Drug–target affinity prediction using applicability domain based on data density

Shunya Sugita
Department of Computer Science
School of Computing
Tokyo Institute of Technology
Kanagawa, Japan
sugita@li.c.titech.ac.jp

Masahito Ohue
Department of Computer Science
School of Computing
Tokyo Institute of Technology
Kanagawa, Japan
ohue@c.titech.ac.jp

***Abstract*—In the pursuit of research and development of drug discovery, the computational prediction of the target affinity of a drug candidate is useful for screening compounds at an early stage and for verifying the binding potential to an unknown target. The chemogenomics-based method has attracted increased attention as it integrates information pertaining to the drug and target to predict drug–target affinity (DTA). However, the compound and target spaces are vast, and without sufficient training data, proper DTA prediction is not possible. If a DTA prediction is made in this situation, it will potentially lead to false predictions. In this study, we propose a DTA prediction method that can advise whether/when there are insufficient samples in the compound/target spaces based on the concept of the applicability domain (AD) and the data density of the training dataset. AD indicates a data region in which a machine learning model can make reliable predictions. By preclassifying the samples to be predicted by the constructed AD into those within (In-AD) and those outside the AD (Out-AD), we can determine whether a reasonable prediction can be made for these samples. The results of the evaluation experiments based on the use of three different public datasets showed that the AD constructed by the k-nearest neighbor (k-NN) method worked well, i.e., the prediction accuracy of the samples classified by the AD as Out-AD was low, while the prediction accuracy of the samples classified by the AD as In-AD was high.**

***Index Terms*—drug–target affinity prediction, chemogenomics, applicability domain, data density, k-nearest neighbor**

I. Introduction

In drug discovery research and development, computational prediction of the target affinity of a drug candidate is useful for screening compounds at the early stage and for the verification of the binding potential to an unknown target. In particular, a prediction method based on machine learning that utilizes information on both the drug and the target can rapidly and comprehensively predict the affinity between the drug and the target [1].

Convolutional neural network (CNN)-based architectures have been employed in deep-learning studies for the prediction of affinity values of drug–target pairs with the use of detailed information obtained from the three-dimensional (3D) structure of protein–ligand complexes [2], [3]. The information obtained from 3D structures can provide a good representation of the structural mechanisms of drug–target interactions, but these studies depend on the availability of the complex structural data.

978-1-6654-0112-8/21/$31.00 ©2021 IEEE

Conversely, chemogenomics [1], which integrates information on both drug and target to predict the drug–target affinity (DTA), does not necessarily require the 3D structure. Chemogenomics methods are attractive because they can predict interactions for unvalidated drug–target pairs from existing drug–target interaction data. Thus, chemogenomics is useful for drug repositioning/repurposing [4]. Chemogenomics methods have been traditionally approached as binary classification problems [1], [5]–[9]. However, with the increase in the number of available data and the improvement in the prediction performance of machine learning, the approach of predicting DTA as a regression problem has recently become popular [10], [11]. SimBoost [10], proposed in 2017, is a method used for the prediction of DTA with the use of a gradient boosting method with features derived from drug-to-drug and target-to-target similarities. This method has achieved high-prediction accuracy.

Conversely, the compound space is huge, and machine learning predictions for drug–target pairs are not always reliable. In fact, in the field of quantitative structure–activity relationship (QSAR), the applicability domain (AD), a region in the compound space wherein QSAR can make reasonable predictions, has been actively studied [12]. QSAR is a prediction and analysis method used for the identification of relationships between chemical structures and biological activities, a field that has been developed and advanced to discover better chemicals [13]. AD is the region wherein a machine learning model can perform reasonably well. By quantifying the distance from the model to the sample, it is possible to estimate the region in which the predictive model can make reliable predictions. AD allows one to estimate the uncertainty of prediction for unknown samples to avoid model misuse.

