TranDTA: Prediction Of Drug–Target Binding Affinity Using Transformer Representations

Mahsa Saadat[1], Armin Behjati[1], Fatemeh Zare-Mirakabad[1*], Sajjad Gharaghani[2]

[1] Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran

[2] Laboratory of Bioinformatics and Drug Design, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

***Corresponding author:** Fatemeh Zare-Mirakabad

Assistant Professor

Department of Mathematics and Computer Science

Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran

Telephone: +982164545674

Email: f.zare@aut.ac.ir

Abstract

Drug discovery is generally difficult, expensive and the success rate is low. One of the essential steps in the early stages of drug discovery and drug repurposing is identifying drug target interactions. Although several methods developed use binary classification to predict if the interaction between a drug and its target exists or not, it is more informative and challenging to predict the strength of the binding between a drug and its target. Binding affinity indicates the strength of drug-target pair interactions. In this regard, several computational methods have been developed to predict the drug-target binding affinity. With the advent of deep learning methods, the accuracy of binding affinity prediction is improving. However, the input representation of these models is very effective in the result. The early models only use the sequence of molecules and the latter models focus on the structure of them. Although the recent models predict binding affinity more accurate than the first ones, they need more data and resources for training.

In this study, we present a method that uses a pre-trained transformer to represent the protein as model input. Although pretrained transformer extracts a feature vector of the protein sequence, they can learn structural information in layers and heads. So, the extracted feature vector by transformer includes the sequence and structural properties of protein. Therefore, our method can also be run without limitations on resources (memory, CPU and GPU). The results show that our model achieves a competitive performance with the state-of-art models.

Data and trained model is available at http://bioinformatics.aut.ac.ir/TranDTA/ .

Keywords:

Drug-target interaction; Drug discovery; Protein representation; Drug repurposing; Computational pharmacology

Introduction

Drugs are often developed to target proteins that are involved in many cellular processes. Binding affinity indicates the strength of drug-target pair interactions. Through binding, drugs can have a positive or negative effect on the functions performed by proteins and affect disease conditions. Since the high binding affinity between a small molecule and a target protein is one of the criteria for selecting a candidate compound in drug discovery[1], the drug-target binding affinity (DTBA) prediction has received much attention in recent years. Furthermore, repurposing existing drugs and identifying novel interacting partners for approved drugs is also gaining interest in drug discovery[2],[3]. The time cost and financial expenses of directly measuring the binding affinity through experimental methods is extremely high. Therefore, there is a need to develop computational models for accurate prediction of binding affinity.

So far, several computational models have been developed to predict the binding affinity of drug-target pairs. "KronRLS"[4] and “SimBoost”[5] are two machine learning models that use the “guilt by association” rule. Using this rule is based on the assumptions that similar drugs tend to interact with similar targets and similar targets are targeted by similar drugs[6].

These two models manually extract the features of target and drug using drug and target similarities. Manually appropriate feature extraction requires a great deal of biological pre-knowledge and intuition about the relationship between drugs (or targets) and different types of biological data as features[6]. In addition, some of these features may be highly correlated and lead to data redundancy. Feature extraction methods such as auto-encoder, transformer are developed using deep learning to overcome these challenges[7].
“DeepDTA”[8] was the first deep learning model for predicting DTBA which use drug SMILES (Simplified Molecular Input Line Entry System) and protein sequence as inputs. DeepDTA uses integers for each character in categories to represent inputs[8]. Therefore, DeepDTA is a character-based model while in both protein sequence and drug SMILES, it is important to consider dependencies between neighbor characters. In the other words, this model has a character-based approach while amino acid dependence is not considered efficiently because there is relationship between amino acids in the structure of the molecule. The use of word-based approach makes more attention to the dependencies between neighboring words and dependencies between neighboring characters are included in the model by default.

Hence, the “WideDTA”[9] was introduced, which is a word-based approach. WideDTA also use drug SMILES and protein sequence as inputs but it assigned a unique integer to the words that is extracted from the sequences[9].

DeepDTA and WideDTA use two convolutional neural networks (CNN) blocks for feature extraction from drug and protein respectively and finally the two representations being concatenated to predict the binding affinity[8], [9].

Although WideDTA[9] performed better than DeepDTA[8], these models use sequences as inputs while the useful biological information for DTBA lies in drug and target structures. This is because of the fact that binding affinity is defined based on the structure of molecules[10].

