Computers in Biology and Medicine 164 (2023) 107283

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

D3CARP: a comprehensive platform with multiple-conformation based docking, ligand similarity search and deep learning approaches for target prediction and virtual screening

Yulong Shi [a,b,1], Xinben Zhang [a,1], Yanqing Yang [a,b,1], Tingting Cai [a], Cheng Peng [a,b], Leyun Wu [a,b], Liping Zhou [a,b], Jiaxin Han [a], Minfei Ma [a,b], Weiliang Zhu [a,b,**], Zhijian Xu [a,b,*]

[a] State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

[b] School of Pharmacy, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Target prediction
Virtual screening
Molecular docking
Ligand similarity search
Deep learning

ABSTRACT

Resource- and time-consuming biological experiments are unavoidable in traditional drug discovery, which have directly driven the evolution of various computational algorithms and tools for drug-target interaction (DTI) prediction. For improving the prediction reliability, a comprehensive platform is highly expected as some previously reported webservers are small in scale, single-method, or even out of service. In this study, we integrated the multiple-conformation based docking, 2D/3D ligand similarity search and deep learning approaches to construct a comprehensive webserver, namely D3CARP, for target prediction and virtual screening. Specifically, 9352 conformations with positive control of 1970 targets were used for molecular docking, and approximately 2 million target-ligand pairs were used for 2D/3D ligand similarity search and deep learning. Besides, the positive compounds were added as references, and related diseases of therapeutic targets were annotated for further disease-based DTI study. The accuracies of the molecular docking and deep learning approaches were 0.44 and 0.89, respectively. And the average accuracy of five ligand similarity searches was 0.94. The strengths of D3CARP encompass the support for multiple computational methods, ensemble docking, utilization of positive controls as references, cross-validation of predicted outcomes, diverse disease types, and broad applicability in drug discovery. The D3CARP is freely accessible at https://www.d3pharma.com/D3CARP/index.php.

1. Introduction

Identification of drug-target interactions (DTIs) is a crucial part in innovative drug discovery and an important basis for molecular structure optimization, drug repurposing, pharmacological mechanism exploration, and early warning of potential side effects. In recent years, with the development of computational chemistry and computational biology, computer-aided drug design (CADD) has become one of the highly focused strategies for DTIs study, effectively saving the time and resources required for wet experiments [1,2]. Nowadays, structure-based, ligand-based and data-driven drug design play an important role in DTIs study for drug discovery and development [3–5]. Molecular docking is a structure-based tool to calculate the binding site complementarity between the ligand and the target [6]. The scoring function [7] is the core of the molecular docking, which can not only be used to determine the protein-ligand binding mode [8,9], but also to predict the binding affinity of the protein-ligand interaction [10,11]. It is widely applied to screen innovative molecular scaffolds and guide structure optimization [12,13]. Meanwhile, multiple-conformation based docking algorithm refers to the use of multiple protein structures to represent a flexible receptor, which can be used to account for protein structural variations [14]. Besides, studies showed that

* Corresponding author. State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

** Corresponding author. State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

E-mail addresses: wlzhu@simm.ac.cn (W. Zhu), zjxu@simm.ac.cn (Z. Xu).

[1] These authors made equal contributions to this work.

https://doi.org/10.1016/j.compbiomed.2023.107283
Received 8 June 2023; Received in revised form 15 July 2023; Accepted 28 July 2023
Available online 29 July 2023

molecular dynamics (MD) simulations combined with free-energy calculations via MM/GBSA or free energy perturbation (FEP) can improve the binding affinity prediction of poses obtained from molecular docking [15–17]. Ligand similarity search is a ligand-based computational strategy which is based on the hypothesis that structurally similar compounds exhibit similar physicochemical properties and biological functions [18]. It shows significant advantages of scaffold hopping which allows a chemical substructure to be substituted for better selectivity, bioactivity and pharmacokinetic properties [19]. In recent years, an exponential increase in the amount of available bioactivity data has been witnessed [20]. Many public or commercial accessible chemical databases, such as ChEMBL [21], PubChem [22] and BindingDB [23], contain millions of bioactivity data primarily extracted from the scientific literature, and serve as the essential resources for cheminformatics, chemical biology and medicinal chemistry. Accordingly, data-driven deep learning has achieved remarkable success due to its ability to perform feature detection from massive amounts of bioactive data and their flexibility as neural network architectures [24]. It also has shown superior performance to other conventional machine learning methods in fields like molecular property prediction and DTIs prediction [25–27].

Over the past two decades, a variety of web tools have been developed for target prediction and virtual screening. For example, Jiang's group developed TarFisDock [28] for automating the procedure of searching for small molecule-protein interactions over 698 protein structures with molecular docking. Grosdidier A et al. constructed SwissDock [29] based on EADock DSS engine that allows users to select proteins from the RCSB Protein Data Bank (PDB) [30] or to upload structure files. D3Targets-2019-nCoV [31] was developed to efficiently discover effective drugs against the SARS-CoV-2 via docking. SwissTargetPrediction [32] and SwissSimilarity [33] were used for accurately predicting the targets of bioactive molecules and virtual screening of ligand libraries based on ligand similarity measures, respectively. In addition, advances in deep learning algorithms have once again boosted the evolution of drug discovery. Öztürk H et al. constructed DeepDTA [34] with convolutional neural networks (CNNs) to predict DTIs by using protein sequences and compound SMILES information. Lim J et al. [35] proposed a graph neural network (GNN) approach for predicting DTIs based on the spatial information from the protein-ligand binding pose, which shows better performance than docking and other deep learning methods for both pose prediction and virtual screening. Stokes JM et al. [36] reported a deep neural network (DNN) capable of predicting molecules with antibacterial activity by using a directed-message passing neural network (D-MPNN) architecture. They successfully identified halicin as a new broad-spectrum antibacterial compound, demonstrating the effectiveness of deep learning methods in drug repositioning. Nonetheless, it should be noted that each of the above approaches has its own drawbacks, such as dependencies on target or ligand structures, model overfitting, or model interpretability. Therefore, a platform integrating these methods together to take advantage of all the available technologies is meaningful and useful.