One of the methods proposed to introduce AD in DTA prediction problems is SimBoostQuant [10], which is a derivative of SimBoost. SimBoostQuant determines the AD from the standard error and confidence interval when performing cross-validation. This method is called an ensemble learning-based method. However, AD based on ensemble learning is

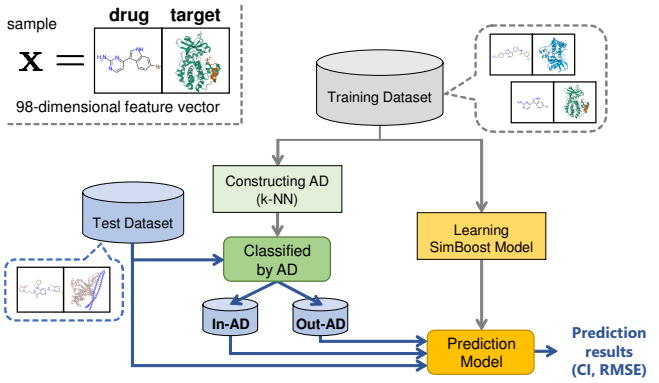

Fig. 1: Application domain (AD)-based prediction methods and performance evaluation.

still problematic in that in instances in which the training data are not informative, i.e., in instances in which the data density is low, the confidence interval itself becomes less reliable. Constructing ADs by taking into account the data density is important for building reliable ADs [14].

In this study, we proposed a DTA prediction method that considers the AD, which is determined by the data density of the training data. We adopted the k-nearest neighbor (k-NN) method, which is an intuitive method using data density as the method to obtain AD. By constructing the AD, it is possible to estimate in advance whether the sample to be predicted will be inside or outside the AD. Our results show that the reliability of the prediction results differs depending on whether the sample is inside or outside the AD.

II. Materials and Methods

An overview of the dataset partitioning, training, AD construction, and inference performed in this study is shown in Fig. 1. The evaluation accuracy was estimated based on the entire test data, samples assigned as In-AD by the AD, and samples assigned as Out-AD by the AD, respectively.

A. Dataset

To evaluate the accuracy of the proposed method, we used the three datasets, Davis [15], Metz [16], and kinase inhibitors bioactivity (KIBA) [17], shown in Table 1, as well as in the publication on SimBoost [10]. These have labels for target activity values (real values) such as IC_{50}. Note that except for the Davis dataset, not all drug-target pairs have labels for their activity values (See Table 1).

The Davis dataset contains binding affinities for all pairs of 68 drugs and 442 targets as measured by K_d values (dissociation constants). The data were converted to pK_d values by the conversion equation $pK_d = -\log(K_d \cdot 10^{-9})$ according to previous studies [10]. The Metz dataset contains the logarithmic values of the enzyme inhibition constants EC_{50}. The KIBA dataset contains a combination of biological activities of enzyme inhibitors from different sources, such as

TABLE I: Drug–target affinity (DTA) datasets used in this study. The values in parentheses indicate the ratio to the product of the number of drugs and the number of targets.

Dataset	#Drug	#Target	#sample
Davis [15]	68	442	30,056 (100%)
Metz [16]	1,421	156	93,326 (42.1%)
KIBA [17]	2,116	229	118,234 (24.4%)

K_i, K_d, IC_{50}, which are unified in the form of the KIBA score [17].

Each dataset was divided into training and test datasets at the ratio of $4:1$, and the training dataset was used to train the prediction model and construct the AD. The split of samples into training and test datasets was also the same as in the publication of He *et al.* [10].

B. Learning prediction models

SimBoost [10] was used to train the predictive model for the DTA. The implementation used the publicly available code [18]. The features were also the same as in the publication of He *et al.* [10], and 98-dimensional feature vectors $\mathbf{x}$ were generated for each sample (drug-target pair). The hyperparameters were the values optimized in the publication of He *et al.* [10].

C. Construction of AD

To estimate in advance whether the prediction model covers the samples in the test dataset, this study utilized AD. k-NN and one-class support vector machine (OCSVM) were used in this study to obtain AD based on data density. In both methods, the percentage of Out-AD in the training sample in the dataset was given in advance as ν.

1) k-NN: k-NN is a method used to determine the AD based on data density using the distance between a test sample and k training samples in the neighborhood of its test sample. The distance between the samples was calculated by the Euclidean distance using $\mathbf{w}^\top \hat{\mathbf{x}}$, which is scaled feature vector $\mathbf{x}$ defined according to the RobustScaler method in scikit-learn [19] multiplied by a weight $\mathbf{w}$. $\mathbf{w}$ is the feature importance and is calculated by SimBoost. The feature importance in a decision tree is calculated from the number of samples that reached the node divided by the total number of samples. The larger this value is, the more important the feature is.