Therefore, later studies such as "GraphDTA"[11] tried to use structural information[12], [13]. In GraphDTA[11], drug compounds are represented as graphs where atoms are nodes and bonds are edges. The construction of these graphs and atom feature extraction process are implemented using the RDKit[14]. In GraphDTA[11], simililar to DeepDTA, CNN was used to extract protein sequence representation. For drugs, graph neural networks were implemented on the molecular graph. Then the two representations being concatenated and ended by a regression layer to estimate the output as the drug-target affinity value[11]. GraphDTA[11] performs better than DeepDTA[8] and WideDTA[9] by using graph representations for drugs in DTBA prediction.

Since in "GraphDTA[11]", CNN was used to obtain protein features through the sequence, "DGraphDTA[12]" was introduced which not only constructed the graph of the drug molecule, but also constructed the protein graph. This model used the protein graph based on the contact map of the protein. In this model, both small drug molecule and protein are encoded using GNN and it improved performance.

Although deep learning models show great performance improvement in DTBA prediction, there are three main challenges as follows:

(i) The use of structural information in these models increased the cost of time and resources.

(ii) While the information that can be obtained from the 3D structure is very valuable in DTBA prediction, these studies depend on the availability of protein-ligand complex structures[9]. In some cases, information about the structure is not available and this information needs to be predicted. The use of predicted information backpropagate the error rate on the main issue.

(iii) It is difficult to analyze deep learning models because of their black-box characteristic[15].

In this paper, we propose TranDTA to predict the drug–target binding affinity. We use transformers for protein feature extraction because they can learn structural information in layers and heads. According to [15], reviews of BERT[16] models on proteins shows that they can consider the structure in their layers and heads (for example the attention in head 7-1 targets binding sites[15]). So without using the structure directly, our model is extracting features from the sequence which are appropriate representation of the structure. ProtAlbert[17] is an extended model of BERT[16] that is pretrained with proteins.

In TranDTA, protein sequences are encoded by ProtAlbert[17]. Then the protein vectors are concatenated with drugs which are represented as molecular fingerprint[18], to predict the binding affinity of drug-target pairs. We test the proposed method on KIBA[19], which is a benchmark

dataset and compare the performance with state-of-the-art approaches in the field. The results show that our proposed method achieves a competitive performance in DTBA prediction task without relying on 3D structure information of the complex or 2D representation of the compound.

In the following, we first introduce our proposed model and the used dataset. Then, in the next section, we explain the implementation details and compare the results with state-of-the art models.

Materials and methods

In this section, first we represent the input and output of drug target binding affinity (DTBA) problem. In DTBA, the sequence of protein P and molecular fingerprint[18] of drug d are given as inputs. A real value is generated as an output to show the binding affinity between target P and drug d.

In the following, we introduce TranDTA and the dataset that we evaluate our proposed model on.

Input representations

We use molecular fingerprint[18] representations to represent drug inputs. In cheminformatics, molecular fingerprints are one of the most common representations of chemical structures. The advantage of this representation is that it increases the calculation speed and reduces the storage space[20]. In molecular fingerprint representation each drug, such as d, is represented as an 881-length vector, as follows:

$$F^d = [f_1^d \dots f_{881}^d]$$

where $f_i^d \in \{0,1\}$ to show the absence (0) or presence (1) of a substructure descriptor associated with a specific molecular feature which is predetermined in the design of the fingerprint.

We use sequence of protein for target inputs. The sequence of protein P is defined as follows:

$$S^P = s_1^P \dots s_n^P$$

where s_i^P shows one out of 20 types of amino acid and n is the length of the protein sequence.

It has recently been shown that encoded protein sequence representations from transformers-based models can capture the biophysical and structural features of the original sequence[15].

We used ProtAlbert [16] as a protein sequence encoder. A transformer-based language model with 12 layers and 64 attention heads that was trained on Uniref100 dataset that includes 216 million protein sequences.

ProtAlbert[17] representations which generates 4096-length vectors for the protein sequence as follows:

$$E^P = [e_1^P \dots e_{4096}^P]$$

where $e_i^d \in \mathbb{R}$ and $\mathbb{R}$ is real number set.