In this work, toward comprehensive perspectives on the DTIs from different computational tools, we integrated multiple-conformation based docking, ligand similarity search and deep learning approaches to construct a comprehensive webserver, namely D3CARP, for predicting potential targets against a given compound, or screening hit compounds against a given target. The brief workflow of the D3CARP is shown in Fig. 1. It can be freely accessible at https://www.d3pharma.com/D3CARP/index.php.

2. Materials and methods

2.1. Construction of the prediction platform based on molecular docking

Target database. Molecular docking is one of the most widely used tools in drug discovery. In contrast to traditional molecular docking, reverse docking is used for identifying targets for a given ligand among various targets. The three-dimensional (3D) target structures are necessary for traditional molecular docking and reverse docking. In this part, the crystal complex structures were obtained from the PDBbind-CN database (version 2020) [37]. Each protein-ligand pair was annotated with experimentally measured binding affinity, and the native ligand was regarded as the positive control. The root mean squared deviation (RMSD) < 2 Å between the docking pose and the native binding pose is considered as a main criterion of molecular docking successfully reproducing the experimental conformation [38,39]. And the scores of active compounds are generally lower than -5 kcal/mol [40,41]. Thus, re-docking was further performed to keep complex systems with the docking score < -5 kcal/mol and the RMSD <2 Å. Due to the induced-fit effect and protein conformational change, the binding pockets of different conformations of the same protein may have significant differences, so these conformations will be preserved to represent the flexibility of the protein structure. In addition, we selected the target conformations that combined with highly active molecules as a representative conformation set for preliminary study. Overall, the D3CARP target database consists of 1970 targets with 9352 conformations, among which 716 conformations were extracted for the representative conformation set.

Docking parameters. The docking process was performed by AutoDock Vina [42], which is one of the fastest and most widely used open-source programs for molecular docking. Both the protein and ligand formats were previously converted to pdbqt by using structural preprocessing scripts in ADFRsuite version 1.0 (https://ccsb.scripps.edu/adfr/downloads/). All of the docking boxes of each protein were generated by extending 5 Å in each dimension based on the coordinates of the native ligand in the crystal complex. The first-ranked docking score and docking pose will be reported with the random seed number set as 1.

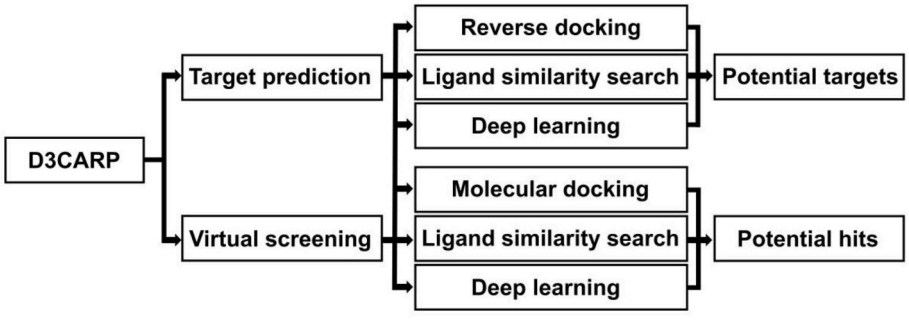


Fig. 1. The brief workflow of the D3CARP.

2.2. Construction of the prediction platform based on ligand similarity search

Ligand database. Ligand similarity search is one of the techniques used for target prediction and virtual screening. This method compares structures of the query compounds to the compounds with known targets to find potential targets or hit compounds. In this part, ligand structures and target-ligand pairs were obtained from BindingDB [43] (updated 2021-11-01), which is a public database containing over two million measured binding affinities for approximately one million small molecules, focusing chiefly on the interactions between proteins and drug-like molecules. For each compound included in the ligand database, detailed information was annotated as the vital basis for ligand-based target prediction and virtual screening, such as the target name, organism, biological activity, and literature sources.

Ligand similarity search. The major challenge of ligand similarity search is to accurately quantify similarity between the query molecule and the reference ligands. The actual quantification of two-dimensional (2D) similarity is usually calculated by the Tanimoto coefficient, which is the number of common positive bits in both strings divided by the total number of positive bits between both strings. Here, three most commonly used molecular fingerprints, namely the path-based fingerprint FP2, the substructure-based fingerprints FP4 and MACCS, are employed for computing 2D similarity of query ligands to the D3CARP ligand database via Open Babel (version 3.1.0) software [44]. In addition, the LS-align program [45] was employed to perform 3D-structure based ligand alignment, which can further capture the physical and functional features required for the biological interaction. Considering the structural and chemical differences and the structural flexibility of the conformational changes, both two modules of Rigid-LS-align and Flexi-LS-align in LS-align were added to the platform for rigid-body and flexible alignments, respectively. The lowest energy conformation of each molecule in the ligand database was generated under MMFF94 force filed by using the RDKit toolkit (version 2020.09.5). Users can customize similarity thresholds for 2D and 3D ligand similarity searches.

2.3. Construction of the prediction platform based on deep learning

Dataset. Deep learning offers an increased expressive power in identifying, processing, and extrapolating new DTIs based on existing DTIs data [34,46]. In this part, DTIs prediction is defined as a classification problem and a regression problem, respectively. In the classification model, ligand-target pairs with biological activities K_i, K_d, IC_{50} or EC_{50} lower than 10 μM were extracted from the BindingDB database, with a total of ~1.2 million positive entries, 787,543 ligand SMILES strings and 5901 protein sequences. The same number of negative entries composed of ligands and random targets were added to the dataset. Numeric labels 0 and 1 represent negative and positive data pairs, respectively. Since the inhibition constant K_i obtained by different research groups and different experiments is more stable than other biological activity indicators, only the K_i is used as the label of drug-target interaction in the regression model. In addition, we constructed a special normalization formula to calculate the molecular activity intensity compared to other active compounds on the same target. Specifically, ligand-protein pairs with K_i lower than 1 M in BindingDB were extracted and converted into the normalized values according to Equation (1). If there are multiple activity data for the same compound with the same target, only the highest bioactivity is taken. The normalized value nK_i reflects the difference in activity intensity between the query compound and the highest active compound for the same target, and it can be converted to K_i by Equation (2). Equation (1) and Equation (2) are defined as follows:

$$nK_i = \frac{lg\ K_i}{lg\ K_{i_{max}}} \tag{1}$$

$$K_i = 10^{nK_i \times lg\ K_{i_{max}}} \tag{2}$$

where K_i and $K_{i_{max}}$ are the inhibition constants of the query compound and the highest active compound, respectively. And nK_i is the normalized inhibition constant.

