Quantitative Biology 2023, 11(3): 320–331
https://doi.org/10.15302/J-QB-022-0321

RESEARCH ARTICLE

Exploration on learning molecular docking with deep learning models

Qin Xie[1,†], Wei Ma[1,2,†], Jianhang Zhang[1], Shiliang Li[2], Xiaobing Deng[3,*], Youjun Xu[1,*], Weilin Zhang[1,*]

[1] Infinite Intelligence Pharma, Beijing 100083, China
[2] Shanghai Key Laboratory of New Drug Design, State Key Laboratory of Bioreactor Engineering, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China
[3] College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China
* Correspondence: dengxb@pku.edu.cn; xuyj@iipharma.cn; zhangwl@iipharma.cn

Received November 23, 2022; Revised January 19, 2023; Accepted January 29, 2023

***Background*: Molecular docking-based virtual screening (VS) aims to choose ligands with potential pharmacological activities from millions or even billions of molecules. This process could significantly cut down the number of compounds that need to be experimentally tested. However, during the docking calculation, many molecules have low affinity for a particular protein target, which waste a lot of computational resources.**

***Methods*: We implemented a fast and practical molecular screening approach called DL-DockVS (deep learning dock virtual screening) by using deep learning models (regression and classification models) to learn the outcomes of pipelined docking programs step-by-step.**

***Results*: In this study, we showed that this approach could successfully weed out compounds with poor docking scores while keeping compounds with potentially high docking scores against 10 DUD-E protein targets. A self-built dataset of about 1.9 million molecules was used to further verify DL-DockVS, yielding good results in terms of recall rate, active compounds enrichment factor and runtime speed.**

***Conclusions*: We comprehensively evaluate the practicality and effectiveness of DL-DockVS against 10 protein targets. Due to the improvements of runtime and maintained success rate, it would be a useful and promising approach to screen ultra-large compound libraries in the age of big data. It is also very convenient for researchers to make a well-trained model of one specific target for predicting other chemical libraries and high docking-score molecules without docking computation again.**

Keywords: molecular docking; ultra-large virtual screening; deep learning

Author summary: A deep learning-powered VS approach combined with two free docking programs are proposed and evaluated for screening an ultra-large compound library to obtain diverse potential active compounds rapidly and efficiently. We found that it is a practical and transferable strategy to significantly reduce computational cost.

INTRODUCTION

Drug discovery is an expensive, difficult, and drawn-out process. Depending on the diseases and treatment methods, the cost of developing approved drugs increased from $300 million to $2.8 billion during the past decades [1]. Virtual screening (VS) as a typical computational technique, has been widely applied in the early stage of drug discovery for accelerating lead discovery [2]. A target protein with a three-dimensional structure is used in docking-based virtual screening (DockVS), which seeks to computationally place and assess small molecules one at a time at the binding site. Molecular docking could explicitly suggest new

† These authors contributed equally.

compounds with high docking scores after continually improving the binding poses of the molecules.

The currently available molecular docking programs including their released years, released organizations, descriptions, and licenses, have been summarized in the Supplementary Table S1 [3–19]. As the scoring functions' definitions and sampling techniques varied greatly, they might provide different docking results for a given protein target. For real applications, it is fairly common to use two docking programs sequentially in the process of molecular virtual screening, which is shown as a well-known funnel-like workflow [20–31]. In the sequential docking workflow, the first docking process used could be the one with relative lower precision and with a fast speed, while the second docking process could be more precision but somehow slowly. Consequently, a balance between efficiency and docking quality is finally reached. After this filter, further evaluations with higher accuracy but low speed as well as human visual inspection were applied to choose compounds for experimental verification.

Recently, it has been reported that docking based VS processes on ultra-large databases can significantly increase the success rate of finding active hits [32]. Meanwhile, the size of compound databases has also increased significantly. For example, the widely used ZINC database has grown from 700 thousand in 2005 to over 1.3 billion in 2019 [33–35]. By using parallel distributed cloud computing capabilities on the Apache Spark engine, Capuccini *et al.* [36] carried out the VS procedure on a large-scale compound library and achieved a good parallel efficiency. However, only a tiny part of the compounds will have high docking scores and the great majority of the low-score compounds that were docked occupied most of the computational resources. It is desirable for researchers to identify high-score compounds using trustworthy machine learning methods in advance, therefore the computational resources for docking could be significantly reduced.