TranDTA

The DTBA problem is considered as a regression model. In this study, we propose a sequence-based model, which we call TranDTA. Figure 1 shows TranDTA architecture. In TranDTA, we use a pre-trained model (ProtAlbert[17]) to transform the proteins sequences into feature vectors. Then, they are concatenated to the molecular fingerprint[18] vectors of drugs and fed into 5 fully connected layers to predict the binding affinity value. We used 2048 nodes in the first FC layers. The next layers have 1024, 512 and 256 nodes, respectively.

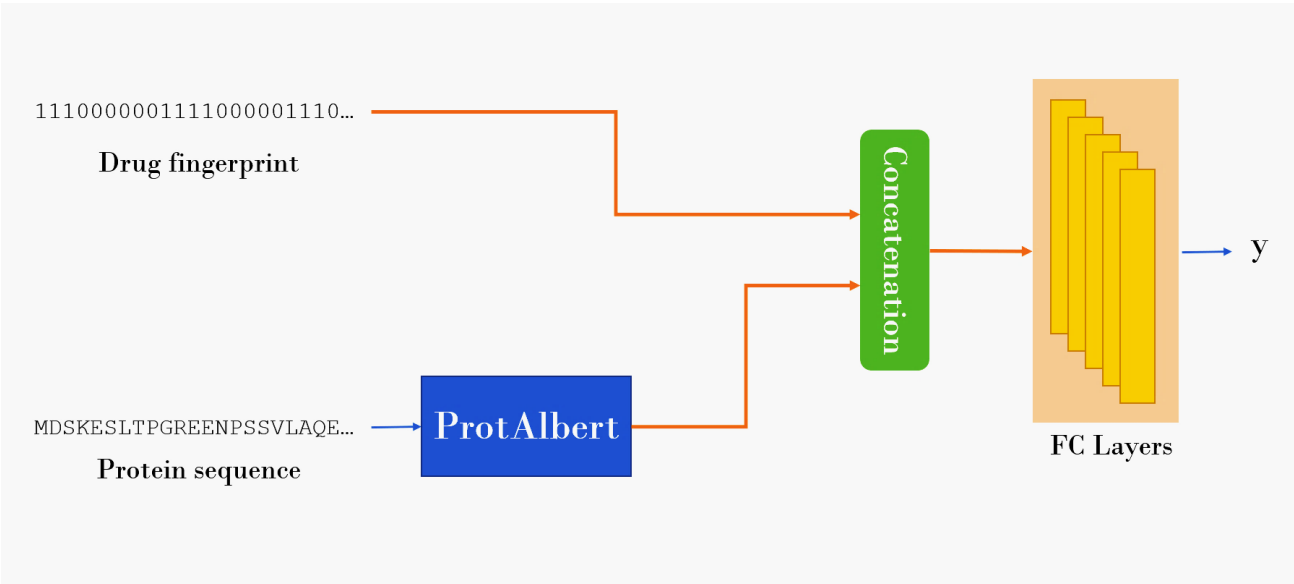

Figure 1 TranDTA architecture.

Dataset

We evaluate our proposed model on KIBA[19] dataset, which has been popularly used as benchmark for binding affinity prediction assessments[4], [5], [8], [9]. The KIBA dataset is one of the large scale biochemical selectivity assays of the kinase inhibitors and originated from an approach called KIBA, in which kinase inhibitor bioactivities from different sources such as K_i, K_d, and IC_{50} were combined[19]. KIBA scores were constructed to optimize the consistency between K_i, K_d, and IC_{50} by utilizing the statistical information they contained. In the KIBA dataset, the lower KIBA score corresponds to the higher binding affinity.

The original KIBA dataset contains a matrix of 467 proteins, and 52,498 drugs, with 246,088 interactions[19]. We used the filtered version of the KIBA dataset, in which each protein and ligand has at least 10 interactions[5]. As a result, this dataset includes 229 unique proteins, 2,111 unique drugs, and 118,254 interactions.

Similar to DeepDTA[8], the KIBA scores pre-processed as follows: (i) negative of each KIBA score was taken, (ii) the minimum value among the negatives was chosen, and (iii) the absolute value of the minimum was added to all negative scores,thus the final form of the KIBA scores are constructed[8] and its value ranges from 0.0 to 17.2.

Experiments and results

In this part, we first describe the Implementation Details. Then, we introduce used evaluation metrics to compare our proposed model with state-of-the art models.

