal, India

[2] Department of Computer Science and Engineering, Manipal Academy of Higher Education, Bengaluru, India

[3] Department of Biotechnology, Manipal Academy of Higher Education, Manipal, India