2) OCSVM: We also examined AD construction with the use of OCSVM, which is a support vector machine that selects a discriminating hyperplane by considering that all the samples belong to the same class. The optimized hyperplane is defined as the AD, and samples that deviate from the discriminating hyperplane are classified to be outside the AD. The features were processed in the same way as in the k-NN method. According to the related work by Kaneko and Funatsu [20], we used a radial basis function as the kernel function, and its parameter γ was chosen to maximize the variance of the Gram matrix.

D. Performance evaluation

To evaluate the performance of the proposed method, we used the concordance index (CI), which evaluates the correctness of the ordering relationship of the DTA predictions, and the root-mean-squared error (RMSE), which evaluates the overall accuracy of the DTA predictions.

CI is calculated by the following equation,

$$\mathrm{CI} = \frac{\sum_{y_i > y_j} h(\tilde{y}_i - \tilde{y}_j)}{\sum_{y_i > y_j} 1} \tag{1}$$

$$h(x) = \begin{cases} 1 & (x > 0) \\ 0.5 & (x = 0) \\ 0 & (x < 0) \end{cases} \tag{2}$$

where y_i, y_j are the true values of the affinities, and $\tilde{y}_i$, $\tilde{y}_j$ are the predicted values of the affinities.

The RMSE is one of the most common metrics that measures the difference between the predicted and true values, and is calculated by the following equation,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2} \tag{3}$$

where n is the number of samples used in the evaluation.

III. Results and Discussion

A. Prediction accuracy on In-AD and Out-AD

Table II lists the performances of DTA predictions with and without AD on three datasets, namely, Davis [15], Metz [16], and KIBA [17]. In kNN, the parameter k was set to five. This is because the RMSE of the samples of out-AD gets worse for smaller values of k. We will discuss parameter k in a later section III-E.

In all datasets, k-NN determined AD more accurately than OCSVM. The larger the parameter ν (AD is narrower) is, the better the RMSE values are. The RMSE for the data of out-AD was worse than the baseline, suggesting that samples that were difficult to predict were determined to be out-AD. We showed that indicating whether or not a sample will enter AD can provide a more reliable prediction. the CI of baseline and in-AD is similar.

B. Distribution of samples of In-AD and Out-AD

The scatter plots for the samples classified as in-AD and out-AD when $\nu = 0.02$ that obtained the worse RMSE value in Out-AD are shown in Fig. 2. In these scatter plots, the horizontal axis is the predicted value, and the vertical axis is the true value.

Although the Out-AD data should not be able to make valid predictions and the prediction error should be larger, all the scatter plots of the Out-AD data for all the datasets yielded good correlations. This can be attributed to the high accuracy of SimBoost and the high-data density of each dataset. Conversely, the scatter plots on Out-AD have few outliers and yield worse correlations. Based on these, we can conclude that AD improves the reliability of the prediction performance.

C. Ratio of Out-AD of test sample for ν

The ν is the ratio of the training sample that is outside of AD. We examined how the parameter ν classified the test samples. Table III shows the ratio of the samples discriminated as Out-AD in the test sample when each ν (ratio of Out-AD samples in the training samples) is given. The results showed that ν and the ratio of the test sample classified as Out-AD were almost identical. This implies that the samples of the training and test data exhibit similar distributions. In practical use, it is rare that the distribution of the training sample is similar to that of the samples whose affinity we want to know. Therefore, the training data used to determine ν will need to be chosen carefully depending on the target to be applied.

D. Performances at various ν values

As shown in Table IV, the RMSE of In-AD by the k-NN method tends to decrease when ν is increased. For consideration, we examined the case of various ν by the k-NN method ($k = 5$) on the Davis dataset.

The k-NN method can determine the AD that improves the RMSE for any ν. Not only the RMSE but also the CI tended to improve slightly. Given that a large ν indicates a narrower AD region, it is suggested that the narrower the AD is based on the data density, the more reliable is the prediction of the samples in the AD. Within large ADs where $\nu = 0.10$, the number of samples classified to be outside the AD is small. The prediction accuracy of these few out-AD samples is quite poor. Conversely, within the comparatively small AD where $\nu = 0.90$, the prediction accuracy of the samples within the AD is good. Therefore, according to the situation wherein the prediction is to be applied, it is required to set the parameter ν to obtain the desired AD based on considerations whether a loose AD or a strict AD is desired.