Gentile *et al.* [37,38] developed a deep learning (DL) classification model named as deep docking (DD) based on the docking scores of a subset of compounds to speed up the prediction of the remaining compounds from an ultra-large library of a billion compounds (ZINC-15). Several active learning-based workflows have recently been proposed [39–43]. Yang *et al*. used DOCK3.7 and Glide SP to do an active learning process. They used a 0.1% random subset of the total library and then iteratively updated the DL models with an active learning strategy. This VS process could recover the majority of the top-ranking compounds. However, it still needs to dock about 5% of the total compounds which is still a greater computational burden since DD that only dock 1%. As suggested by Graff *et al.* [44], pruning the searching space is necessary to accelerate the whole screening process. Therefore, it is highly valuable to investigate and develop an accelerated VS approach for reducing the computational burden of molecular docking on low-score compounds in real-world scenarios. Scoring function is often used to estimate potential binding affinities of given compounds against given targets of interest. It is intuitive and straight-forward to build regression models to fit such affinity scores produced by certain docking program. Machine learning- and deep learning-based scoring functions have been developed with a pronounced trend [45,46]. The speed and accuracy of prediction have been improved in these processes. However, scoring functions in many real-world cases do not correlate well with the binding affinities. Moreover, a significant obstacle is how effectively one scoring function can discriminate true actives from numerous inactives [45].

To improve the reliability of docking assessments, multiple docking programs have been used in this study. Considering the fundamental principles of docking programs and the potential advantages of deep learning, we developed a new rapid and practical VS strategy, called DL-DockVS, which is composed of a set of two deep learning models (DL-DockVS-R as regression models and DL-DockVS-C as classification models) in tandem to mimic the funnel-like VS procedure which could eliminate many decoy compounds. We utilized the output scores from two docking programs to construct DL-DockVS-R and DL-DockVS-C models. The rationale for building one regression model of percentile and a classification model is illustrated as follows: for different scoring functions, various scales of them often make it hard to determine a special threshold value. The ranking order instead of the absolute scores of docking scores are essential for the first filtration, as the regression task for DL-DockVS-R models. And it can easily be used to predict and select top ranking compounds from an ultra-large library. Subsequently, these top-ranking compounds are predicted by the DL-DockVS-C models to return potential active labels which would be chosen for further evaluation. With comprehensive evaluations, the DL-DockVS approach was shown to have a decent capacity to find known active compounds for 10 specific protein targets. The active enrichment factor was in range of 5 and 17. Further analysis of the target BRAF revealed that DL-DockVS had a high screening speed. These findings demonstrate that this approach can confidently and swiftly screen out molecules with low docking scores which are probably inactive. It is considered to be a good and useful filter in routine VS processes on the ultra-large chemical libraries.

RESULTS AND DISCUSSION

DL-DockVS model performance

To reflect the consistency of the model's performance and its generalizability, we summarized the training statistical metrics of those models for the 10 selected targets in Supplementary Fig. S1. The area under curve (AUC) and root mean squared error (RMSE) metrics on the test sets of the DL-DockVS models for each target are presented in Table 1. For those DL-DockVS-R models, the RMSE values range from 0.11 to 0.14, illustrating an acceptable error in ranking prediction. For those DL-DockVS-C models, the AUC values range from 0.89 to 0.96, indicating an excellent ability to distinguish between positive data (top 20% ranked compounds by Dock-2) and negative data (the rest of the compounds by Dock-2).

Evaluations of DL-DockVS-R models

Model evaluation on external test sets

Considering there may be topological bias in the DUD-E dataset, we did not intentionally incorporate these compounds into the training set, and just used it as one external test set only [47]. We labelled these as follows: DUD-E actives/decoys were docked by VinaLC and compounds with docking scores above the top 20% percentile of the Training-Set1 (See Experimental) were selected. The selected compounds were compared with the compounds selected by the DL-DockVS-R models. The results are listed in Table 2. It is shown that among the 10 targets, except the target of ACE, the consistency rates between the VinaLC docking and DL-DockVS-R models were more than 75.0%, with the highest consistency rate of 98% for the CDK2. These results suggest that the DL-DockVS-R models are actually good at learning the filter process of the top 20% ranking and achieve an acceptable performance compared to corresponding docking methods.

