JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Application Note

CODD-Pred: A Web Server for Efficient Target Identification and Bioactivity Prediction of Small Molecules

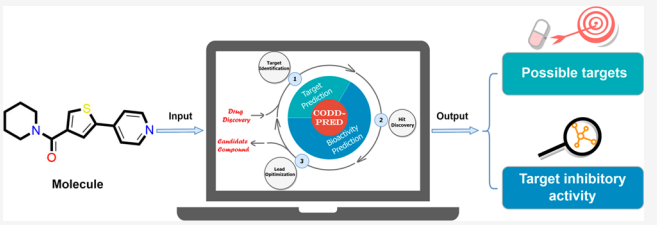
Xiaodan Yin,[▽] Xiaorui Wang,[▽] Yuquan Li, Jike Wang, Yuwei Wang, Yafeng Deng, Tingjun Hou, Huanxiang Liu, Pei Luo, and Xiaojun Yao*

Cite This: *J. Chem. Inf. Model.* 2023, 63, 6169−6176

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Target identification and bioactivity prediction are critical steps in the drug discovery process. Here we introduce CODD-Pred (COmprehensive Drug Design Predictor), an online web server with well-curated data sets from the GOSTAR database, which is designed with a dual purpose of predicting potential protein drug targets and computing bioactivity values of small molecules. We first designed a double molecular graph perception (DMGP) framework for target prediction based on a large library of 646 498 small molecules interacting with 640 human targets. The framework achieved a top-5 accuracy of over 80% for hitting at least one target on both external validation sets. Additionally, its performance on the external validation set comprising 200 molecules surpassed that of four existing target prediction servers. Second, we collected 56 targets closely related to the occurrence and development of cancer, metabolic diseases, and inflammatory immune diseases and developed a multi-model self-validation activity prediction (MSAP) framework that enables accurate bioactivity quantification predictions for small-molecule ligands of these 56 targets. CODD-Pred is a handy tool for rapid evaluation and optimization of small molecules with specific target activity. CODD-Pred is freely accessible at http://codd.iddd.group/.



INTRODUCTION

Drug research and development (R&D) is an expensive and time-consuming process, with statistics showing that it takes about 15 years and more than $2 billion to successfully develop a drug.[1,2] Drug discovery, as the first step in drug R&D, typically assumes that activating or inhibiting a target will have a therapeutic effect on the disease in its traditional process, and then high throughput screening (HTS) experiments are carried out to screen out hit compounds with expected target activity in the synthesized compound library.[3,4] These two phases usually involve target identification and hit discovery. However, experiment-based target identification and hit discovery methods are often limited by long experimental cycles and the availability of protein targets and synthetic compounds.[5,6] Therefore, computational alternative methods are used to guide and accelerate target identification and hit discovery, such as target prediction[7] and bioactivity prediction.[8−10]

Target prediction can help elucidate the mechanism of action of a bioactive compound[11] and also detect the polypharmacology of a drug[12] and promote drug repositioning.[13] Target prediction methods include ligand-based[14−22] and structure-based methods,[23] with the former having higher prediction accuracy, lower computational costs, and greater flexibility.[11,24] Notably, the application area of ligand-based target prediction methods is limited by existing chemical and biological data,[25] and the data sets of the current target prediction tools are mainly from ChEMBL and other public open-source databases. It is worth noting that some commercialized databases have also collected a large amount of valuable compound data, such as GOSTAR,[26] which is the largest manually annotated structure−activity relationship (SAR) database of small molecules, containing over 8 million compounds with over 28 million SAR points. Therefore, we curated a target prediction data set consisting of 646 498 small molecules interacting with 640 human targets from the GOSTAR database, and we adopted a double molecular graph perception (DMGP) framework combining the multi-task binary classification algorithm TrimNet[27] developed in our group recently and a multi-classification algorithm based on directed message passing neural network (DMPNN)[28] to develop a precise and efficient ligand-based target prediction method.

With the ever-growing chemical bioactivity data and the continuous improvement of computer processing power,

Received: May 12, 2023
Published: October 11, 2023

https://doi.org/10.1021/acs.jcim.3c00685
J. Chem. Inf. Model. 2023, 63, 6169−6176

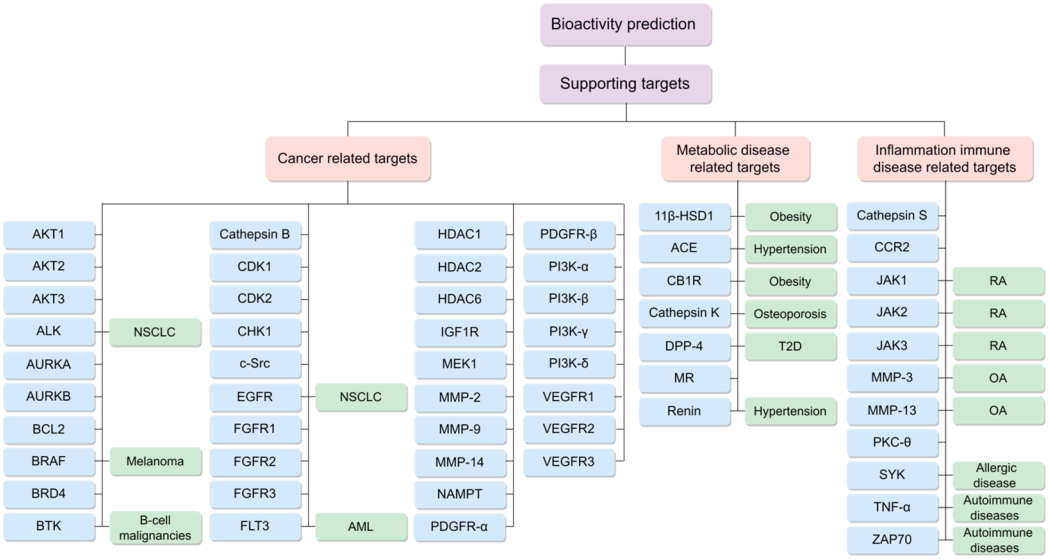


Figure 1. Fifty-six disease-related targets. The blue block represents target name and the corresponding green block is the typical disease affected by the target; only blue block indicates that the target is related to multiple diseases in this disease type.