Implementation Details

We used the Python programming language. Our experiments were run on Google Colaboratory[21]. The proposed framework was implemented using PyTorch[22] backend and ADAM optimization.

For drugs, we used Pubchem[23] to represent the molecular fingerprint and it is contained data for only 2065 of KIBA[19] drugs. ProtAlbert[17] needed more than 32 gigabytes of RAM to run on protein sequences longer than 1000, which was not available to us. For this reason, we considered protein sequences with a length of less than 1000 in the dataset, which is 185 proteins.

Due to limited resources, we used a sample set of 1,512 interactions to train and test. We used 20% of the data for testing and the rest for training. More details are shown in Table 1.

No. Drugs	No. Proteins	No. Interactions	Split	No. samples
38	163	1512	Train	1209
			Test	303

Table 1- details of preprocessed KIBA dataset.

To construct the sample set of interactions, we used the systematic random sampling method in which the sampling interval calculated by dividing the entire population size by the desired sample size. We also used the Cochran's formula[24] to calculate the sample size (error rate = 0.025 , Z = 1.96). Therefore, the sample set size is equal to 1,512. Figure 2 shows the total data distribution in the dataset and the data distribution in the sample set.

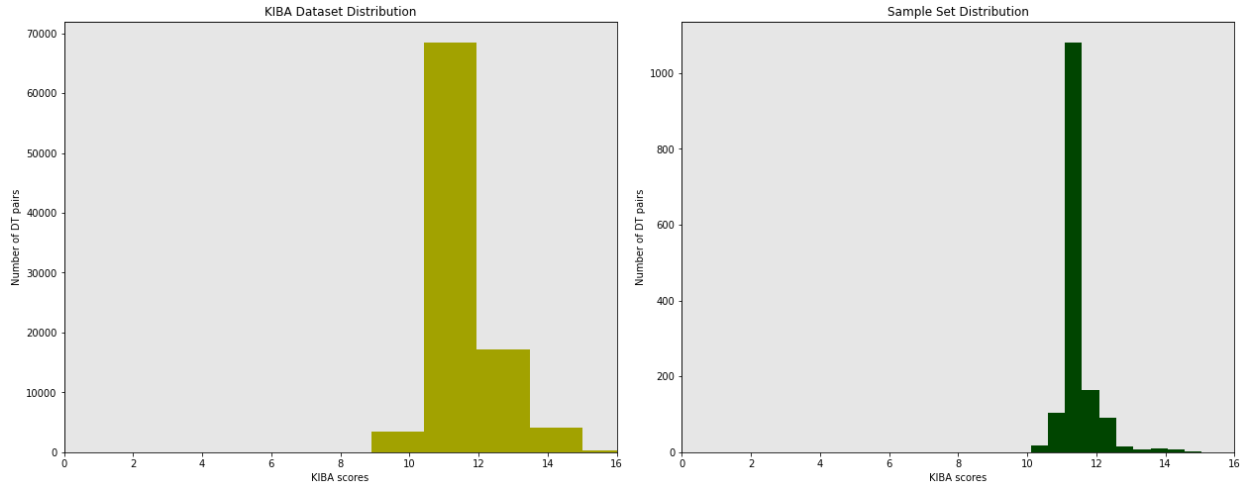

Figure 2 KIBA dataset and TranDTA sample set distributions.

We achieved a high performance of the proposed model with a relatively small range on hyperparameter tuning. The detailed settings are summarized in Table 2.

Hyperparameter	Setting
Epoch	55
Batch size	1
Learning rate	0.003
Optimizer	Adam
Number of Hidden neurons	2048;1024; 512; 256
Activation function	Rectified Linear Unit (ReLU)
Loss function	Root Mean Square Error (RMSE)

Table 2- details of hyperparameter settings.

Evaluation Metrics

In this study, we compare the performance of KronRLS[4], SimBoost[5], DeepDTA[8], WideDTA[9], GraphDTA[11], DGraphDTA[12] and TranDTA (our proposed model), using the KIBA[19] dataset for evaluation. We use three metrics such as CI[25], MSE and RMSE for the evaluation of the performance in these models.