Model evaluation on the ChEMBL subset

We used DL-DockVS-R models to predict the ranking percentage of the compounds in the ChEMBL subset. The compounds with predicted top 20% rankings were further docked to their targets using VinaLC to test their real distributions. Figure 1 and Supplementary Fig. S2 show the VinaLC docking score distributions of those selected compounds from the ChEMBL subset, the whole ChemDiv dataset, and the top 20% ranked compounds against each protein target. In Fig. 1, we can see that most of the selected ChEMBL compounds are located in the top 20% ranked compounds region of the

Table 1 Performance of DL-DockVS-R and DL-DockVS-C models based on the test set from ChemDiv subset against 10 protein targets

Target	DL-DockVS-R RMSE	DL-DockVS-C AUC
ACE	0.1237±0.0008	0.9601±0.0030
ADRB1	0.1210±0.0008	0.9478±0.0043
BRAF	0.1323±0.0007	0.9473±0.0047
CDK2	0.1177±0.0009	0.9448±0.0028
DRD3	0.1118±0.0010	0.9478±0.0037
DPP4	0.1210±0.0006	0.9257±0.0027
EGFR	0.1235±0.0006	0.9611±0.0035
JAK2	0.1340±0.0007	0.9124±0.0039
LCK	0.1220±0.0008	0.9604±0.0014
VGFR2	0.1425±0.0005	0.9556±0.0071

Table 2 Performance of DL-DockVS-R models and VinaLC docking results on the external test set from DUD-E

Model	Comp.	No. of comp. selected by Dock-1 (top 20%)	No. of comp. selected by DL-DockVS-R	Consistency rate*
ACE	Actives	55	31	56.4%
	Decoys	745	643	86.3%
ADRB1	Actives	60	51	85.0%
	Decoys	679	575	84.7%
BRAF	Actives	98	88	89.8%
	Decoys	949	779	82.1%
CDK2	Actives	245	240	98.0%
	Decoys	1341	1091	81.4%
DRD3	Actives	170	136	80.0%
	Decoys	1437	1214	84.5%
DPP4	Actives	84	66	78.6%
	Decoys	1500	1190	79.3%
EGFR	Actives	225	216	96.0%
	Decoys	2788	2308	82.8%
JAK2	Actives	34	26	76.5%
	Decoys	395	302	76.5%
LCK	Actives	291	253	86.9%
	Decoys	2096	1,737	82.9%
VGFR2	Actives	158	120	76.0%
	Decoys	1312	1086	82.8%

* Consistency rate is used to evaluate the performance of models to retrieve the top compounds ranked by Dock-1

$$\text{Consistency rate} = \frac{N_{DL\_DockVS\_R}}{N_{Dock\_1}}$$

where $N_{DL\_DockVS\_R}$ is the number of compounds selected by DL_DockVS_R, N_{Dock_1} is the number of top 20% ranked compounds selected by Dock_1.