Hence, 340,817 positive entries, 189,849 ligand SMILES strings and 2568 protein sequences were extracted for the regression model. Both the labelled classification and regression datasets were split into the training, validation and test sets by a ratio of 98:1:1.

Deep learning architecture. As a kind of graph neural network, Message passing neural networks (MPNNs) [47] perform well in learning graph topological properties and are often used in molecular property prediction [48–50]. Convolutional neural networks (CNN) [51] have been proven effective in extracting useful information from protein sequences and are widely used in DTIs prediction [27,52–54]. Thus, we employ MPNNs and CNN to learn the key features of compounds and proteins by using the DeepPurpose [55], which is a comprehensive deep learning library that can rapidly generate deep learning models for DTIs prediction (Fig. 2). The platform provides two prediction methods to evaluate the DTIs, where MPNNs-CNN model is a classification model to predict their binding possibility, and MPNNs–CNN–R model is a regression model to predict their binding strength. Furthermore, 6 evaluation metrics, accuracy (Acc), precision (Pre), recall, F1 score (F1), area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUPRC) were used to evaluate the prediction of MPNNs-CNN model, while the Pearson correlation coefficient and concordance index were used to evaluate the prediction of MPNNs–CNN–R model. The formulae of these evaluation metrics are given below:

$$\text{Acc} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$\text{Pre} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1} = \frac{2 \times Pre \times Recall}{Pre + Recall} \tag{6}$$

where TP (True positive) and TN (True negative) represent the numbers of correctly predicted positive and negative samples, respectively, while FP (False positive) and FN (False negative) represent the numbers of wrong predicted positive and negative samples, respectively.

$$\text{Pearson correlation} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}} \tag{7}$$

where N represents the number of all samples, while x_i and y_i represent the labels and predicted values of samples, respectively.

$$\text{Concordance index} = \frac{\sum_{i,j} 1_{T_j<T_i} \bullet 1_{\eta_j>\eta_i} \bullet \delta_j}{\sum_{i,j} 1_{T_j<T_i} \bullet \delta_j} \tag{8}$$

where η_i represents the risk score of a unit i, $1_{T_j<T_i}$ means $1_{T_j<T_i} = 1$ if $T_j < T_i$ else 0, $1_{\eta_j>\eta_i}$ means $1_{\eta_j>\eta_i} = 1$ if $\eta_j > \eta_i$ else 0.

2.4. Chemical space visualization of D3CARP ligand database

The complexity and diversity of compounds in the ligand database influence the prediction performance of ligand similarity search and deep learning approaches. Chemical space visualization provides a practical means to characterize the distribution of tremendous amount of compounds by reducing high-dimensional data to low-dimensional space. In this section, the D3CARP ligand database was clustered using

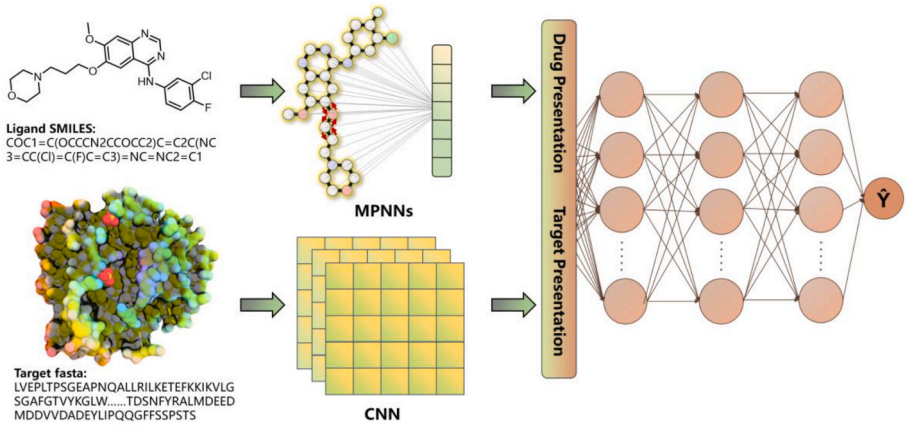


Fig. 2. Frameworks of the MPNNs-CNN model.

the *k*-means algorithm according to the molecular Morgan fingerprints [56], constituting a subset of representative 50,000 compounds. Afterwards, principal component analysis (PCA) [57] and t-distributed stochastic neighbor embedding (*t*-SNE) [58] were used to compare this representative subset to the 12,449 approved, investigational and experimental drugs in the DrugBank database (version 5.1.9) [59]. Descriptors for PCA are molecular weight (MW), topological polar surface area (TPSA), calculated octanol-water partition coefficient (LogP), number of H-bond acceptors (HBA), number of H-bond donors (HBD), and number of rotatable bonds (nRotB) calculated by RDKit, which provide an overview of the drug-like properties of the molecule. And descriptors for *t*-SNE are extended-connectivity fingerprints (ECFPs) [56] with a bit-vector length of 2048 bits and radius of 2 adjacent atoms, where each bit represents the presence or absence of a particular substructure to provide an overview of the 2D structural diversity of compounds. Besides, the principal moments of inertia (PMI) analysis [60] was further used to calculate the 3D molecular shape distribution of the compound in the lowest energy conformation. It classified all of the molecules as rods, discs or spheres around the triangle to demonstrate the molecular shape diversity of the ligand library.

2.5. Disease information related to targets

Mapping of therapeutic targets with associated diseases can contribute to the understanding of target function. Thus, the related disease types of all targets in D3CARP target database were collected from the UniProt Knowledgebase (UniProtKB) [61] and the Therapeutic Target Database (TTD) [62], both of which were extensively used due to their rich annotation on proteins.