- **Concordance Index (CI)**

  As suggested in [4], the CI can be used as an evaluation metric for the prediction accuracy in DTBA. CI[25] is a ranking metric for continuous values. The intuition behind the CI is whether the predicted binding affinity values of two random drug-target pairs were predicted in the same order as their actual values were or not. The CI ranges between 0.5 and 1.0, where 1.0 corresponds to perfect prediction accuracy and 0.5 corresponds to a random predictor. CI is computed as follows:

$$CI = \frac{1}{Z} \sum_{s_i > s_j} h(b_i - b_j)$$

  where b_i is the prediction value for the larger affinity s_i, b_j is the prediction value for the smaller affinity s_j, Z is a normalization constant that equals the number of data pairs with different label values, and $h(x)$ is the Heaviside step function[4]. It is a discontinuous function and defined as:

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

- **Mean Square Error (MSE)**

  The MSE is a commonly used metric for the error in continuous prediction. It used in regression task to measure how close the fitted line, that is represented by connecting the

estimated values, is to the actual data points. Since DTBA is a regression task, we use the MSE as a metric:

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

where y_i is the actual output, p_i corresponds to the prediction and n indicates the number of samples.

- **Root Mean Square Error (RMSE)**

  RMSE is one of regressor metrics. It is the distance, on average, of data points from the fitted line and computed as the square root of MSE:

$$RMSE = \sqrt[2]{MSE}$$

Comparison Results

In this study, we propose a drug-target binding affinity prediction model, named TranDTA, based on only sequence information of drugs and proteins. Table 3 shows the average CI, MSE and RMSE values for KIBA[19] dataset.

Model	CI	MSE	RMSE
KronRLS	0.782	0.411	0.620
SimBoost	0.836	0.222	0.471
DeepDTA	0.863	0.194	0.440
WideDTA	0.875	0.179	0.423
GraphDTA	0.891	0.139	0.372
DGraphDTA	0.904	0.126	0.354
TransDTA (error rate of sampling = 0.025)	**0.911**	**0.110**	**0.333**

Table 3- Prediction performance on KIBA dataset, sorted by MSE.

The results show that the TranDTA performed much better than the similarity-based models (KronRLS[4] and SimBoost[5]) and the sequence-based models (DeepDTA[8] and WideDTA[9]) in all three metrics. It also performed closely with the structure-based models that use graphs (GraphDTA[11] and DGraphDTA[12]) and was slightly better in this experiment.

Our model is competitive not only with sequence-based models[8], [9] but also with models that use structural information[11], [12]. In addition, in our experience, TranDTA costs less time and calculations than other models. This is because TranDTA does not learn input representations when model is training, it uses pretrained models for input representations. This reduces computation and speeds up. Therefore, this model can be trained and tested without limitations on resources (memory, CPU and GPU).

Conclusion

So far, computational DTBA prediction models have been extracting features by using deep learning approaches. First sequence-based models and then graph-based models were introduced. Although the use of graph performs better in predicting, it requires more resources to execute. Also, deep learning models that have been introduced so far, learn the representations of inputs during training.

In this paper, we proposed the novel method, named TranDTA, for the problem of predicting DTBA. There are new types of transformers that are pre-trained on proteins and can use sequences to show structural features on their heads and layers.

To the best of our knowledge, TranDTA is the first model that applies transformers to extract features of protein sequence and uses transformer representations in DTBA prediction. Experimental results show that TranDTA outperforms other existing sequence-based methods in prediction performance on the KIBA dataset. Moreover, it performed closely with the structure-based models and was slightly better in this experiment.

Because of the success of transformers in NLP and the results of this study, we believe that TranDTA is an effective approach for DTBA prediction and can be quite helpful in drug development process.

As future work, there is a need to fine-tune TranDTA on the dataset. In addition, to improve TranDTA, a specific transformer can be used to represent the drug features.

References

[1] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug," pp. 1239–1249, 2011, doi: 10.1111/j.1476-5381.2010.01127.x.

[2] T. I. Oprea and J. Mestres, "Drug Repurposing : Far Beyond New Targets for Old Drugs," vol. 14, no. 4, pp. 759–763, 2012, doi: 10.1208/s12248-012-9390-1.

[3] M. Moridi, M. Ghadirinia, A. Sharifi-zarchi, and F. Zare-mirakabad, "The assessment of efficient representation of drug features using deep learning for drug repositioning," pp. 1–11, 2019.