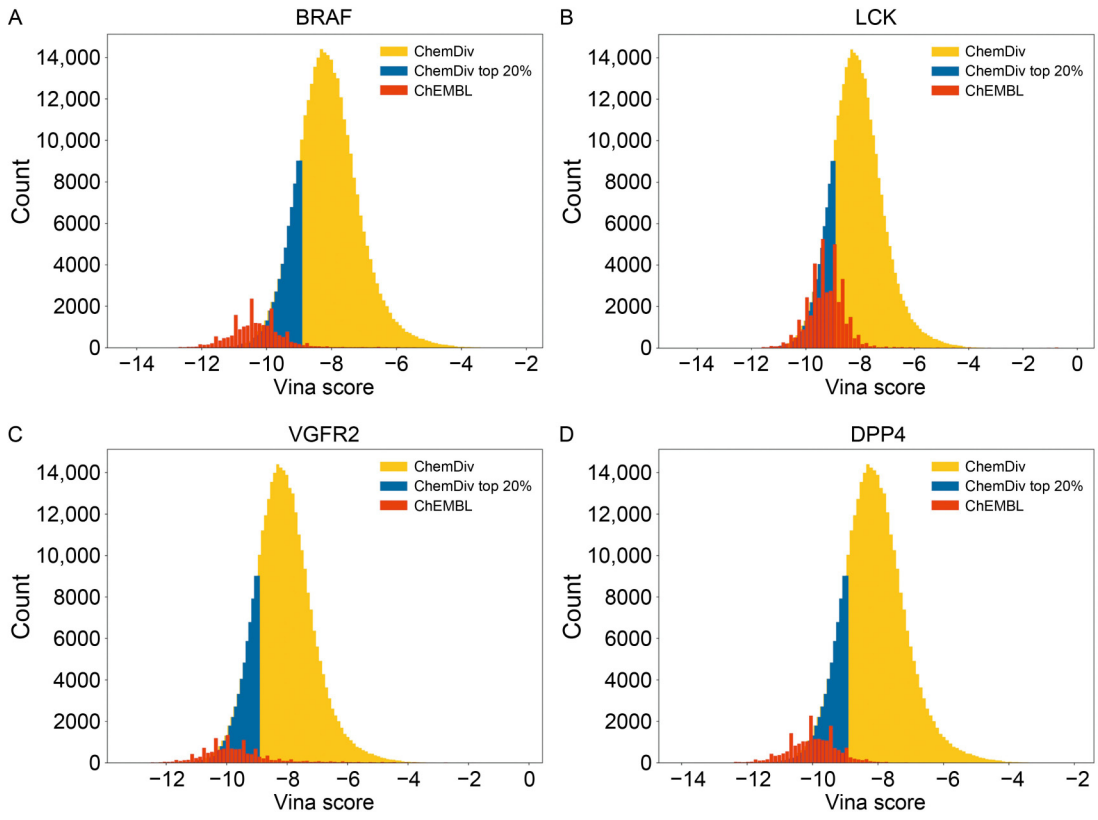

Figure 1. Docking score distributions of the selected compounds from DL-DockVS-R model, ChEMBL and ChemDiv for four targets, (A) BARF (B) LCK (C) VGFR2 and (D) DPP4. The yellow bins represent the distribution for ChemDiv, the blue bins represent the distribution for the top 20% compounds from ChemDiv, and the red bins represent the distribution from the DL-DockVS-R prediction results of the selected ChEMBL subset compounds..

ChemDiv subset. It suggests that our DL-DockVS-R models have a strong ability to enrich compounds with a high docking score. Therefore, the models can distinguish compounds with potentially high docking scores for specific targets in a quick manner without real docking calculation for the whole dataset.

Evaluation of DL-DockVS-C models on external test sets

In the previous section, the DUD-E actives/decoys docked by VinaLC and selected compounds with the top 20% ranking percentile of the Training Set1 were chosen then further docked by rDock. The compounds with docking scores of above the top 20% percentile of the Training Set2 were labeled as positive, while the bottom 20% of them were labelled as negatives. These compounds were further predicted by the DL-DockVS-C models. Table 3 shows the statistical results of the various metrics on the DUD-E external data set. The results show that, except the target of DPP4, the accuracies of the other 9 targets are above 70%. The specificity values of the 8 targets are above 72%, except for the two targets of DPP4 and BRAF. As for the true positive rates (TPRs), the maximum and minimum values were 89.5% of BRAF and 27.5% of JAK2. It is obvious that the performances of DL-DockVS-C on various protein targets differ. Generally, the DL-DockVS-C models show good performance in terms of accuracy and specificity on the external test sets from DUD-E.

Performance of DL-DockVS models on retrieving active compounds

To further verify whether the DL-DockVS models can identify active compounds towards specific targets, the active compounds for 10 targets were collected from ChEMBL. The active compounds for 10 targets were filtered according to a set of activity thresholds. For instance, if a compound has an IC_{50}/EC_{50} value of less than 50 μM, it is labeled as a positive sample, otherwise negative.

To validate the abilities of recalling actives, about 1.9 million molecules from the ChEMBL database were first predicted with the DL-DockVS-R models. Then the compounds with the predicted top 20% ranking were predicted by the DL-DockVS-C models. The results are

described in Table 4, showing the percentage of the retrieved active compounds under the different activity cutoffs. The statistical results indicated that most of the models can identify over 50% of actives towards specific targets, such as the recall percentages of ADRB1, BRAF, EGFR, JAK2, LCK, and VGFR2 were 65.78%, 79.30%, 68.63%, 51.88%, 74.61%, and 52.72%, respectively. It means most of the active compounds could be recalled by our DL-DockVS models for most targets.