3. Results

3.1. D3CARP overview

The D3CARP server was developed based on PHP, and hosted on a Linux server. Target prediction and virtual screening are two main functions of this platform. The former is to identify potential targets that can interact with a certain molecule from various targets, and the latter is to screen out potential active compounds that can bind to a specific target. Specifically, 2D ligand similarity methodologies can be carried out using three alternative molecular fingerprints, namely FP2, FP4, and MACCS. Since the FP2 fingerprints of the ligand database have been generated and saved in advance, the similarity calculation based on FP2 is about 5 s per compound, which is 20–30 times faster than the other two fingerprints. Therefore, it is recommended to try the FP2 first for the 2D similarity calculation. The 3D similarity module is divided into the rigid-body and flexible ligand structural alignments, and the rigid superposition is about 1 h per compound, which is 10 times faster than the other. As for the deep learning model, the platform provides MPNNs-CNN and MPNNs–CNN–R models to evaluate the DTIs of the target and ligand, both of which take time generally less than 1 min. The whole time for the target prediction and virtual screening of each module is summarized in Table S1. The D3CARP is free to all users and there is no login requirement. But it is recommended to register the account before using the D3CARP to keep data private and not viewable by other users.

3.2. Molecular physicochemical properties and chemical space of ligand database

The distributions of the compound MW, TPSA, LogP, HBA, HBD, and nRotB were analyzed to describe whole physicochemical property profile of the D3CARP ligand database (Fig. 3A–F). The majority of compounds have MW between 300 and 600 Da (80.68%), TPSA between 30 and 150 $Å^2$ (88.75%), LogP between 2 and 6 (76.16%), HBA between 2 and 10 (92.16%), HBD fewer than 4 (89.26%), nRotB between 2 and 10 (82.92%). Notably, more than half of the compounds (62.58%) have druggable potential according to Lipinski's rule of five (MW $<$ 500; log P $\leq$ 5; HBD$\leq$5; HBA$\leq$10). To further investigate whether the ligand database covers the chemical space of most of the marketed drugs and clinical drug candidates, we compared the physicochemical properties and structural characteristics of 50,000 representative compounds clustered from the ligand database with drugs from the DrugBank database (version 5.1.9). DrugBank is a freely accessible database containing information on approved drugs and drug candidates, including 2568 approved drugs, 3660 investigational drugs and 6221 experimental drugs. The space occupied by the physicochemical properties was analyzed by PCA, and the result showed that the D3CARP has a high degree of overlap with the properties of approved and in-research drugs in DrugBank according to Fig. 3G. Moreover, *t*-SNE analysis showed that the representative compound set was broader in the chemical space distribution of molecular structure, and the clusters were more compact and continuous, which could be beneficial for the target prediction of the compounds and their derivatives (Fig. 3H). As for the shape diversity, approved and in-research drugs are concentrated in rod-like and disc-like regions, whereas the D3CARP database fills the majority of the 3D shape chemical space (Fig. 3I), suggesting that it contains a greater variety of scaffold types, implying that an advantage in shape diversity would enable 3D ligand similarity methods perform well.

3.3. Target-related diseases and disease-related targets

In the D3CARP target database, there are 1447 therapeutic targets out of all targets, corresponding to 2168 clinical diseases, including cancer, stroke, hypertension, alzheimer, pain, etc. We tend to focus on a

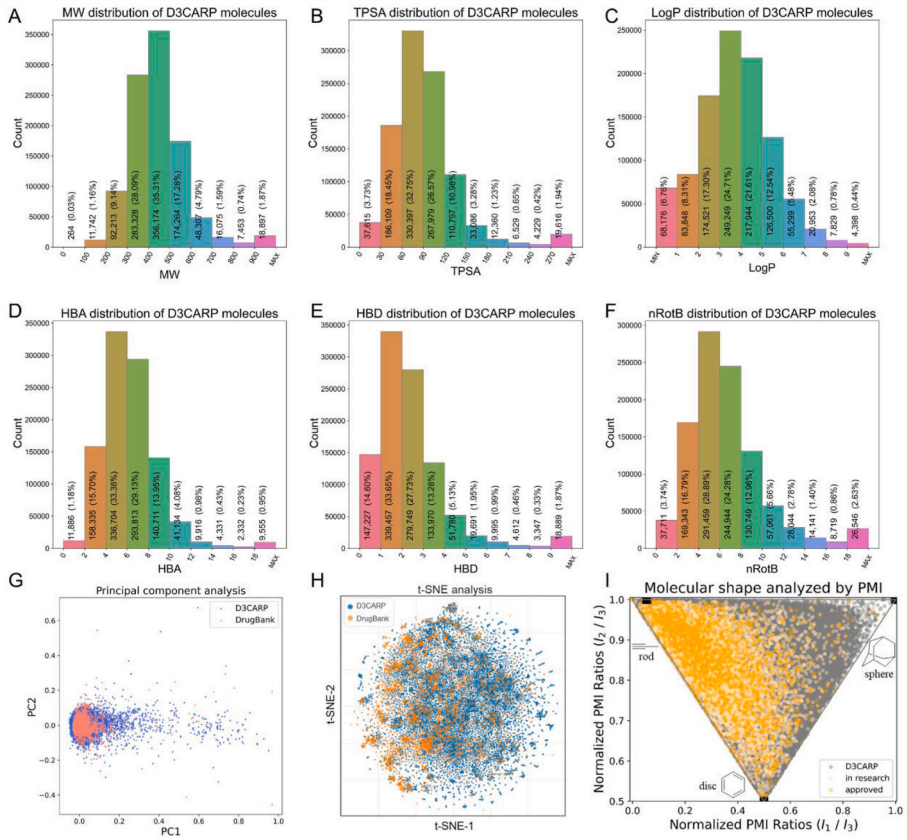


Fig. 3. Distribution of physicochemical properties and chemical space visualization. (A) MW distribution; (B) TPSA distribution; (C) LogP distribution; (D) HBA distribution; (E) HBD distribution; (F) nRotB distribution; (G) PCA of the chemical space of the D3CARP (royalblue) and DrugBank database (salmon); (H) *t*-SNE of the chemical space of the D3CARP (orange) and DrugBank database (blue). (I) PMI of the shape diversity of the D3CARP (grey) and DrugBank database (approved: orange, in research: light orange).

