Scuola universitaria professionale
della Svizzera italiana

SUPSI

University of Applied Sciences and Arts of Southern Switzerland
Department of Innovative Technologies

Applied Case Studies of Machine Learning and Deep Learning in Key Areas II

DRUG TARGET INTERACTIONS

Andrea Wey
andrea.wey@student.supsi.ch

Carlo Grigioni
carlo.grigioni@student.supsi.ch

Christian Pala
christian.pala@student.supsi.ch

Professor: Gianvito Grasso
SUPSI, Lugano Switzerland

22/05/2023

Table of Contents

1. Problem definition **1**

2. State-of-the-art **1**
2.1 DTI literature review . . . 1
2.1.1 Modeling architectures . . . 1
2.2 Literature evaluation metrics . . . 1
2.3 Challenges and Opportunities in DTI Prediction . . . 2

3. Data **3**
3.1 DAVIS . . . 3
3.2 KiBA . . . 3
3.3 BindingDB . . . 3

4. Data preprocessing **3**

5. Metrics **4**
5.1 Concordance Index . . . 4
5.2 Mean Squared Error . . . 4
5.3 Pearson Correlation . . . 4

6. Modeling **4**
6.1 Drug encoding . . . 4
6.2 Target encoding . . . 5
6.3 CNNs . . . 5
6.4 DeepPurpose Results . . . 5
6.5 Modeling results . . . 6

7. Deployment **6**

8. Limitations **6**

9. Conclusions **6**

References **8**

List of Figures

1 DeepDTA model architecture with CNN encodings Source: Özgür et al.[5] 2

List of Tables

1 Feature Encodings 3
2 DeepPurpose KiBA Results with CNN Protein Encoding 5
3 DeepPurpose Davis Results with CNN Protein Encoding 5
4 Modeling results on 10% BindgingDB 6

1. Problem definition

Prediction of drug-target interactions (DTI) plays a vital role in drug development. It is integral to a variety of areas, such as virtual screening, drug repurposing, and the identification of potential drug side effects. The accurate prediction of these interactions can make the difference between discovering a potentially life-saving drug and spending significant resources developing a drug with an incorrect mechanism of action. Traditional methods of identifying these interactions, typically biological experiments, are often costly and time-consuming. Therefore, computational approaches have been developed to improve the efficiency of this process and reduce costs. These methods aim to narrow down the search space of drug and protein candidates, which is crucial for accelerating drug discovery and development [1].

In this project, we aim to predict the binding affinity of a small molecule and of a protein target, starting from the ligand SMILES description and the protein amino acid sequence.

2. State-of-the-art

Artificial Intelligence (AI) and Machine Learning (ML) have made remarkable strides in the realm of Drug-Target Interaction (DTI) tasks. These advancements have been fueled by diverse methodologies and strategies, drawing inspiration from Natural Language Processing (NLP) models employed for sentence classification. Prominent techniques in DTI encompass Recurrent Neural Networks (RNNs), Transformers, and Convolutional Neural Networks (CNNs) influenced by the field of Computer Vision. These models have undergone adaptation, and have demonstrated their effectiveness in DTI research. Looking at performances on a large dataset such as BindingDB, the best results have been obtained by recent publications exploiting graph learning models [2] and transformers with self-attention [3], for instance achieving area-under-the-curve (AUC) scores of up to 0.971, for certain tasks.

2.1 DTI literature review

We conducted a review of the literature, with a particular emphasis on a 2021 survey of drug-target interaction predictions[4]. We also focused on two recent studies: "DeepDTA: Deep Drug-Target Binding Affinity Prediction"[5], which was evaluated on the KiBA benchmark dataset[6], and "Interpretable Drug Target Prediction Using Deep Neural Representation"[7], evaluated on BindingDB[8], another very relevant source, which is also the dataset professor Grasso provided us for the project

2.1.1 Modeling architectures

In their study, Özgür et al. [5] proposed a novel architecture for predicting drug-target interactions. This architecture uses CNNs for encoding the chemical structure of the drug ligand, represented as a SMILES string, and the sequence of the protein target, expressed as an amino acid string. Once encoded, these representations are fed into a Fully Connected Neural Network (FCN) for prediction.