E. Performances at various k values

Table V shows the values of RMSE for various values of k in the Davis dataset for In-AD and Out-AD, respectively. From Table V(a), we can say that the value of k has a minor effect on the reliability of the samples in AD between $k = 1$ and 5. Conversely, Table V(b) shows that the RMSE of the samples outside the AD becomes worse as the value of k becomes smaller, thus suggesting that the application area becomes narrower as the value of k becomes larger. However, the RMSE of In-AD is not improved even when k is increased to 10 or 100, thus suggesting that $k = 1$ to 5 is appropriate. For the k-NN-based AD constructed in this study, we recommend that the parameter k is set to five and ν is adjusted by the users according to the actual data and situation of their application.

F. Comparison with related work - SimBoostQuant

The SimBoostQuant model is a method used to set the AD based on the use of quantile regression. This model was proposed together with SimBoost in [10]. SimBoostQuant is effective owing to the high density of the three studied datasets (Davis, Metz, and KIBA). The quantile regression method is not effective when the datasets are sparse because the

TABLE II: Prediction performances for all tested datasets. "Baseline" is the prediction accuracy when the prediction model learned by SimBoost is directly applied to the entire test dataset without AD. Boldface denotes the minimum RMSE value for In-AD in each dataset, and italics denote the worst RMSE value for Out-AD in each dataset.

Method	ν	Davis dataset				Metz dataset				KIBA dataset			
		In-AD		Out-AD		In-AD		Out-AD		In-AD		Out-AD	
		CI	RMSE	CI	RMSE	CI	RMSE	CI	RMSE	CI	RMSE	CI	RMSE
Baseline	-	0.901	0.488	-	-	0.849	0.413	-	-	0.835	0.452	-	-
k-NN	0.02	0.901	0.482	0.873	*0.742*	0.849	0.408	0.834	*0.636*	0.835	0.437	0.791	*0.944*
	0.04	0.899	0.476	0.873	0.705	0.848	0.405	0.846	0.577	0.836	0.429	0.801	0.841
	0.06	0.898	0.476	0.879	0.664	0.847	0.402	0.850	0.560	0.836	0.422	0.810	0.796
	0.08	0.899	0.470	0.878	0.653	0.846	0.399	0.857	0.549	0.836	0.417	0.814	0.749
	0.10	0.900	**0.468**	0.879	0.633	0.846	**0.397**	0.855	0.542	0.834	**0.411**	0.821	0.731
OCSVM	0.02	0.901	0.487	0.903	0.506	0.850	0.411	0.806	0.523	0.834	0.450	0.826	0.560
	0.04	0.902	0.487	0.894	0.511	0.850	0.410	0.825	0.492	0.834	0.447	0.845	0.567
	0.06	0.902	0.488	0.895	0.484	0.851	0.410	0.827	0.464	0.834	0.446	0.839	0.544
	0.08	0.901	0.487	0.907	0.495	0.851	0.410	0.827	0.452	0.834	0.445	0.842	0.531
	0.10	0.901	0.486	0.906	0.501	0.852	0.410	0.828	0.446	0.833	0.445	0.843	0.513