specific target for drug design, which may lead to ignoring possible off-target effects or underestimating the synergistic effect of the multi-target effect. For example, Alzheimer's disease is associated with 15 therapeutic targets in the database, and finding one or more drugs that can act on multiple of these therapeutic targets at the same time may greatly improve the efficacy of drug treatment (Fig. 4A). In addition to targeting human-derived therapeutic targets, certain diseases can also be treated by acting on viral or bacterial targets like human immunodeficiency virus (HIV) disease (Fig. 4B). The complex and vast network of targets and diseases discovered so far suggests that a single target is often associated with multiple diseases, and different diseases may share multiple identical therapeutic targets, which are in agreement with many reports [63–65]. For example, epidermal growth factor receptor (EGFR) is associated with 20 diseases, among which lung cancer and breast cancer are closely related to the same therapeutic targets, such as ABL1, ALK, ERBB2, CDK4, CDK6, PD1L1, PDCD1, ROS1 and VGFR2 in addition to EGFR (Fig. 4C). Therefore, in the docking module, users can not only select the specified target, but also select many targets corresponding to the same disease for subsequent calculation. The ligand similarity module and the deep learning module perform calculations with all in-house ligands or targets by default due to the fast calculation speed.

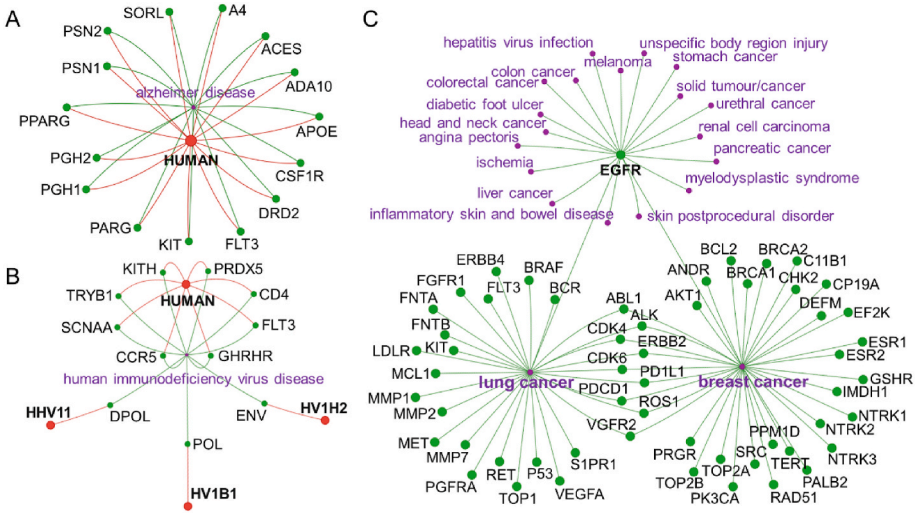


Fig. 4. Network node relationships between specific targets and diseases in the D3CARP database. (A) Association of the Alzheimer's disease with human-derived targets. (B) Association of the HIV disease with human and viral targets. (C) Relationship between diseases with EGFR and other targets.

3.4. Prediction performance evaluation

For deep learning-based methods, the ROC curve and precision recall curve of the MPNNs-CNN model are shown in Fig. 5A and B, with AUC 0.99 and AUPRC 0.99. Other performance indicators are summarized in Fig. 5C, including Acc (0.96), Pre (0.97), recall (0.95) and F1 (0.96), demonstrating the precise predictability of the MPNNs-CNN model. Besides, the Pearson correlation coefficient and concordance index of MPNNs–CNN–R model are 0.81 and 0.82, respectively.

To further evaluate the prediction performance of each D3CARP computational module, we collected 18 widely used FDA-approved drugs and their indication-related targets as an external dataset (Table S2), with targets covering enzymes, kinases, and ion channels, etc. Molecular docking and MPNNs–CNN–based deep learning modules took the top 10% of the predicted targets as predicted positive targets. For ligand similarity searches, targets with ligand similarity above 0.9 were taken as the predicted positive target, and if there was no ligand with similarity higher than 0.9, the top five targets would be selected. Since the MPNNs–CNN–R model predicted normalized activity values, which was not suitable for inter-target comparisons, it was not included in the test list. The prediction results showed that ligand similarity searches performed the best prediction, among which three 2D molecular fingerprints and 3D flexible method successfully predicted the targets of all drugs, and 3D rigid method successfully predicted 17 drugs (Table 1). Even after excluding the same ligand structure as the query compound from the prediction results, the average accuracy of the five ligand similarity approaches was still as high as 0.94. Besides, deep learning methods had an accuracy as high as 0.89. However, reverse docking relying only on the docking score as a criterion was significantly worse than the other two modules. Due to the superiority of molecular docking in protein-ligand interaction mode analysis and visualization capabilities, we retain the reverse docking function of the web pages.

3.5. Computational pipeline for the target prediction and virtual screening

D3CARP provides a concise and user-friendly web interface. The task submission of each method is similar, which has three steps: 1) Set the job name; 2) Upload the compound structure; 3) Check the specified parameters and submit job. The input and output items of each prediction method are shown in Fig. 6. Users can provide the query molecule by uploading the compound file in format of mol2, mol, sdf, or smi, and the platform will automatically perform the calculation. In the molecular docking studies, the output includes not only the docking scores and atom efficiencies of the query molecule against each target, but also the docking scores, atom efficiencies, and ligand potency of the positive control, as well as 2D and 3D similarities between the query molecule and the positive control. In the ligand similarity studies, the output contains structures of similar molecules, and related target name, organism, biological activity (K_i, IC_{50}, K_d, and EC_{50}), and literature sources. In the deep learning studies, the output is the binding possibility in activity prediction, and binding strength in activity value prediction, respectively. Among them, diseases of therapeutic targets will be output together. After the calculation is completed, the user can view and download the results on the web page.