ligand-based bioactivity prediction methods have been vigorously developed, which accurately identify new hits and promising lead compounds from enormous chemical libraries through QSAR modeling.[8,9,29] Despite the significant progress in medical technology, the prevalence of diseases such as cancer, metabolic diseases, and inflammatory immune diseases remains a pressing concern for human health worldwide. To facilitate the discovery of hit compounds with potential therapeutic activity for these diseases, we collected 56 targets related to them through literature retrieval (Figure 1). Based on the carefully curated data sets from the GOSTAR database, we developed a robust multi-model self-validating activity prediction (MSAP) framework combining various graph neural network (GNN) algorithms and traditional machine learning (ML) algorithms to model structure−activity data for small molecular inhibitors of the above 56 disease-related targets. This framework can provide reliable quantitative bioactivity prediction of potential active small molecules for cancer, metabolic diseases, and inflammatory immune diseases.

In order to comprehensively and conveniently evaluate the potential of small molecules to become drug candidates, we have integrated target prediction and bioactivity prediction of small molecules into an online sever, CODD-Pred (COmprehensive Drug Design Predictor). The user-friendly graphical interface of CODD-Pred makes it easy for specialists or nonspecialists to select the properties of compounds they are interested in for prediction. We believe that CODD-Pred will hopefully accelerate the drug discovery by facilitating the rapid evaluation and optimization of compounds with specific target activity.

MATERIALS AND METHODS

Data Set Curation. The data for constructing small molecule target prediction and bioactivity prediction models were extracted from the GOSTAR database, and details of data set curation can be found in the Supporting Information.

Model Construction. *DMGP Framework.* We constructed a double molecular graph perception framework using TrimNet[27] and DMPNN[28] in the target prediction module, which integrates the predictive results of the two algorithms to rank the probable targets of the query molecule. First, we designed a multi-task binary classification model using TrimNet to learn the effect of one molecule on multiple targets (positive or negative molecule). TrimNet is a graph-based approach with few parameters and high prediction accuracy recently proposed by our group which adopts a novel triplet message mechanism to effectively learn molecular representations. We randomly divided the target prediction data set-*bin* into training, validation, and test sets according to the molecular scaffold in a ratio of 8:1:1, respectively. The molecules in the data sets were preprocessed by RDKit[30] into molecular graphs with atomic and bond characteristics and adjacency matrixes, and TrimNet extracted molecular graphs into feature vectors that can represent molecule structures through message passing and readout stages. Different from most message passing neural networks (MPNNs),[31,32] TrimNet explicitly dropped the matrix mapping of edge features in the message stage, calculated messages from atom−bond−atom information through a triplet message mechanism, and updated the hidden state of neural networks, thus avoiding the problems of large number of parameters and insufficient extraction of edge information in MPNN methods. When a molecule is input, the output form of TrimNet is a 640-dimensional 0∼1 probability vector $\vec{a}$ corresponding to 640 targets, and each dimension vector represents the probability of the query molecule to become a positive molecule for the corresponding target. DMPNN model, as another branch of the DMGP framework, was used to estimate the high dimensional similarity of the query molecule to 640 target positive molecules. We randomly divided the data set-*multi* into training (0.8), validation (0.1), and test (0.1) sets according to the molecular scaffold, and the well-trained DMPNN can be used to measure the query molecule more similar to the positive molecule of which target. The similarity measure here is different from the traditional 2D or 3D molecular similarity, and it gives 640 similarity probability values. When a molecule is input, the output form of DMPNN is also a 640-dimensional 0∼1 probability vector $\vec{b}$, and the sum of elements in the vector $\vec{b}$ is 1. Finally, by elementwise

https://doi.org/10.1021/acs.jcim.3c00685
J. Chem. Inf. Model. 2023, 63, 6169−6176