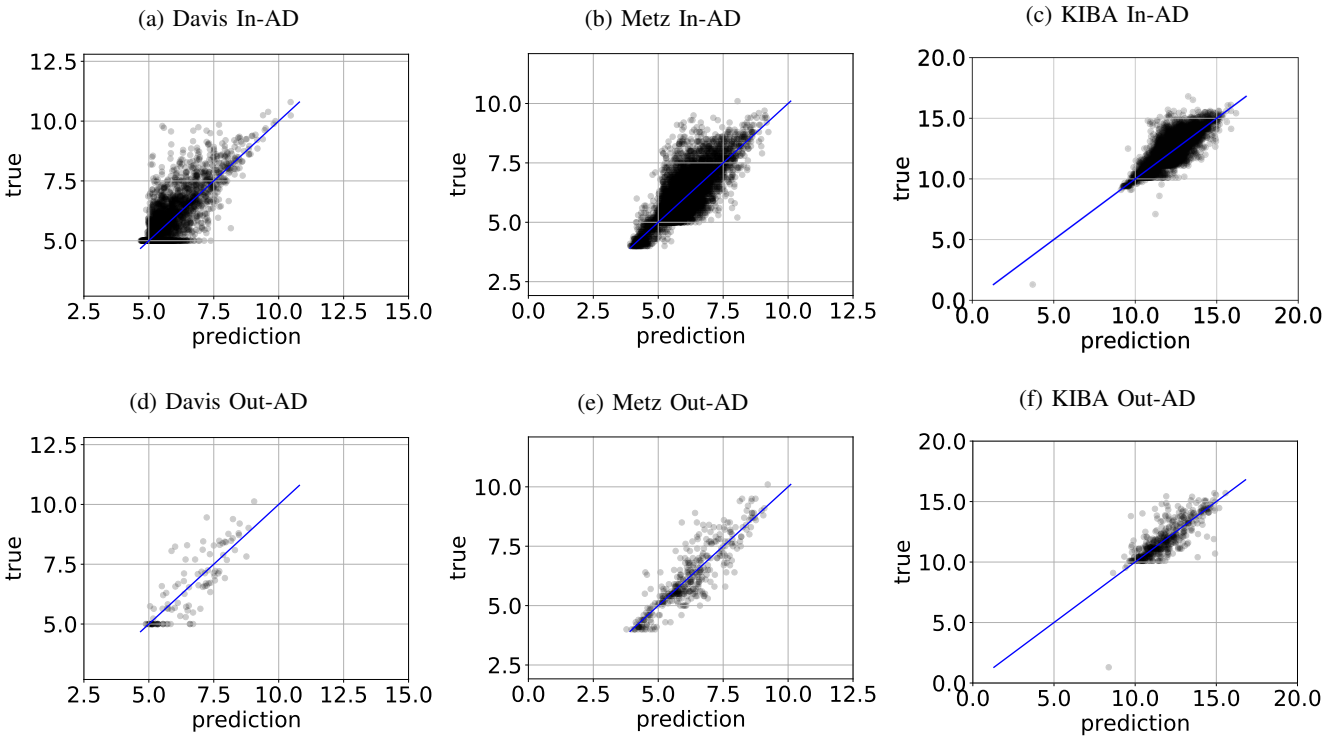

Fig. 2: Scatter plots of three studied datasets ($\nu = 0.02$)

TABLE III: Ratio of test samples that were determined to be outside of the AD for each ν.

ν	Davis	KIBA	Metz
0.02	0.018	0.019	0.020
0.04	0.043	0.039	0.040
0.06	0.055	0.058	0.060
0.08	0.082	0.079	0.080
0.10	0.104	0.098	0.098

prediction intervals in the regions with low-data density are not reliable. Conversely, as pointed out in the literature [10], the SimBoostQuant model does not provide accurate predictions compared with the original SimBoost because the prediction interval is an output, and the prediction accuracy is calculated by the median of the confidence interval. The proposed method is superior as it uses the training data and the model to provide confidence in its capacity to predict results of the test data while using the SimBoost model. However, as shown in Table II, there is no major difference between the prediction accuracy of the entire test data and the data classified as In-AD. However, it should be noted that the situation in the benchmark dataset is somewhat far from reality. It is difficult to maintain a high-data density in actual data in the drug discovery field, and a method that works effectively in low-data density conditions is required. In this sense, the proposed

TABLE IV: Prediction performances for various ν values according to the Davis dataset. "Baseline" is the prediction accuracy when the prediction model learned by SimBoost is directly applied to the entire test dataset without the AD. Boldface denotes the minimum root-mean-square error (RMSE) value for In-AD, and italics denotes the worst RMSE value for Out-AD.

method	ν	In-AD		Out-AD	
		CI	RMSE	CI	RMSE
Baseline	-	0.901	0.488	-	-
k-NN	0.1	0.900	0.468	0.879	*0.633*
	0.2	0.903	0.459	0.879	0.591
	0.3	0.903	0.450	0.885	0.566
	0.4	0.905	0.438	0.888	0.555
	0.5	0.906	0.433	0.891	0.536
	0.6	0.906	0.404	0.894	0.533
	0.7	0.907	0.386	0.896	0.524
	0.8	0.910	0.376	0.897	0.512
	0.9	0.916	**0.324**	0.897	0.502

TABLE V: RMSE of the Davis dataset for various values of k. Boldface denotes the minimum RMSE value for In-AD, and italics denote the worst RMSE value for Out-AD.