The FCN used in the DeepDTA model [5] comprises two layers, each with 1024 neurons. Each of these layers is followed by a dropout layer for regularization, with a dropout rate of 0.1. The network then concludes with a final layer consisting of 512 nodes, followed by the output layer. Our aim is to not only replicate this architecture but also to investigate alternative methods for encoding the drug ligand and protein target.

2.2 Literature evaluation metrics

In the field of Drug-Target Interaction (DTI), one of the commonly used evaluation metrics is the Concordance Index (CI). The CI is particularly useful for DTI tasks because the interaction affinities, which form the basis of these interactions, are continuous values. The CI calculates the probability that, for two randomly selected drug-target pairs with different actual values, the

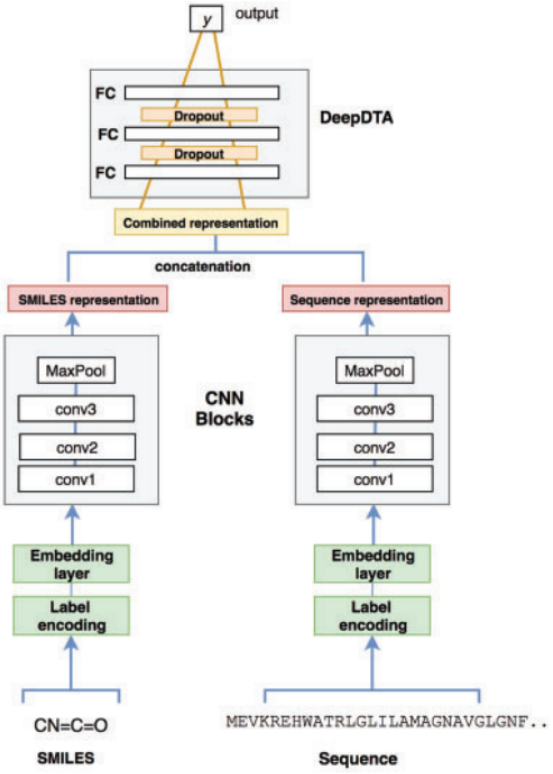

Figure 1: DeepDTA model architecture with CNN encodings
Source: Özgür et al.[5]

pair with the higher predicted value also has the higher actual value. This measure allows for the evaluation of the model's ability to correctly rank the interactions [9].

2.3 Challenges and Opportunities in DTI Prediction

Despite the advancements in DTI prediction, several challenges remain. Firstly, the sheer complexity and variability of biological systems make the prediction task difficult. Each drug and protein target has a unique, intricate structure that defines its behavior, and small alterations can dramatically impact the outcome.

Secondly, while the quantity of available bioactivity data is vast, it is also very imbalanced. Only a small fraction of possible drug-target pairs have been experimentally validated, which leaves a significant portion of the drug-target space unexplored. This imbalance in the data can cause models to be biased toward predicting non-interactions.

Lastly, the interpretability of machine learning models is a significant concern. While these models can achieve high predictive performance, understanding how they arrive at their predictions is critical, particularly in the context of drug discovery, where insights into the nature of the drug-target interaction can guide further research and development. This makes DTI an active and interesting field of research, particularly now with the explosion of large language models and our ability to encode language, including chemical representation, in our machine-learning models.

3. Data

For our models, we conducted tests on benchmark datasets, which are detailed in the subsequent sections.

3.1 DAVIS

Dataset Description: This dataset encompasses the interaction of 72 kinase inhibitors with 442 kinases, which accounts for more than 80% of the human catalytic protein kinome.

Dataset Statistics: DAVIS comprises 25,772 DTI pairs, inclusive of 68 drugs and 379 proteins.[10]

3.2 KiBA

Dataset Description: KiBA amalgamates various types of bioactivity information, including IC50, K(i), and K(d), to generate an integrated drug-target bioactivity matrix. This integrative approach was introduced by Tang et al. [6].