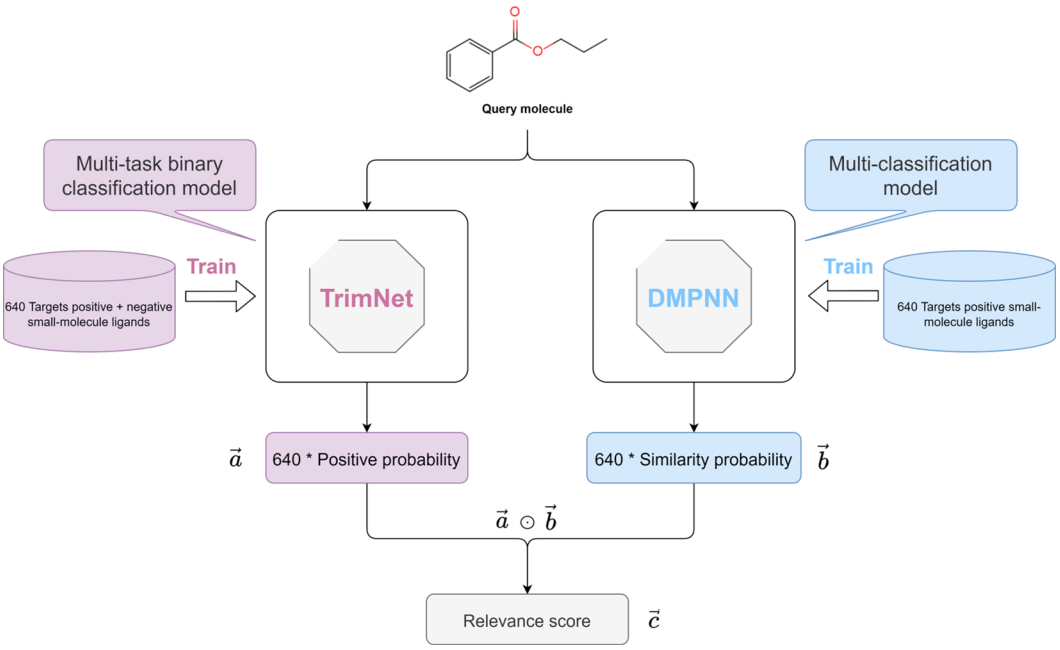


Figure 2. Workflow of DMGP framework for target prediction. DMGP framework is composed of a multi-task binary classification model, TrimNet, and a multi-classification model, DMPNN, and combines the predictive results of the two branch models to rank the probable targets of the query molecule.

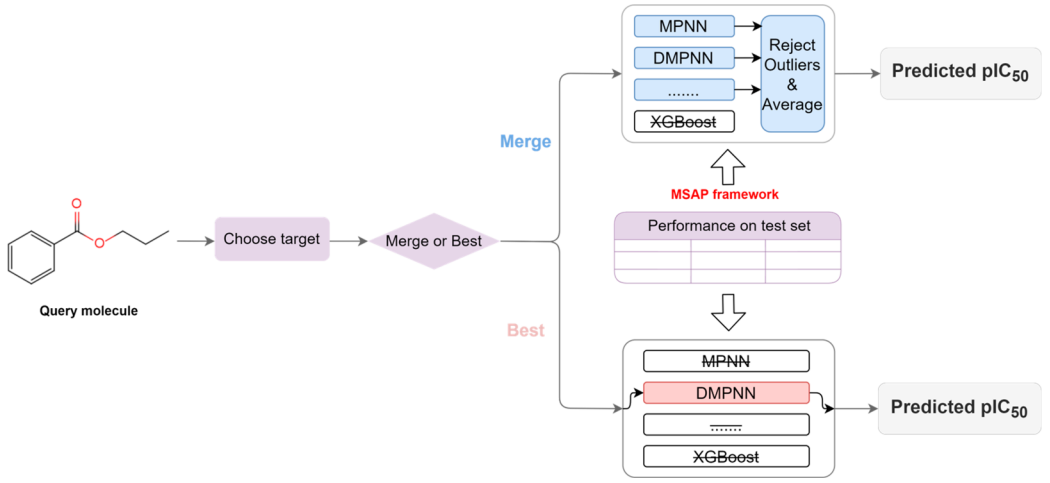


Figure 3. Workflow of bioactivity prediction. Bioactivity prediction starts by selecting a target of interest, then choosing a prediction mode, and finally predicting the pIC_{50} value of the query molecule through the MSAP framework.

multiplication of vector $\vec{a}$ and $\vec{b}$, we obtained a 640-dimensional 0∼1 relevance score vector $\vec{c}$ (formula 1), and the 640 elements in the vector $\vec{c}$ represent the final relevance scores of the query molecule to the 640 targets, respectively. When the relevance score corresponding to a target is greater, the target is more probably to be the target of the query molecule, and the workflow of DMGP framework is shown in Figure 2.

$$\vec{c} = \vec{a} \odot \vec{b} \tag{1}$$

MSAP Framework. We developed a multi-model self-validation activity prediction framework consisting of seven ML regression models for bioactivity prediction, including four graph-based deep learning models, MPNN,[31] DMPNN,[28] graph attention network (GAT),[33] and graph isomorphism network (GIN),[34] and three traditional ML models based on molecular fingerprinting, namely support vector machine (SVM),[35] random forest (RF),[36] and eXtreme Gradient Boosting (XGBoost).[37] We randomly divided the structure−activity data set of small-molecule inhibitors of each target into training, validation, and testing sets in the ratio of 8:1:1 by stratified sampling of the activity data and trained, validated, and tested the MSAP framework. For each query molecule, after selecting a target of interest, we provide it with two prediction modes, namely Best-mode and Merge-mode. Based on the performance of MSAP framework on the test set, the Best-mode selects the best performing model to predict the pIC_{50} value of the query molecule, while the Merge-mode selects several models in the framework whose performances meet the established criteria to predict the pIC_{50} value at the same time. Specifically, the models with the top-5 perform-

https://doi.org/10.1021/acs.jcim.3c00685
J. Chem. Inf. Model. 2023, 63, 6169−6176

ances (mean average error, MAE) and $R^2 > 0.6$ are selected for bioactivity prediction; when the number of models with $R^2 > 0.6$ is greater than three but less than five, the model with $R^2 > 0.6$ will be used for the prediction; and when there are less than three models with $R^2 > 0.6$, the top three models with R^2 ranking are selected for prediction. After excluding abnormal predicted values, take the average of the predicted values of multiple models as the final predicted pIC_{50} value of the query molecule in Merge-mode. The calculation process of the predicted value of the Merge-mode is shown in eqs S1−S3, and the workflow of bioactivity prediction is shown in Figure 3. The performance test metrics used in the above model frameworks are shown in the Supporting Information.