The final number of compounds with positive label as about 4% of the total compounds predicted. Which means our model's real behavior is somehow mimic the step-by-step docking process (20%×20% as the real process returned). Therefore, the enrichment factor here we denoted as $EF_{0.04}$. The summarized results for 10 targets are listed in Table 5. From this table, the $EF_{0.04}$ value ranges from the lowest 5.86 (CDK2) to the highest 17.03 (ADRB1). It revealed that DL-DockVS models had a good enrichment ability of active compounds.

Running time comparison

To demonstrate the speed advantages of DL-DockVS, the target BRAF was selected as a case. The drug-like molecules (containing 981,247,974 compounds) from ZINC-15 were selected in this case. The statistical results in Table 6 show that, based on the above available computing resources, the prediction speed of the DL-DockVS is greatly improved compared to traditional docking processes, with an average speed improvement about 2000 times. DL-DockVS could quickly filter out low-score compounds against specific targets. Moreover, the compounds with high docking scores (top 4%) were obtained at an acceptable rate.

In the DD protocol [38], DD needed to dock 1% of the molecules while retrieving 90% of the best-scoring structures. The DL-DockVS approach has similar performance as shown on an example in Table 6. It weeded out 99.67% low-score compounds on the BRAF targets from the whole dataset.

Such high rate of clearance suggested that the chance to obtain more tight binding/high docking score compounds is at a computational cost of even more low scored one for an ultra-large library virtual screening. Based on the result from Table 6, for example, when screening 10–20 million purchasable compounds

Table 3 Performance of DL-DockVS-C models on external test sets from DUD-E

No.	Target	Accuracy	Precision	Specificity	FPR	TPR	F1-score
1	CDK2	73.3%	59.3%	82.2%	17.8%	53.8%	56.4%
2	EGFR	72.9%	49.7%	75.2%	24.8%	66.3%	56.8%
3	DPP4	60.4%	41.8%	53.1%	46.9%	77.2%	54.3%
4	ACE	73.9%	52.0%	84.1%	15.9%	46.3%	49.0%
5	DRD3	76.0%	60.2%	89.7%	10.3%	40.4%	48.3%
6	ADRB1	79.9%	66.5%	85.5%	14.5%	54.6%	60.0%
7	LCK	77.5%	63.4%	90.3%	9.7%	44.1%	52.0%
8	JAK2	74.3%	46.3%	89.5%	10.5%	27.5%	34.5%
9	VGFR2	72.9%	62.2%	90.7%	9.3%	33.8%	43.8%
10	BRAF	82.8%	87.9%	62.1%	37.9%	89.5%	88.7%

Table 4 Active compounds prediction statistics of the DL-DockVS models

Target	Threshold of active compounds retrieved from ChEMBL (IC50)							
	<50 nM	<100 nM	<200 nM	<500 nM	<1 μM	<10 μM	<20 μM	<50 μM
ACE	30.00%	31.56%	32.73%	34.46%	33.42%	32.35%	31.83%	31.10%
ADRB1	44.62%	55.81%	60.68%	62.89%	64.44%	66.81%	65.34%	65.78%
BRAF	79.01%	79.18%	79.37%	79.26%	79.14%	79.23%	79.32%	79.30%
CDK2	23.29%	23.08%	23.34%	24.22%	24.35%	29.41%	29.41%	28.80%
DPP4	34.81%	34.64%	34.24%	33.12%	32.55%	32.18%	32.34%	32.34%
DRD3	28.97%	29.45%	33.51%	35.68%	38.91%	42.17%	42.04%	42.43%
EGFR	74.45%	74.30%	73.17%	72.46%	69.84%	68.88%	68.96%	68.63%
JAK2	55.95%	55.03%	53.59%	52.39%	51.97%	51.78%	51.82%	51.88%
LCK	77.97%	78.19%	78.94%	78.31%	77.74%	74.77%	74.66%	74.61%
VGFR2	36.53%	40.93%	43.62%	46.96%	48.65%	52.58%	52.59%	52.72%

Table 5 Statistics for top 4% enrichment factor against 10 targets

Target	No. of actives	No. of positive predictions	No. of true positives	$EF_{4\%}$
ACE	553	72,638	172	8.32
ADRB1	529	74,995	348	17.03
BRAF	5222	94,479	4141	16.34
CDK2	1684	95,447	485	5.86
DPP4	3779	87,247	1222	7.21
DRD3	337	85,286	143	9.66
EGFR	8232	134,286	5650	9.96
LCK	1713	93,253	1278	15.55
JAK2	5181	108,679	2688	9.29
VGFR2	9329	71,504	4918	14.38