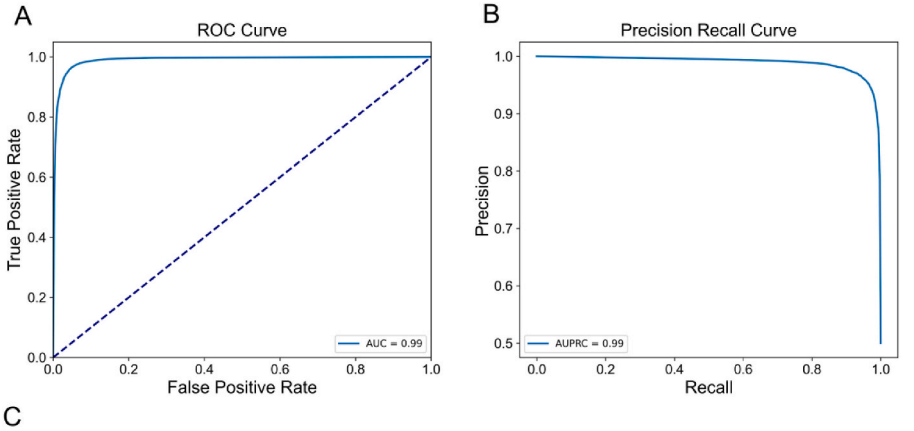
Table 1
Prediction performance of each method in D3CARP.

Type	Method	Accuracy
Reverse Docking	AutoDock Vina	0.44
Ligand Similarity (2D)	FP2	1.00 (0.94)[a]
	FP4	1.00 (0.94)
	MACCS	1.00 (0.94)
Ligand Similarity (3D)	Rigid-LS-align	0.94 (0.89)
	Flexi-LS-align	1.00 (1.00)
Deep Learning	MPNNs-CNN	0.89

[a] Values in parentheses are the accuracy after removing ligands with the same structure as the query compounds.

4. Conclusion and discussion

In this study, we developed a webserver called D3CARP, where a series of high-quality DTI prediction models were embedded into. Although the three prediction methods represent the current mainstream target prediction and virtual screening technologies, each of them still has its own drawbacks. For example, ligand similarity approaches are based on the hypothesis that similar molecules tend to have



C

Method	AUC	AUPR	Acc	Pre	Recall	F1
MPNNs-CNN	0.99	0.99	0.96	0.97	0.95	0.96
Method	Pearson Correlation			Concordance Index		
MPNNs-CNN-R	0.81			0.82		

Fig. 5. Prediction performance evaluation of the deep learning models. (A) The ROC curve of the MPNNs-CNN model. (B) The precision recall curve of the MPNNs-CNN model. (C) Performance of the MPNNs-CNN and MPNNs–CNN–R models.

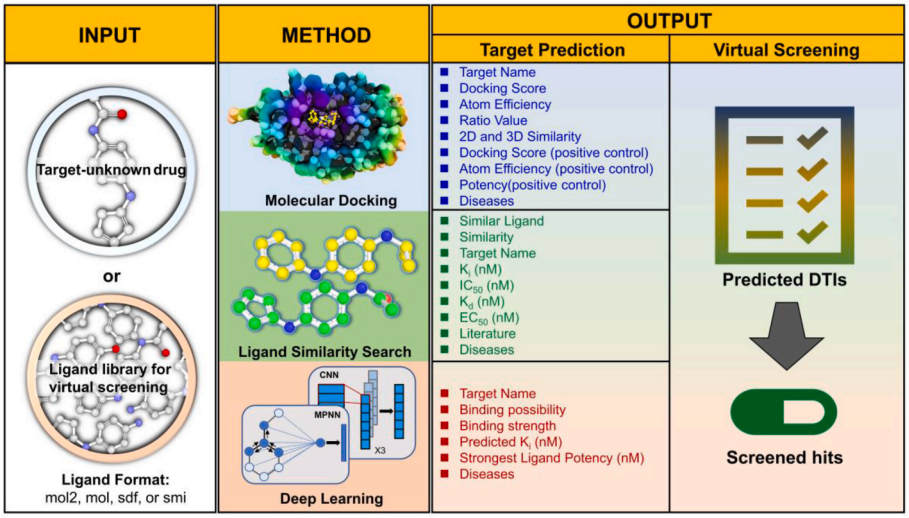


Fig. 6. The input and output items of each method in D3CARP.

similar physicochemical properties and biological activities. However, compounds exhibiting good similarity with active molecules sometimes are inactive if they are different in key positions that are crucial for protein interactions. When performing predictions, we also introduce some additional computations to help users analyze the reasonability of the results. For instance, multiple-conformation docking is supported as the ensemble docking to highlight protein conformational changes, with an average of 5 conformations per target. The analysis of docking scores and atomic efficiency scores can be used to screen atom-economical small molecules. The docking score and biological activity information of the ligand in the crystal structure can serve as a standard positive control for the query molecule. In the ligand similarity search, the structures of various similar compounds and their literature sources are presented together in the results. In the deep learning regression model, the activity range intervals of different targets are considered. The establishment of D3CARP is an effort to overcome the shortcomings of the individual computational approach, such as the low accuracy of the molecular docking, the inability of the ligand similarity search to reflect binding strength, and the poor model interpretability of the deep learning. Additionally, the D3CARP provides feasibility for comprehensive analysis of the prediction results from multiple different methods, which is beneficial for the improvement of the prediction accuracy, such as the combination of the molecular docking and deep learning [66], or the combination of the molecular docking and ligand similarity [67], etc. In the end, the D3CARP could serve as a reliable tool for the DTIs-based target prediction, virtual screening and interaction mechanism exploration.

Author contributions

Zhu W and Xu Z conceived and designed the study. Shi Y and Zhang X wrote the script code for the webserver. Shi Y, Cai T, Peng C, Han J, Wu L, Zhou L, and Ma M collected the data. Shi Y and Yang Y developed docking models, ligand similarity search models, and deep learning models. Shi Y performed the data analysis. Zhu W, Xu Z, and Shi Y wrote the paper.

Funding

This work was supported by Key project at central government level: The ability establishment of sustainable use for valuable Chinese medicine resources (2,060,302); National Key Research and Development Program of China (2022YFA1004304); and National Natural Science Foundation of China (82,273,851, 81,302,699).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.107283.