Web Server Implementation. The CODD-Pred web server is publicly accessible via a web browser. It was built using Flask framework on Ubuntu Linux, deployed on Tencent cloud. Nginx enables web access, while uwsgi supports interactions with the proxy server. CODD-Pred uses PyTorch, scikit-learn, and RDKit for model implementation and molecular data processing. It is compatible with popular browsers like Edge, Chrome, and Safari.

RESULTS AND DISCUSSION

Profile of Data Set. We curated two data sets in the target prediction section (data sets *bin* and *multi*); their details are shown in Table 1. It can be observed that the two data sets contain 640 human protein targets and 508 855 positive interactions. Compared with some target prediction tools with similar functions, such as SwissTargetPrediction (human targets, 2092; number of interactions, 494 196), our data sets contain fewer targets, but the number of interactions far exceeds that with SwissTargetPrediction, which provides a data basis for obtaining an accurate target predictor. Moreover, we classified the targets of positive molecules according to their biochemical types, which can be divided into 12 different types, including enzyme, membrane receptor, ion channel, epigenetic regulator, transcription factor, transporter, other cytosolic protein, unclassified protein, other nuclear protein, secreted protein, auxiliary transport protein, and adhesion. Then we visualized the distribution of positive molecule structures according to the biochemical types of their targets using tmap,[38] and the results are shown in Figure S1. The same color in the figure represents the same target biochemical type, each dot represents one molecule, and molecules on the same branch are structurally similar. It can be seen that the target biochemical types of most molecules are enzymes (61.08%), and there is also a significant proportion of molecules whose targets are membrane receptors (21.48%). The blue and orange dots are spread throughout Figure S1, reflecting that numerous studies have been conducted on enzymes and membrane receptors, and a large number of active molecules with rich structural types of enzymes or membrane receptors have emerged. In addition, it can be noted that the molecules whose targets belong to transcription factor or transporter are mainly distributed on several different branches; it can be inferred that the active molecules of these two types of targets may have several major structural types.

Table 1. Volume of Bioactivity Data of Two Target Prediction Data Sets

data set	targets	positive interactions	interactions
data set-*bin*	640	508855	874371
data set-*multi*	640	508855	508855

The 56 disease-related targets we collected in the bioactivity prediction section are shown in Figure 1. First, there are 38 cancer-related targets, such as Bruton's tyrosine kinase (BTK), a key therapeutic target for B-cell malignancies,[39] and anaplastic lymphoma kinase (ALK)[40] and epidermal growth factor receptor (EGFR),[41] two important targets for clinical treatment of non-small cell lung cancer (NSCLC). Next, there are seven targets related to metabolic diseases, such as 11β-hydroxysteroid dehydrogenase type 1 (11β-HSD1)[42] and cannabinoid receptor 1 (CB1R),[43] which are considered as attractive targets for the treatment of obesity and related metabolic diseases. Finally, there are 11 targets related to inflammatory immune diseases, such as Janus kinases (JAK1−JAK3),[44] the popular targets for rheumatoid arthritis (RA) research, and two important target genes in the development of osteoarthritis (OA): MMP-3 and MMP-13.[45,46] Detailed descriptions of these 56 targets can be found in the Supporting Information, and details of the structure−activity data sets of small-molecule inhibitors of 56 disease-related targets are in Table S1, in which there are 24 targets with inhibitors over 10 000 and 16 targets with inhibitors between 5000 and 10 000; this indicates that the vast majority of targets are supported by sufficient structure−activity data about their inhibitors.

Model Performance of Target Prediction. We analyze the reliability of the DMGP framework by evaluating the performance of two branching models. First, TrimNet demonstrated promising performance on the data set-*bin* test set, with AUROC, AUPRC, precision, accuracy, sensitivity, and specificity values of 0.884, 0.883, 0.823, 0.847, 0.838, and 0.821, respectively. Additionally, as another branch of the DMGP framework, the DMPNN model achieved an accuracy of 0.6664 on the data set-*multi* test set. To validate the predictive capability of the entire target prediction pipeline, we collected two external validation data sets, both containing 1500 experimentally active molecules, but with different activity thresholds (<1 and <10 nM), for testing its performance. These molecules were randomly selected from the ChEMBL 30 database and were not included in the training, validation, and test sets of TrimNet and DMPNN. Additionally, each molecule must meet the following criteria:[18] first, it is annotated as a direct binder; second, it contains <80 heavy atoms; third, it only binds to a single protein or protein complex; fourth, its assay labels with a confidence score of >3; fifth, it has at least two different reported human targets; sixth, in the external validation set with a threshold of 1 nM, all molecules have activity values (K_i, K_d, IC_{50}, or EC_{50}) of <1 nM, while in the external validation set with a threshold of 10 nM the activity values are <10 nM. The test criterion of the DMGP framework on the two external validation sets is the top-*N* accuracy of hitting at least one or two targets (Table 2). It can be observed that the DMGP framework achieved top-1 and top-5 accuracies exceeding 60 and 80%, respectively, on two external validation sets with activity thresholds of 1 and 10 nM when hitting at least one target. When hitting at least two targets, the DMGP framework achieved a top-5 accuracy of over 60% on both external validation sets; this suggests that the framework can provide references for polypharmacology and repositioning of tested molecules. In order to provide a more objective evaluation of the target prediction performance of the DMGP framework, we randomly selected 200 molecules from

https://doi.org/10.1021/acs.jcim.3c00685
J. Chem. Inf. Model. 2023, 63, 6169−6176