(a) In-AD (baseline RMSE = 0.488)

$k \setminus \nu$	0.01	0.02	0.03	0.04	0.05	0.10
1	0.482	0.478	0.477	0.477	0.475	0.469
2	0.484	0.478	0.477	0.476	0.476	**0.468**
3	0.484	0.479	0.477	0.476	0.476	0.469
4	0.485	0.480	0.477	0.475	0.475	**0.468**
5	0.485	0.481	0.476	0.476	0.476	**0.468**
10	0.485	0.482	0.479	0.477	0.477	0.470
100	0.487	0.487	0.485	0.483	0.483	0.472

(b) Out-AD (baseline RMSE = 0.488)

$k \setminus \nu$	0.01	0.02	0.03	0.04	0.05	0.10
1	*0.901*	0.831	0.752	0.708	0.687	0.631
2	0.806	0.831	0.760	0.712	0.672	0.641
3	0.758	0.807	0.756	0.695	0.670	0.634
4	0.733	0.790	0.760	0.709	0.687	0.640
5	0.741	0.746	0.773	0.718	0.675	0.640
10	0.695	0.726	0.708	0.706	0.664	0.624
100	0.607	0.544	0.560	0.546	0.565	0.614

method has merits.

IV. Conclusions

In this study, we proposed a prediction method with the use of AD to improve the accuracy and reliability of drug–target affinity prediction. k-NN and OCSVM were used to construct AD based on data density, and the prediction accuracy was tested on three datasets. We have shown that AD constructed by k-NN can classify samples that were difficult to predict as out-AD.

The three datasets have high densities. If the drug and target spaces covered by the training data become more extensive and sparse, it is unknown how this will affect AD. When using a mixture of several datasets or a database such as BindingDB [21], the compound and target spaces constructed by the training data will be extensive and sparse. In this case, the AD of the data density may work more effectively.

The application to DTA prediction methods other than SimBoost is also important. The concept of AD can also be applied to deep learning, which learns feature representations of samples. Recent methods for DTA prediction have used deep learning [11], [22]–[26], and we would also like to verify AD in the future.

In addition, we may use approaches from other AD constructions, for example, a method [27] that combines ensemble learning and data density. The combination of ensemble learning and data density is effective because ensemble learning works well in regions with a high-training data density [14].

Acknowledgments

This work was partially supported by KAKENHI (Grant No. 20H04280) from the Japan Society for the Promotion of Science (JSPS), ACT-X (Grant No. JPMJAX20A3) from the Japan Science and Technology Agency (JST), and the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) (Grant No. JP20am0101112) from the Japan Agency for Medical Research and Development (AMED). The authors thank Editage (www.editage.com) for English language editing.

References

[1] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug-target interaction prediction: databases, web servers and computational models," *Brief. Bioinform.*, vol. 17, no. 4, pp. 696–712, Jul. 2016, doi: 10.1093/bib/bbv066.

[2] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, "Development and evaluation of a deep learning model for protein-ligand binding affinity prediction," *Bioinformatics*, vol. 34, no. 21, pp. 3666–3674, Nov. 2018, doi: 10.1093/bioinformatics/bty374.

[3] J. Jiménez, M. Škalič, G. Martínez-Rosell, and G. De Fabritiis, "K_{DEEP}: Protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks," *J. Chem. Inf. Model.*, vol. 58, no. 2, pp. 287–296, Feb. 2018, doi: 10.1021/acs.jcim.7b00650.

[4] T. N. Jarada, J. G. Rokne, and R. Alhajj, "A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions," *J. Cheminform.*, vol. 12, no. 1, article no. 46, Dec. 2020, doi: 10.1186/s13321-020-00450-7.

[5] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, Jul. 2008, doi: 10.1093/bioinformatics/btn162.

[6] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput. Biol.*, vol. 8, no. 5, article no. e1002503, May 2012, doi: 10.1371/journal.pcbi.1002503.

[7] T. Ban, M. Ohue, and Y. Akiyama, "Efficient hyperparameter optimization by using Bayesian optimization for drug-target interaction prediction," In *Proc. IEEE ICCABS2017*, Nov. 2017, doi: 10.1109/ICCABS.2017.8114299.