References

[1] A. Baldi, Computational approaches for drug design and discovery: an overview, Sys. Rev. Pharm. 1 (2010) 99.
[2] V. Sharma, S. Wakode, H. Kumar, Structure-and ligand-based drug design: concepts, approaches, and challenges, Chemoinformatics and bioinformatics in the pharmaceutical sciences (2021) 27–53.
[3] L. Pinzi, G. Rastelli, Molecular docking: shifting paradigms in drug discovery, Int. J. Mol. Sci. 20 (2019).
[4] K. Szilagyi, B. Flachner, I. Hajdu, M. Szaszko, K. Dobi, Z. Lorincz, S. Cseh, G. Dorman, Rapid identification of potential drug candidates from multi-million compounds' repositories. Combination of 2D similarity search with 3D ligand/structure based methods and in vitro screening, Molecules (2021) 26.
[5] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (2019) 463–477.
[6] D.B. Kitchen, H. Decornez, J.R. Furr, J. Bajorath, Docking and scoring in virtual screening for drug discovery: methods and applications, Nat. Rev. Drug Discov. 3 (2004) 935–949.
[7] X.Y. Meng, H.X. Zhang, M. Mezei, M. Cui, Molecular docking: a powerful approach for structure-based drug discovery, Curr. Comput. Aided Drug Des. 7 (2011) 146–157.
[8] J. Lyu, S. Wang, T.E. Balius, I. Singh, A. Levit, Y.S. Moroz, M.J. O'Meara, T. Che, E. Algaa, K. Tolmachova, A.A. Tolmachev, B.K. Shoichet, B.L. Roth, J.J. Irwin, Ultra-large library docking for discovering new chemotypes, Nature 566 (2019) 224–229.
[9] J.C. Hermann, R. Marti-Arbona, A.A. Fedorov, E. Fedorov, S.C. Almo, B. K. Shoichet, F.M. Raushel, Structure-based activity prediction for an enzyme of unknown function, Nature 448 (2007) 775–779.
[10] I. Parveen, P. Khan, S. Ali, M.I. Hassan, N. Ahmed, Synthesis, molecular docking and inhibition studies of novel 3-N-aryl substituted-2-heteroarylchromones targeting microtubule affinity regulating kinase 4 inhibitors, Eur. J. Med. Chem. 159 (2018) 166–177.
[11] H.H.M. Abdu-Allah, S.C. Wu, C.H. Lin, Y.Y. Tseng, Design, synthesis and molecular docking study of alpha-triazolylsialosides as non-hydrolyzable and potent CD22 ligands, Eur. J. Med. Chem. 208 (2020), 112707.
[12] L. Hu, Q. Ren, L. Deng, Z. Zhou, Z. Cai, B. Wang, Z. Li, Design, synthesis, and biological studies of novel 3-benzamidobenzoic acid derivatives as farnesoid X receptor partial agonist, Eur. J. Med. Chem. 211 (2021), 113106.
[13] Y. Wang, W. Hu, Y. Yuan, Protein arginine methyltransferase 5 (PRMT5) as an anticancer target and its inhibitor discovery, J. Med. Chem. 61 (2018) 9429–9441.
[14] R.E. Amaro, J. Baudry, J. Chodera, Ö. Demir, J.A. McCammon, Y. Miao, J.C. Smith, Ensemble docking in drug discovery, Biophys. J. 114 (2018) 2271–2278.

[15] Z. Li, X. Li, Y.Y. Huang, Y. Wu, R. Liu, L. Zhou, Y. Lin, D. Wu, L. Zhang, H. Liu, X. Xu, K. Yu, Y. Zhang, J. Cui, C.G. Zhan, X. Wang, H.B. Luo, Identify potent SARS-CoV-2 main protease inhibitors via accelerated free energy perturbation-based virtual screening of existing drugs, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 27381–27387.

[16] I. Maffucci, X. Hu, V. Fumagalli, A. Contini, An efficient implementation of the nwat-MMGBSA method to rescore docking results in medium-throughput virtual screenings, Front. Chem. 6 (2018) 43.

[17] G. Rastelli, L. Pinzi, Refinement and rescoring of virtual screening results, Front. Chem. 7 (2019) 498.

[18] M. Johnson, G.M. Maggiora, Concepts and Applications of Molecular Similarity, John Wiley & Sons, New York, 1990.

[19] Y. Hu, D. Stumpfe, J. Bajorath, Recent advances in scaffold hopping, J. Med. Chem. 60 (2017) 1238–1246.

[20] G.S. Sittampalam, S.D. Kahl, W.P. Janzen, High-throughput screening: advances in assay technologies, Curr. Opin. Chem. Biol. 1 (1997) 384–391.

[21] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) D1100–D1107.

[22] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2019 update: improved access to chemical data, Nucleic Acids Res. 47 (2019) D1102–D1109.

[23] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, M.K. Gilson, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, Nucleic Acids Res. 35 (2007) D198–D201.

[24] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[25] A. Koutsoukas, K.J. Monaghan, X. Li, J. Huan, Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data, J. Cheminf. 9 (2017) 42.

[26] E.B. Lenselink, N. Ten Dijke, B. Bongers, G. Papadatos, H.W.T. van Vlijmen, W. Kowalczyk, I.J. Ap, G.J.P. van Westen, Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set, J. Cheminf. 9 (2017) 45.

[27] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, D.R. Koes, Protein-ligand scoring with convolutional neural networks, J. Chem. Inf. Model. 57 (2017) 942–957.

[28] H. Li, Z. Gao, L. Kang, H. Zhang, K. Yang, K. Yu, X. Luo, W. Zhu, K. Chen, J. Shen, X. Wang, H. Jiang, TarFisDock: a web server for identifying drug targets with docking approach, Nucleic Acids Res. 34 (2006) W219–W224.

[29] A. Grosdidier, V. Zoete, O. Michielin, SwissDock, a protein-small molecule docking web service based on EADock DSS, Nucleic Acids Res. 39 (2011) W270–W277.

[30] S.K. Burley, H.M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J.M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D.S. Goodsell, R.K. Green, V. Guranovic, D. Guzenko, B.P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prlic, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva, C. Zardecki, RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy, Nucleic Acids Res. 47 (2019) D464–D474.

[31] Y. Shi, X. Zhang, K. Mu, C. Peng, Z. Zhu, X. Wang, Y. Yang, Z. Xu, W. Zhu, D3Targets-2019-nCoV: a webserver for predicting drug targets and for multi-target and multi-site based virtual screening against COVID-19, Acta Pharm. Sin. B 10 (2020) 1239–1248.