Table 2. Performance of DMGP Framework on the Two External Validation Data Sets

activity threshold	hit target	top-N accuracy (%)			
		N = 1	N = 5	N = 10	N = 15
<1 nM	1	65.0	84.6	89.5	91.3
	2	–	65.3	73.5	77.4
<10 nM	1	62.5	83.2	89.8	92.6
	2	–	63.9	74.2	78.8

the external validation set with an activity threshold of 1 nM (external validation data set 200). Subsequently, we conducted tests on the DMGP framework and compared its performance with that of four representative target prediction tools: SEA,[22] PASS,[47] SwissTargetPrediction,[18] and Super-PRED.[48] It is worth noting that, for the sake of fairness, when testing SEA and Super-PRED, we only considered their prediction results and did not include the database matching results. The data in Table 3 demonstrates that, compared to the other four methods, DMGP achieved the highest top-N accuracy, followed by SwissTargetPrediction, PASS, SEA, and Super-PRED. This indicates that DMGP has a competitive target prediction performance, which can be attributed to sufficient data collection for each target and the effectiveness of the design of the double molecular graph perception framework.

Table 3. Comparative Performance of Five Target Prediction Methods on External Validation Data Set 200

method	top-N accuracy (%)			
	N = 1	N = 5	N = 10	N = 15
DMGP	66.0	84.0	90.0	90.0
SwissTargetPrediction	57.5	78.0	80.5	81.0
PASS	25.5	43.0	51.0	55.5
SEA[a]	18.5	25.0	27.5	28.5
Super-PRED[a]	8.5	16.0	19.0	19.0

[a]In the calculation of top-N accuracy for SEA and Super-PRED, we excluded their database match results.

Model Performance of Bioactivity Prediction. In the bioactivity prediction module, we provide two prediction modes for the query molecule and the model invoked in the Merge-mode is defined as the "available model". The MAE, MSE, RMSE, and R^2 of the available models for each target on its test set are shown in the Table S2. From the table we can see that, after training and filtering the models for each target, there are a total of 259 available models for 56 targets. Among them, there are 125 models with R^2 values above 0.8, 109 models with R^2 values between 0.7 and 0.8, and only a few models with R^2 values between 0.6 and 0.7. In addition, the MAEs and MSEs of 259 available models mainly fall in the range 0.2−0.4, accounting for 58 and 52%, respectively, and the RMSEs are mainly in the range 0.4−0.8 (86%). In general, these models yield satisfactory performances, which can give relatively accurate bioactivity predictions for small molecules. MAE values of available models for each target data set are presented separately in Table S3. Analyzing the data in the table, we first find that RF achieves the best performance on the remaining 52 data sets, except for the four target data sets AKT1, HDAC6, MMP-14, and TNF-α, and the best performing models for AKT1, HDAC6, MMP-14, and TNF-α are SVM, DMPNN, XGBoost, and DMPNN, respectively. Then, there are 42 target data sets where the number of available models is five; these five models are RF, SVM, XGBoost, DMPNN, and MPNN. However, GAT and GIN in the MSAP framework are not included in the available models and they do not seem to be suitable for these data sets. Finally, the average MAEs of RF, SVM, XGBoost, DMPNN, and MPNN on the 56 target data sets are 0.346, 0.376, 0.391, 0.382, and 0.435, respectively. Obviously, RF shows the best predictive performance on these data sets, and MPNN has a relatively weak performance. This reflects that on these 56 regression tasks, compared with the highly complicated and specialized graph-based deep learning models, the lightweight traditional ML models based on molecular fingerprint can achieve better prediction accuracy.[49]

Web Usage. CODD-Pred is composed of two functional modules: Target Prediction and Bioactivity Prediction.

The Target Prediction module allows users to obtain possible targets for query molecules. As shown in Figure 4a, first, users need to submit a valid SMILES string in the target prediction interface ((i) enter a SMILES string; (ii) draw a molecule) and then select Start Prediction. By default, the prediction results rank the possible targets of the query molecule according to the relevance score, and we also provide information about the predicted targets, such as the UniProt ID and target class. It is worth noting that, under this module, we also integrated a DMGP model with similar accuracy training data from ChEMBL 30 for target prediction (see section ChEMBL-DMGP in the Supporting Information for details). Moreover, this module also features a functionality to return known experimental information for the input molecules. Currently, the data sources include GOSTAR and ChEMBL 30 target prediction data sets. If the experimentally known targets of the input molecule are found in the target prediction data sets, the Database Query Results section will return the experimentally known targets of the input molecule (Figure 4b).

In the Bioactivity Prediction module, users can select the target they are interested in from 56 targets, and obtain the pIC_{50} predicted value of the query molecule for the target. As shown in Figure 4c, the module supports single or batch small molecule activity prediction. First, users need to submit one or more query molecules in the bioactivity prediction interface ((i) enter a SMILES string; (ii) draw a molecule; (iii) enter multiple small molecules in smi or sdf file format), then select a target they are interested in, then select a prediction mode, and finally select Start Prediction. The prediction results are shown in Figure 4d.