Table 6 Comparison of running time from docking process and DL-DockVS prediction

No. compounds	Time cost (h)	No. of predicted high-score compounds[a]	Estimated time of docking[b] (h)	Time acceleration ratio[c]
200 K	0.018	1263±35	33	1823
500 K	0.041	3169±47	82	2001
1 M	0.079	6336±78	164	2077
2 M	0.157	12,666±106	328	2090
5 M	0.391	31,643±191	820	2098
10 M	0.779	63,300±233	$1.6*10^3$	2054
20 M	1.557	126,689±343	$3.3*10^3$	2119
50 M	3.891	316,570±585	$8.2*10^3$	2107
100 M	7.779	633,155±709	$1.6*10^4$	2057
500 M	38.891	3,166,069±1247	$8.2*10^4$	2108
981 M	77.586	6,213,291	$1.6*10^5$	2062

Notes: Model prediction computing resource: 10 * CPUs (Xeon Gold 5118) +1 * GPU (Tesla V100); Molecular docking computing resources: 20 * CPUs (Xeon E5-2670 V2); speed improvement based on the above resource allocation. [a]The average and standard deviation of the predicted high scoring compounds number by randomly unbiased sample 100 times. [b]The docking time is linearly predicted according to the docking speed of 30 K ChemDiv compounds. [c]Time acceleration ratio= Estimated time of docking / time cost-prediction.

worldwide [48] about 63−126 thousand compounds are predicted by DL-DockVS for further evaluation, such an order of magnitude within the range of the ability of normal docking computation.

DISCUSSION

Due to the availability and accessibility of the docking software, we only tested the combination of VinaLC and rDock, which tends to be orthogonal to each other. Conceptionally, this approach is believed to be general to other combinations of docking programs, and flexible to accommodate different requirements in variant projects.

To explore a reasonable threshold in the process of obtaining the training dataset, the DUD-E active compounds were selected to carry out statistical analysis on the relationships between the enrichment ratio of the active compounds and the top value of the two docking programs for 10 selected targets, respectively. Representative results of 4 targets were listed in Fig. 2. Others are shown in Supplementary Fig. S3. The *x*-axis represents the percentile of the top 20% ranked compounds sorted by the docking score, and the *y*-axis represents the proportion of DUD-E active compound enrichment status for the corresponding target. For most of targets, top 20% of the compounds docking by two docking programs can enrich 40% of active compounds in DUD-E. Thus, the top 20% was selected for the preparation of the training dataset. Overall, the active compound enrichment ability of rDock is better than VinaLC

To obtain a higher enrichment ratio, we further explored the usage ordering of two docking programs. The enrichment ratios of different-ordering combinations are summarized in Fig. 2. The red curve represents the enrichment ratio of the corresponding target by the rDock program, the blue one represents that of the VinaLC program, and the green one represents the re-docking of the top 20% compounds using the rDock after docking with VinaLC. Consequently, the orange one represents the enrichment ratio in the order of rDock first and then VinaLC.

As shown in Fig. 2A and 2B, when the performance of two programs is relatively similar under a certain target, it can be observed that the combination of two docking programs is better than the use of a single docking program. Nevertheless, for a common circumstance, programs with good performance may be limited in use for various reasons and cannot be used on a large scale. In this case, we can only use a slightly weaker program for large-scale screening. According to Fig. 2C and 2D, it can be observed that the end points on the red lines are higher than that of the blue lines, indicating that the performance of the two docking programs is quite different and the red one get a good enrichment performance. Because of the two targets docked by VinaLC had a low enrichment ratio of DUD-E active compounds, VinaLC is a step of limiting enrichment ratio. Therefore, the serial use of VinaLC and rDock does not lead to an improvement in the final performance. In this instance, the green line can be considered as a simulation of selecting a program (such as VinaLC) with a slightly lower enrichment rate that can be quickly docked for rough screening in the first step, and then use the docking program (such as rDock)