[8] T. Ban, M. Ohue, and Y. Akiyama, "NRLMFβ: Beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug-target interaction prediction," *Biochem. Biophys. Rep.*, vol. 18, article no. 100615, Jul. 2019, doi: 10.1016/j.bbrep.2019.01.008.

[9] M. Ohue, T. Yamazaki, T. Ban, and Y. Akiyama, "Link mining for kernel-based compound-protein interaction predictions using a chemogenomics approach," In *Proc. ICIC2017*, *Lecture Notes in Comput. Sci.*, vol. 10362, pp. 549–558, Jul. 2017, doi: 10.1007/978-3-319-63312-1_48

[10] 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, article no. 24, Dec. 2017, doi: 10.1186/s13321-017-0209-z.

[11] M. Jiang, Z. Li, S. Zhang, S. Wang, X. Wang, Q. Yuan, and Z. Wei, "Drug-target affinity prediction using graph neural network and contact maps," *RSC Adv.*, vol. 10, no. 35, pp. 20701–20712, 2020, doi: 10.1039/D0RA02297G.

[12] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela, and O. Mekenyan O. "A stepwise approach for defining the applicability domain of SAR and QSAR models," *J. Chem. Inf. Model.*, vol. 45, no. 4, pp. 839–849, Jul. 2005, doi: 10.1021/ci0500381.

[13] T. Scior, J. Medina-Franco, Q.-T. Do, K. Martinez-Mayorga, J. Yunes Rojas, and P. Bernard, "How to recognize and workaround pitfalls in QSAR studies: A critical review," *Curr. Med. Chem.*, vol. 16, no. 32, pp. 4297–4313, Nov. 2009, doi: 10.2174/092986709789578213.

[14] H. Kaneko, and K. Funatsu, "Applicability domain based on ensemble learning in classification and regression analyses," *J. Chem. Inf. Model.*, vol. 54, no. 9, pp. 2469–2482, Sep. 2014, doi: 10.1021/ci500364e.

[15] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar, "Comprehensive analysis of kinase inhibitor selectivity," *Nat. Biotechnol.*, vol. 29, no. 11, pp. 1046–1051, Nov. 2011, doi: 10.1038/nbt.1990.

[16] J. T. Metz, E. F. Johnson, N. B. Soni, P. J. Merta, L. Kifle, and P. J. Hajduk, "Navigating the kinome," *Nat. Chem. Biol.*, vol. 7, no. 4, pp. 200–202, Apr. 2011, doi: 10.1038/nchembio.530.

[17] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, and T. Aittokallio, "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.

[18] SimBoost. https://github.com/hetong007/SimBoost/

[19] Scikit-learn, RobustScaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html.

[20] H. Kaneko, and K. Funatsu, "Fast optimization of hyperparameters for support vector regression models with highly predictive ability," *Chemom. Intell. Lab. Syst.*, vol. 142, pp. 64–69, Mar. 2015, doi: 10.1016/j.chemolab.2015.01.001.

[21] M. K. 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 Res., vol. 44, no. D1, pp. D1045–D1053, Jan. 2016, doi: 10.1093/nar/gkv1072.

[22] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug-target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, Sep. 2018, doi: 10.1093/bioinformatics/bty593.

[23] L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, and M. Zheng, "TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments," *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, Aug. 2020, doi: 10.1093/bioinformatics/btaa524.

[24] S. Li, F. Wan, H. Shu, T. Jiang, D. Zhao, and J. Zeng, "MONN: a multi-objective neural network for predicting compound-protein interactions and affinities," *Cell Systems*, vol. 10, no. 4, pp. 308–322, Apr. 2020, doi: 10.1016/j.cels.2020.03.002.

[25] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, "GraphDTA: Predicting drug-target binding affinity with graph neural networks," *Bioinformatics*, Oct. 2020, doi: 10.1093/bioinformatics/btaa921.

[26] K. Wang, R. Zhou, Y. Li, and M. Li, "DeepDTAF: a deep learning method to predict protein-ligand binding affinity," *Brief. Bioinform.*, Apr. 2021, doi: 10.1093/bib/bbab072.

[27] R. P. Sheridan, "Using random forest to model the domain applicability of another random forest model," *J. Chem. Inf. Model.*, vol. 53, no. 11, pp. 2837–2850, Nov. 2013, doi: 10.1021/ci400482e.