CONCLUSIONS

Based on the GOSTAR database, we proposed a comprehensive online web site, CODD-Pred, which can provide researchers with small-molecule compound target prediction and bioactivity prediction. The functionality of the web site can complement some existing target prediction and bioactivity prediction tools that data sets extracted from public databases. Notably, one of the main challenges in creating accurate and applicable ML models is that the available experimental data is usually heterogeneous, noisy, and sparse;[50] therefore, during the collection of data sets, we carefully analyzed, evaluated, and processed the bias and noise of the data to ensure the robustness and reliability of the data sets used for modeling. Although the web site is fully functional and has proven performance, there are still some limitations, such as the targets included in our current target prediction

https://doi.org/10.1021/acs.jcim.3c00685
J. Chem. Inf. Model. 2023, 63, 6169−6176

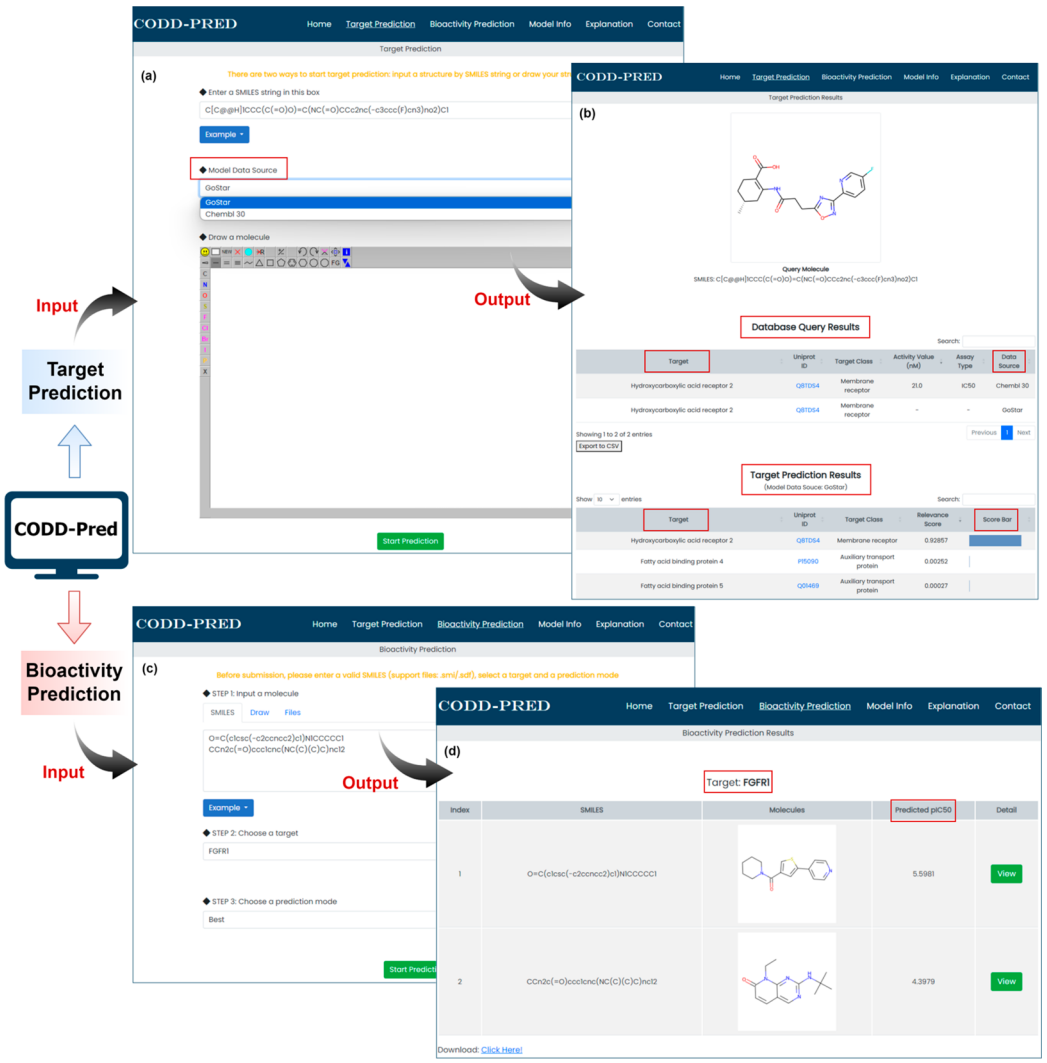


Figure 4. Graphical interface for input and output of CODD-Pred. (a) Input of target prediction. (b) Output of target prediction. (c) Input of bioactivity prediction. (d) Output of bioactivity prediction.

module are all human targets. In the future, in addition to regular updates, we will add prediction models of targets from other sources.

■ ASSOCIATED CONTENT

Data Availability Statement

Source code, external validation data sets, data set construction methods and GOSTAR data IDs are available at https://github.com/xiaodanyin/CODD-Pred. GOSTAR target prediction data sets were collected from the GOSTAR database (data updated until July 2021, https://www.gostardb.com/). ChEMBL 30 target prediction data sets are available at https://drive.google.com/file/d/1R8IIGBfo1ClfAcrAKtfAM0HgGroG7MZA/view?usp=sharing. The trained models are stored at https://drive.google.com/file/d/11BzN6rotyb4mYWnITdLtZg3bXbY1ixze/view?usp=sharing and https://drive.google.com/file/d/1BVe1XS5929g1GeDAJDqaNlAwZSoj7W2n/view, and the models built in this study are freely available via an interface at http://codd.iddd.group/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00685.