[32] D. Gfeller, A. Grosdidier, M. Wirth, A. Daina, O. Michielin, V. Zoete, SwissTargetPrediction: a web server for target prediction of bioactive small molecules, Nucleic Acids Res. 42 (2014) W32–W38.

[33] V. Zoete, A. Daina, C. Bovigny, O. Michielin, SwissSimilarity: a web tool for low to ultra high throughput ligand-based virtual screening, J. Chem. Inf. Model. 56 (2016) 1399–1404.

[34] H. Ozturk, A. Ozgur, E. Ozkirimli, DeepDTA: deep drug-target binding affinity prediction, Bioinformatics 34 (2018) i821–i829.

[35] J. Lim, S. Ryu, K. Park, Y.J. Choe, J. Ham, W.Y. Kim, Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation, J. Chem. Inf. Model. 59 (2019) 3981–3988.

[36] J.M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N.M. Donghia, C. R. MacNair, S. French, L.A. Carfrae, Z. Bloom-Ackermann, V.M. Tran, A. Chiappino-Pepe, A.H. Badran, I.W. Andrews, E.J. Chory, G.M. Church, E. D. Brown, T.S. Jaakkola, R. Barzilay, J.J. Collins, A deep learning approach to antibiotic discovery, Cell 180 (2020) 688–702 e613.

[37] R. Wang, X. Fang, Y. Lu, C.Y. Yang, S. Wang, The PDBbind database: methodologies and updates, J. Med. Chem. 48 (2005) 4111–4119.

[38] Y. Fukunishi, H. Nakamura, Prediction of protein-ligand complex structure by docking software guided by other complex structures, J. Mol. Graph. Model. 26 (2008) 1030–1033.

[39] Z. Wang, H. Sun, X. Yao, D. Li, L. Xu, Y. Li, S. Tian, T. Hou, Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power, Phys. Chem. Chem. Phys. 18 (2016) 12964–12975.

[40] A. Mishra, S. Dey, Molecular docking studies of a cyclic octapeptide-cyclosaplin from sandalwood, Biomolecules 9 (2019).

[41] G. Samykannu, P. Vijayababu, C.B. Antonyraj, S.B. Narayanan, S. Ahamed, Investigations of Binding Mode Insight in Salmonella typhi Type-III Secretion System Tip Protein (SipD): A Molecular Docking and MD Simulation Study, 2017.

[42] J. Eberhardt, D. Santos-Martins, A.F. Tillack, S. Forli, AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings, J. Chem. Inf. Model. 61 (2021) 3891–3898.

[43] M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Res. 44 (2016) D1045–D1053.

[44] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminf. 3 (2011) 33.

[45] J. Hu, Z. Liu, D.J. Yu, Y. Zhang, LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening, Bioinformatics 34 (2018) 2209–2218.

[46] Y. Yang, D. Zhou, X. Zhang, Y. Shi, J. Han, L. Zhou, L. Wu, M. Ma, J. Li, S. Peng, Z. Xu, W. Zhu, D3AI-CoV: a deep learning platform for predicting drug targets and for virtual screening against COVID-19, Briefings Bioinf. 23 (2022).

[47] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural Message Passing for Quantum Chemistry, 2017.

[48] J. Jo, B. Kwak, H.S. Choi, S. Yoon, The message passing neural networks for chemical property prediction on SMILES, Methods 179 (2020) 65–72.

[49] T. Hasebe, Knowledge-embedded message-passing neural networks: improving molecular property prediction with human knowledge, ACS Omega 6 (2021) 27955–27967.

[50] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing learned molecular representations for property prediction, J. Chem. Inf. Model. 59 (2019) 3370–3388.

[51] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90.

[52] J. Jimenez, M. Skalic, G. Martinez-Rosell, G. De Fabritiis, K(DEEP): protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks, J. Chem. Inf. Model. 58 (2018) 287–296.

[53] A.S. Rifaioglu, E. Nalbat, V. Atalay, M.J. Martin, R. Cetin-Atalay, T. Dogan, DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations, Chem. Sci. 11 (2020) 2531–2557.

[54] S. Hu, C. Zhang, P. Chen, P. Gu, J. Zhang, B. Wang, Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks, BMC Bioinf. 20 (2019) 689.

[55] K. Huang, T. Fu, L.M. Glass, M. Zitnik, C. Xiao, J. Sun, DeepPurpose: a deep learning library for drug-target interaction prediction, Bioinformatics 36 (2021) 5545–5547.

[56] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754.

[57] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52.

[58] V.D.M. Laurens, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[59] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082.

[60] W.H. Sauer, M.K. Schwarz, Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity, J. Chem. Inf. Comput. Sci. 43 (2003) 987–1003.

[61] C. UniProt, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res. 49 (2021) D480–D489.

[62] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan, C. Qin, Y. Li, X. Li, Y. Chen, F. Zhu, Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics, Nucleic Acids Res. 48 (2020) D1031–D1041.

[63] M. Pasha, A.H. Eid, A.A. Eid, Y. Gorin, S. Munusamy, Sestrin2 as a novel biomarker and therapeutic target for various diseases, Oxid. Med. Cell. Longev. 2017 (2017), 3296294.

[64] S.C. Johnson, P.S. Rabinovitch, M. Kaeberlein, mTOR is a key modulator of ageing and age-related disease, Nature 493 (2013) 338–345.

[65] R.X. Leng, H.F. Pan, J.H. Tao, D.Q. Ye, IL-19, IL-20 and IL-24: potential therapeutic targets for autoimmune diseases, Expert Opin. Ther. Targets 15 (2011) 119–126.

[66] M.U. Anwaar, F. Adnan, A. Abro, R.A. Khan, A.U. Rehman, M. Osama, C. Rainville, S. Kumar, D.E. Sterner, S. Javed, S.B. Jamal, A. Baig, M.R. Shabbir, W. Ahsan, T. R. Butt, M.Z. Assir, Combined deep learning and molecular docking simulations approach identifies potentially effective FDA approved drugs for repurposing against SARS-CoV-2, Comput. Biol. Med. 141 (2022), 105049.

[67] A. Anighoro, J. Bajorath, Three-dimensional similarity in molecular docking: prioritizing ligand poses on the basis of experimental binding modes, J. Chem. Inf. Model. 56 (2016) 580–587.