[4] T. Pahikkala *et al.*, "Toward more realistic drug-target interaction predictions," *Brief. Bioinform.*, vol. 16, no. 2, pp. 325–337, Mar. 2015, doi: 10.1093/bib/bbu010.

[5] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester, "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines," *J. Cheminform.*, vol. 9, no. 1, p. 24, Dec. 2017, doi: 10.1186/s13321-017-0209-z.

[6] M. Thafar, A. Bin Raies, S. Albaradei, M. Essack, and V. B. Bajic, "Comparison Study of Computational

Prediction Tools for Drug-Target Binding Affinities," *Front. Chem.*, vol. 7, no. November, pp. 1–19, Nov. 2019, doi: 10.3389/fchem.2019.00782.

[7] B. Tang, Z. Pan, K. Yin, and A. Khateeb, "Recent advances of deep learning in bioinformatics and computational biology," *Front. Genet.*, vol. 10, no. MAR, pp. 1–10, 2019, doi: 10.3389/fgene.2019.00214.

[8] H. Öztürk, E. Ozkirimli, and A. Özgür, "DeepDTA: Deep Drug-Target Binding Affinity Prediction," *Bioinformatics*, vol. 35, no. 18, pp. 3329–3338, Jan. 2018, doi: 10.1093/bioinformatics/bty593.

[9] H. Öztürk, E. Ozkirimli, and A. Özgür, "WideDTA: prediction of drug-target binding affinity," Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.04166.

[10] D. S. Induced-fit-theorie, D. E. Koshland, and G. Schlussel-schlor-prinzips, "Das Schussel-SchloQ-Prinzip und die Induced-fit-Theorie" *," no. August, 1994.

[11] T. Nguyen, H. Le, and S. Venkatesh, "GraphDTA: prediction of drug–target binding affinity using graph convolutional networks," *bioRxiv*, p. 684662, 2019, doi: 10.1101/684662.

[12] M. Jiang *et al.*, "Drug–target affinity prediction using graph neural network and contact maps," *RSC Adv.*, pp. 20701–20712, 2020, doi: 10.1039/d0ra02297g.

[13] S. Wang, P. Shan, Y. Zhao, and L. Zuo, "GanDTI: A multi-task neural network for drug-target interaction prediction," *Comput. Biol. Chem.*, vol. 92, p. 2021, 2021, doi: 10.1016/j.compbiolchem.2021.107476.

[14] G. Landrum, "RDKit: Open-source cheminformatics," [Online]. Available: http://www.rdkit.org.

[15] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, "BERTology meets biology: Interpreting attention in protein language models," *bioRxiv*, 2020, doi: 10.1101/2020.06.26.174417.

[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[17] A. Elnaggar *et al.*, "ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing," *bioRxiv*, 2020, doi: 10.1101/2020.07.12.199554.

[18] R. Todeschini and V. Consonni, "Handbook of Molecular Descriptors," *Wiley-VCH*, vol. 11, p. 688, 2000.

[19] J. Tang *et al.*, "Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis," *J. Chem. Inf. Model.*, vol. 54, no. 3, pp. 735–743, Mar. 2014, doi: 10.1021/ci400709d.

[20] E. Fernández-De Gortari, C. R. García-Jacas, K. Martinez-Mayorga, and J. L. Medina-Franco, "Database fingerprint (DFP): an approach to represent molecular databases," *J. Cheminform.*, vol. 9, no. 1, pp. 1–9, 2017, doi: 10.1186/s13321-017-0195-1.

[21] T. Carneiro, R. V. M. Da Nobrega, T. Nepomuceno, G. Bin Bian, V. H. C. De Albuquerque, and P. P. R. Filho, "Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications," *IEEE Access*, vol. 6, pp. 61677–61685, 2018, doi: 10.1109/ACCESS.2018.2874767.

[22] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *arXiv*, no. NeurIPS, 2019.

[23] S. Kim *et al.*, "PubChem in 2021 : new data content and improved web interfaces," no. 1, pp. 1–8, 2021, doi: 10.1093/nar/gkaa971.

[24] W. G. Cochran, *Sampling Techniques*, Third. New York, New York, USA: John Wiley & Sons, 1977.

[25] G. Heller and M. Gönen, "Concordance Probability and Discriminatory Power in Proportional Hazards Regression," *Biometrika*, vol. 92, no. 4, pp. 965–970, 2005.