Additional information on data sets, model performance, and calculation details (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Xiaojun Yao − *Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China;* orcid.org/0000-0001-9958-8438; Email: xjyao@mpu.edu.mo

Authors

Xiaodan Yin − *Dr. Neher's Biophysics Laboratory for Innovative Drug Discovery, State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macao 999078, China; Carbon−Silicon AI Technology Co., Ltd, Zhejiang, Hangzhou 310018, China;* orcid.org/0000-0001-9282-5816

Xiaorui Wang − *Dr. Neher's Biophysics Laboratory for Innovative Drug Discovery, State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macao 999078, China;*

https://doi.org/10.1021/acs.jcim.3c00685
J. Chem. Inf. Model. 2023, 63, 6169−6176

*Carbon−Silicon AI Technology Co., Ltd, Zhejiang, Hangzhou 310018, China; orcid.org/0000-0001-6893-2013

Yuquan Li − *College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou 730000, China*

Jike Wang − *College of Pharmaceutical Sciences and Cancer Center, Zhejiang University, Hangzhou 310058, China*

Yuwei Wang − *College of Pharmacy, Shaanxi University of Chinese Medicine, Xianyang 712000, China*

Yafeng Deng − *Carbon−Silicon AI Technology Co., Ltd, Zhejiang, Hangzhou 310018, China*

Tingjun Hou − *College of Pharmaceutical Sciences and Cancer Center, Zhejiang University, Hangzhou 310058, China; orcid.org/0000-0001-7227-2580

Huanxiang Liu − *Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China*

Pei Luo − *Dr. Neher's Biophysics Laboratory for Innovative Drug Discovery, State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macao 999078, China; orcid.org/0000-0003-3095-9223

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00685

Author Contributions

[▽]These authors contributed equally to this work.

Funding

This work was funded by The Science and Technology Development Fund, Macau SAR (file no. 0056/2020/AMJ, 0114/2020/A3, 0015/2019/AMJ) and Dr. Neher's Biophysics Laboratory for Innovative Drug Discovery, State Key Laboratory of Quality Research in Chinese Medicine, Macau University of Science and Technology, Macau, China (001/2020/ALC).

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Berdigaliyev, N.; Aljofan, M. An Overview of Drug Discovery and Development. *Future Med. Chem.* **2020**, *12*, 939−947.

(2) Mullard, A. New Drugs Cost Us $2.6 Billion to Develop. *Nature Reviews. Drug Discovery* **2014**, *13*, 877.

(3) Deng, J.; Yang, Z.; Ojima, I.; Samaras, D.; Wang, F. Artificial Intelligence in Drug Discovery: Applications and Techniques. *Brief. Bioinform.* **2022**, *23*, bbab430.

(4) Kimber, T. B.; Chen, Y.; Volkamer, A. Deep Learning in Virtual Screening: Recent Applications and Developments. *Int. J. Mol. Sci.* **2021**, *22*, 4435.

(5) Rix, U.; Superti-Furga, G. Target Profiling of Small Molecules by Chemical Proteomics. *Nat. Chem. Biol.* **2009**, *5*, 616−624.

(6) Zheng, X. S.; Chan, T.-F.; Zhou, H. H. Genetic and Genomic Approaches to Identify and Study the Targets of Bioactive Small Molecules. *Chem. Biol.* **2004**, *11*, 609−618.

(7) Rognan, D. Structure-Based Approaches to Target Fishing and Ligand Profiling. *Mol. Inform.* **2010**, *29*, 176−187.

(8) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.

(9) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.

(10) Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 1013−1026.

(11) Galati, S.; Di Stefano, M.; Martinelli, E.; Poli, G.; Tuccinardi, T. Recent Advances in in Silico Target Fishing. *Molecules* **2021**, *26*, 5124.

(12) AbdulHameed, M. D. M.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring Polypharmacology Using a Rocs-Based Target Fishing Approach. *J. Chem. Inf. Model.* **2012**, *52*, 492−505.

(13) Ashburn, T. T.; Thor, K. B. Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. *Nat. Rev. Drug Discovery* **2004**, *3*, 673−683.

(14) Czodrowski, P.; Bolick, W.-G. Ocean: Optimized Cross Reactivity Estimation. *J. Chem. Inf. Model.* **2016**, *56*, 2013−2023.

(15) Liu, X.; Xu, Y.; Li, S.; Wang, Y.; Peng, J.; Luo, C.; Luo, X.; Zheng, M.; Chen, K.; Jiang, H. In Silico target Fishing: Addressing a "Big Data" Problem by Ligand-Based Similarity Rankings with Data Fusion. *J. Cheminform.* **2014**, *6*, 33.

(16) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Identifying Compound-Target Associations by Combining Bioactivity Profile Similarity Search and Public Databases Mining. *J. Chem. Inf. Model.* **2011**, *51*, 2440−2448.

(17) Gallo, K.; Goede, A.; Preissner, R.; Gohlke, B.-O. Superpred 3.0: Drug Classification and Target Prediction—a Machine Learning Approach. *Nucleic Acids Res.* **2022**, *50*, W726.

(18) Daina, A.; Michielin, O.; Zoete, V. Swisstargetprediction: Updated Data and New Features for Efficient Prediction of Protein Targets of Small Molecules. *Nucleic Acids Res.* **2019**, *47*, W357−W364.

(19) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the Macromolecular Targets of De Novo-Designed Chemical Entities through Self-Organizing Map Consensus. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 4067−4072.

(20) Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X.-Q. Targethunter: An in Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *AAPS journal* **2013**, *15*, 395−406.

(21) Liu, X.; Vogt, I.; Haque, T.; Campillos, M. Hitpick: A Web Server for Hit Identification and Target Prediction of Chemical Screenings. *Bioinformatics* **2013**, *29*, 1910−1912.

(22) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.

(23) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.

(24) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. Tools for in Silico Target Fishing. *Methods* **2015**, *71*, 98−103.