Dataset Statistics: The KiBA dataset consists of 117,657 DTI pairs, 2,068 drugs, and 229 proteins.

3.3 BindingDB

Dataset Description: BindingDB is a publicly accessible database that holds approximately 1.84 million measured binding affinity observations. The data primarily focuses on the interactions of proteins (which are considered as potential drug targets) with small, drug-like molecules. Our analysis was performed on a 10% subsample of the dataset provided by Gilson et al. [8].

Dataset Statistics: The subsample of the BindingDB dataset used in our analysis contains 183,673 DTI pairs, including 97,559 drugs and 13,712 proteins.

Please note: BindingDB is a comprehensive collection of various assays that have been collated and made accessible for research purposes.[8]

4. Data preprocessing

The two key preprocessing steps required for drug target interaction (DTI) predictions are encoding the ligand drug SMILES and protein target sequences. In our codebase [11], we implemented the approaches presented in Table 1, following solutions from literature and using the built-in methods exposed by DeepPurpose.

Table 1: Feature Encodings

SMILES	**Sequences**
Implemented	
Morgan Fingerprint	Label Encoding
Morgan Fingerprint	Conjoint Triad
Morgan Fingerprint	Pretrained Transformer
Pretrained Transformer	Pretrained Transformer
From DeepPurpose	
Morgan Fingerprint	Conjoint Triad and CNN
PubChem Fingerprint	Conjoint Triad and CNN
Daylight Fingerprint	Conjoint Triad and CNN
rdkit2dnormalized	Conjoint Triad and CNN
CNN	Conjoint Triad and CNN
CNN_RNN	Conjoint Triad and CNN
Transformer	Conjoint Triad and CNN

We organized our data to properly manage and batch the encoded SMILES, encoded protein sequences, and normalized affinity scores, for the modeling phase.

5. Metrics

In this section, we present the evaluation metrics we tested our models on.

5.1 Concordance Index

As mentioned in our literature review, we evaluated our models on the CI, computed as follows:

$$C = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{h_i > h_j}{2} + \frac{h_i = h_j}{2} \frac{d_i < d_j}{2} \right)$$

5.2 Mean Squared Error

Frequently employed in regression analysis, the Mean Squared Error (MSE) is a robust metric that quantifies the average squared discrepancies between the predicted and true values. It essentially calculates the mean of the squares of the prediction errors, where an error is defined as the difference between the actual value and the model's estimated value. A smaller MSE indicates a more precise prediction. The mathematical formulation of MSE is as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

5.3 Pearson Correlation

The Pearson correlation coefficient, denoted as rr, is a measure of the linear relationship between two variables. It quantifies the strength and direction of the linear association between two sets of numerical data. The coefficient ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. The Pearson correlation is widely used in statistics and research to assess the degree of association between variables.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

6. Modeling

We carried out the training and testing of the models provided by DeepPurpose on our three benchmark datasets to establish baseline values for our drug-target interaction (DTI) regression task. Furthermore, we fine-tuned the most effective architecture on the BindingDB dataset to examine how hyperparameter tuning might influence our outcomes. We adopted the built-in cold protein split method to preclude data leakage.

We replicated the architecture from Özgür et al. [5], adding modularity in the encoding selection, allowing us to evaluate different encoding and embedding strategies, for both the drug ligands and protein sequences.

6.1 Drug encoding

For the ligands, we examined two encoding strategies: producing morgan fingerprints using the RDKit [12] implementation, and tokenizing the ligand using a pre-trained version of ChemBERTa **Zbiciak2023-lh** available on HuggingFace [13]. Our results indicated that morgan fingerprints as encoding outperformed other techniques, which was unexpected. However, it's plausible that our computational constraints played a part in this finding.

6.2 Target encoding

Our baseline method involved label-encoding the 21 values found in the amino-acid sequences (20 amino acids and 1 X character, representing unknown values). As suggested in [14], we encoded the sequence information into a 512-dimensional vector, with each position indicating the presence or absence of a conjoint-triad characteristic of the sequence. While the vector length in this encoding is typically 343, we incorporated an additional component for the unknown token, thereby increasing the possible permutations to 512. Finally, we also experimented with encoding the protein sequence using ProtBert [15], a version of the BERT transformer trained specifically on protein sequences.