(25) Shaikh, F.; Tai, H. K.; Desai, N.; Siu, S. W. LigTMap: Ligand and Structure-Based Target Identification and Activity Prediction for Small Molecular Compounds. *J. Cheminform.* **2021**, *13*, 44.

(26) *GOSTAR*. https://www.gostardb.com/.

(27) Li, P.; Li, Y.; Hsieh, C.-Y.; Zhang, S.; Liu, X.; Liu, H.; Song, S.; Yao, X. TrimNet: Learning Molecular Representation from Triplet Messages for Biomedicine. *Brief. Bioinform.* **2021**, *22*, bbaa266.

(28) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(29) Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 1013−1026.

(30) *RDKit*; 2021. https://www.rdkit.org/.

(31) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *International Conference on Machine Learning, 2017*; PMLR: 2017; pp 1263−1272.

(32) Withnall, M.; Lindelöf, E.; Engkvist, O.; Chen, H. Building Attention and Edge Message Passing Neural Networks for Bioactivity and Physical-Chemical Property Prediction. *J. Cheminform.* **2020**, *12*, 1.
(33) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph Attention Networks. *arXiv (Statistics.Machine Learning)*, October 30, 2017, 1710.10903, ver. 1. https://arxiv.org/abs/1710.10903v1.
(34) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? *arXiv (Computer Science.Machine Learning)*, October 1, 2018, 1810.00826, ver. 1. https://arxiv.org/abs/1810.00826v1.
(35) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273−297.
(36) Ho, T. K. Random Decision Forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995*; IEEE: 1995; Vol. *1*; pp 278−282.
(37) Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. *Xgboost: Extreme Gradient Boosting. R*, package ver. 0.4-2; 2015.
(38) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminform.* **2020**, *12*, 12.
(39) Liu, J.; Chen, C.; Wang, D.; Zhang, J.; Zhang, T. Emerging Small-Molecule Inhibitors of the Bruton's Tyrosine Kinase (Btk): Current Development. *Eur. J. Med. Chem.* **2021**, *217*, 113329.
(40) Chuang, C.-H.; Chen, H.-L.; Chang, H.-M.; Tsai, Y.-C.; Wu, K.-L.; Chen, I.-H.; Chen, K.-C.; Lee, J.-Y.; Chang, Y.-C.; Chen, C.-L.; et al. Systematic Review and Network Meta-Analysis of Anaplastic Lymphoma Kinase (Alk) Inhibitors for Treatment-Naïve Alk-Positive Lung Cancer. *Cancers (Basel)* **2021**, *13*, 1966.
(41) He, J.; Zhou, Z.; Sun, X.; Yang, Z.; Zheng, P.; Xu, S.; Zhu, W. The New Opportunities in Medicinal Chemistry of Fourth-Generation Egfr Inhibitors to Overcome C797s Mutation. *Eur. J. Med. Chem.* **2021**, *210*, 112995.
(42) Xu, Z.; Liu, D.; Liu, D.; Ren, X.; Liu, H.; Qi, G.; Zhou, Y.; Wu, C.; Zhu, K.; Zou, Z.; et al. Equisetin Is an Anti-Obesity Candidate through Targeting 11β-Hsd1. *Acta Pharmaceutica Sinica B* **2022**, *12*, 2358−2373.
(43) Behl, T.; Chadha, S.; Sachdeva, M.; Sehgal, A.; Kumar, A.; Venkatachalam, T.; Hafeez, A.; Aleya, L.; Arora, S.; Batiha, G. E.-S.; et al. Understanding the Possible Role of Endocannabinoid System in Obesity. *Prostaglandins Other Lipid Mediat.* **2021**, *152*, 106520.
(44) Sk, M. F.; Jonniya, N. A.; Roy, R.; Kar, P. Unraveling the Molecular Mechanism of Recognition of Selected Next-Generation Antirheumatoid Arthritis Inhibitors by Janus Kinase 1. *ACS omega* **2022**, *7*, 6195−6209.
(45) Plsikova Matejova, J.; Spakova, T.; Harvanova, D.; Lacko, M.; Filip, V.; Sepitka, R.; Mitro, I.; Rosocha, J. A Preliminary Study of Combined Detection of Comp, Timp-1, and Mmp-3 in Synovial Fluid: Potential Indicators of Osteoarthritis Progression. *Cartilage* **2021**, *13*, 1421S−1430S.
(46) Wang, M.; Sampson, E. R.; Jin, H.; Li, J.; Ke, Q. H.; Im, H.-J.; Chen, D. Mmp13 Is a Critical Target Gene During the Progression of Osteoarthritis. *Arthrit. Res. Ther.* **2013**, *15*, R5.
(47) Pogodin, P.; Lagunin, A.; Filimonov, D.; Poroikov, V. Pass Targets: Ligand-Based Multi-Target Computational System Based on a Public Data and Naïve Bayes Approach. *SAR QSAR Environ. Res.* **2015**, *26*, 783−793.
(48) Gallo, K.; Goede, A.; Preissner, R.; Gohlke, B.-O. Superpred 3.0: Drug Classification and Target Prediction—a Machine Learning Approach. *Nucleic Acids Res.* **2022**, *50*, W726−W731.
(49) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J. Cheminform.* **2021**, *13*, 12.
(50) Huang, D. Z.; Baber, J. C.; Bahmanyar, S. S. The Challenges of Generalizability in Artificial Intelligence for Adme/Tox Endpoint and Activity Prediction. *Expert Opinion on Drug Discovery* **2021**, *16*, 1045−1056.

https://doi.org/10.1021/acs.jcim.3c00685
J. Chem. Inf. Model. 2023, 63, 6169−6176