6.3 CNNs

Both our architecture reference paper [5] and our results with DeepPurpose indicated that employing convolutional neural networks (CNNs) would enhance the feature engineering phase of the modeling. As a result, we developed two CNN encoding modules to embed the ligand and protein encodings. For network inputs, we selected morgan fingerprints for the SMILES sequences and extracted features using a 1-dimensional CNN. Additionally, we chose a matrix representation of the $20 + 1$ amino acids, normalized to a length of 1200 for the protein sequence. We then extracted features from this representation using a 2-dimensional CNN. We used flattening and simple concatenation to feed the resulting data into the DeepDTA-like fully connected network (FCN).

6.4 DeepPurpose Results

Table 2: DeepPurpose KiBA Results with CNN Protein Encoding

Drug Encoding	MSE	Pearson Correlation	Concordance Index
CNN	0.4615	0.5596	**0.6868**
CNN and RNN	0.5326	0.5334	0.5465
Daylight	0.4455	0.5758	0.6979
Morgan	**0.4455**	**0.5780**	0.6967
MPNN	0.6286	0.3508	0.6680
Pubchem	0.4835	0.5439	0.6838
RDKit_2d_normalized	0.4650	0.5627	0.6902
Transformer	0.6164	0.3615	0.6778

The results reported in Table 2 indicate that morgan fingerprint and CNNs are good drug encodings in our setting.

Table 3: DeepPurpose Davis Results with CNN Protein Encoding

Drug Encoding	MSE	Pearson Correlation	Concordance Index
CNN	**0.5905**	0.6074	0.7894
CNN and RNN	0.6493	0.5695	0.7769
Daylight	0.6362	0.5871	0.7829
Morgan	0.6141	0.5878	0.7829
MPNN	0.8290	0.2959	0.6312
Pubchem	0.6388	0.5745	0.7796
RDKit_2d_normalized	0.6029	**0.6097**	**0.7971**
Transformer	0.8362	0.2657	0.5883

As presented in Table 3, the best-performing drug encoding protein, with the target protein encoded via a CNN, is also a CNN, added together with the KiBa [6] results, this was a deciding factor for implementing our own CNNs.

6.5 Modeling results

Despite experimenting with different strategies, we were unable to outperform DeepPurpose on the downsampled BindingDB dataset, Table 4 summarizes our findings.

Table 4: Modeling results on 10% BindgingDB

Encoding	MSE	Pearson Correlation	Concordance Index
Morgan and Conjoint	0.516	2.042	0.249
Morgan and ProtBert	0.566	0.724	0.443
CNN and CNN	**0.502**	**0.582**	**0.694**

We obtained the best results using CNNs.

7. Deployment

We tested our DeepPurpose models on three different benchmark datasets. We consider DAVIS too small to define an applicability domain, but the other two are challenging and diverse, as shown in our codebase [11]. Protein length can be a factor, with our training data being mostly in the 0-1200 amino-acid sequence range. For our in-house models, the sample of the BindingDB dataset we worked on is comparable to KiBA and the same general considerations apply. In particular, for the CNN model, we decided to truncate protein sequences longer than 1200 amino acids in the encoding phase, in this case, the applicability domain is quite rigid.

8. Limitations

Incorporating heterogeneous meta-data to enrich protein sequences with annotations is a common practice we have seen, which would likely increase the predictive power of our models [7]. Due to hardware limitations, we were only able to test our models on comparatively small datasets, which biases our analysis towards simpler encoding schemes. We did not perform more specific analyses on sub-groups of protein targets and ligand drugs, which might have yielded insights into specific classes of DTI.

9. Conclusions

During the development and implementation, we were able to apply our newly acquired knowledge to the DTI task. We tested different strategies, including some more advanced methods we had not explored before. We enjoyed the challenge posed by the DTI task, although we did not manage to obtain the modeling results we were aiming for.

References

[1] Q. Ye, C.-Y. Hsieh, Z. Yang, *et al.*, "A unified drug–target interaction prediction framework based on knowledge graph and recommendation system," *Nature Communications*, vol. 12, p. 6775, 2021.

[2] Y. Li, K. Hsieh, R. Lu, *et al.*, *Glam: An adaptive graph learning method for automated molecular interactions and properties predictions*, Dec. 2021. DOI: 10.21203/rs.3.rs-1172418/v1.

[3] M. Yazdani-Jahromi, N. Yousefi, A. Tayebi, *et al.*, "AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification," *Briefings in Bioinformatics*, vol. 23, no. 4, Jul. 2022, bbac272, ISSN: 1477-4054. DOI: 10.1093/bib/bbac272. eprint: https://academic.oup.com/bib/article-pdf/23/4/bbac272/45017994/bbac272.pdf. [Online]. Available: https://doi.org/10.1093/bib/bbac272.

[4] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, "Machine learning approaches and databases for prediction of drug–target interaction: a survey paper," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 247–269, Jan. 2020, ISSN: 1477-4054. DOI: 10.1093/bib/bbz157. eprint: https://academic.oup.com/bib/article-pdf/22/1/247/35935006/bbz157.pdf. [Online]. Available: https://doi.org/10.1093/bib/bbz157.

[5] H. Öztürk, E. Ozkirimli, and A. Ozgur, "Deepdta: Deep drug-target binding affinity prediction," *Bioinformatics*, vol. 34, Jan. 2018. DOI: 10.1093/bioinformatics/bty593.

[6] J. Tang, A. Szwajda, S. Kumar Shakyawar, *et al.*, "Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis," *Journal of chemical information and modeling*, vol. 54, Feb. 2014. DOI: 10.1021/ci400709d.

[7] K. Y. Gao, A. Fokoue, H. Luo, A. Iyengar, S. Dey, and P. Zhang, "Interpretable drug target prediction using deep neural representation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, pp. 3371–3377. DOI: 10.24963/ijcai.2018/468. [Online]. Available: https://doi.org/10.24963/ijcai.2018/468.

[8] M. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic acids research*, vol. 44, Oct. 2015. DOI: 10.1093/nar/gkv1072.

[9] T. Pahikkala, A. Airola, S. Pietilä, *et al.*, "Toward more realistic drug–target interaction predictions," *Briefings in Bioinformatics*, vol. 16, no. 2, pp. 325–337, Apr. 2014, ISSN: 1467-5463. DOI: 10.1093/bib/bbu010. eprint: https://academic.oup.com/bib/article-pdf/16/2/325/681876/bbu010.pdf. [Online]. Available: https://doi.org/10.1093/bib/bbu010.

[10] M. Davis, J. Hunt, S. Herrgard, *et al.*, "Comprehensive analysis of kinase inhibitor selectivity," *Nature Biotechnology*, vol. 29, no. 11, pp. 1046–1051, 2011. DOI: 10.1038/nbt.1990. [Online]. Available: https://doi.org/10.1038/nbt.1990.

[11] C. Grigioni, C. Pala, and A. Wey, *Github project repository*, https://github.com/ChristianPala/protein_target_affinity, Accessed on May 15th, 2023.

[12] G. L. et al., *Rdkit/rdkit: 2023*$_0$3$_1$(*Q*12023)*Release* − − − *zenodo.org*, https://zenodo.org/record/7880616, [Accessed 21-May-2023], 2017.

[13] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[14] H. Wang and X. Hu, "Accurate prediction of nuclear receptors with conjoint triad feature," *BMC Bioinformatics*, vol. 16, no. 1, Dec. 2015. DOI: 10.1186/s12859-015-0828-1. [Online]. Available: https://doi.org/10.1186/s12859-015-0828-1.

[15] A. Elnaggar, M. Heinzinger, C. Dallago, *et al.*, "Prottrans: Towards cracking the language of life's code through self-supervised learning," *bioRxiv*, 2021. DOI: 10.1101/2020.07.12.199554. eprint: https://www.biorxiv.org/content/early/2021/05/04/2020.07.12.199554.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2021/05/04/2020.07.12.199554.