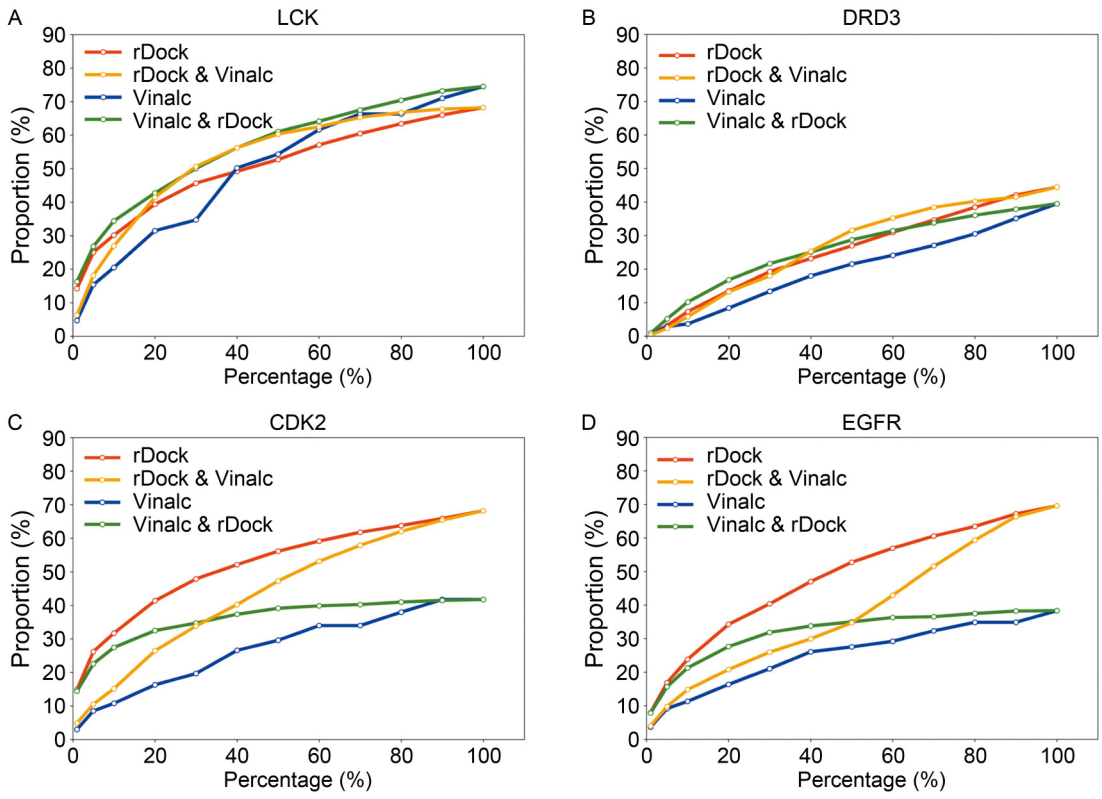

Figure 2. Enrichment ratio of DUD-E active compounds for two docking programs.

with a higher enrichment rate for docking in the second step. The figures show that the end points of the green lines are higher than that of the blue ones and illustrates that this strategy can improve the active enrichment capability compared with a single docking program (blue lines). However, the loss of active enrichment could not be avoided compared to the programs that have a good performance (red lines).

CONCLUSION

In this work, we demonstrated a fast and reusable approach, DL-DockVS, which mimics the funnel-like screening process using deep learning models. We implemented it by utilizing the docking results of certain targets on a commercially purchasable ChemDiv subset with two open-source docking programs (VinaLC and rDock). Various statistical results showed that the models could have good performance in the VS process. Moreover, compared with the traditional VS process, the DL-DockVS we constructed can realize virtual screening on an ultra-large compound library with an improvement of runtime speed. These well-trained models could be easily extended to predict other chemical libraries. With the applications of molecular generative models in the field of drug development in recent years, it is believed that DL-DockVS would be used as a rapid scoring metric for deep generative models to efficiently produce a set of diverse compounds with high docking scores towards a specific protein target.

METHODS

Model construction and usages

The workflow of the DL-DockVS construction and usage is illustrated in Fig. 3. For a certain protein target**,** two supervised training processes are implemented. For the first training process, the compounds from the training library (N compounds) are docked onto the given target. Dock-1 is used as the first docking program, and all the docking results are used as the first training set (Training Set1) for this target. The compounds with the top 20% ranking from the first docking results are selected for the second docking process, and the docking results from this setup are used as the second training set for this target (Training Set2). We then label these training sets according to the docking results. For the first regression task, the ranking percentage values of the compounds in Training Set1 are used as ground truth labels. For the classification task, the top 20% ranked compounds within the Training Set2 are selected as positive samples, and the remaining

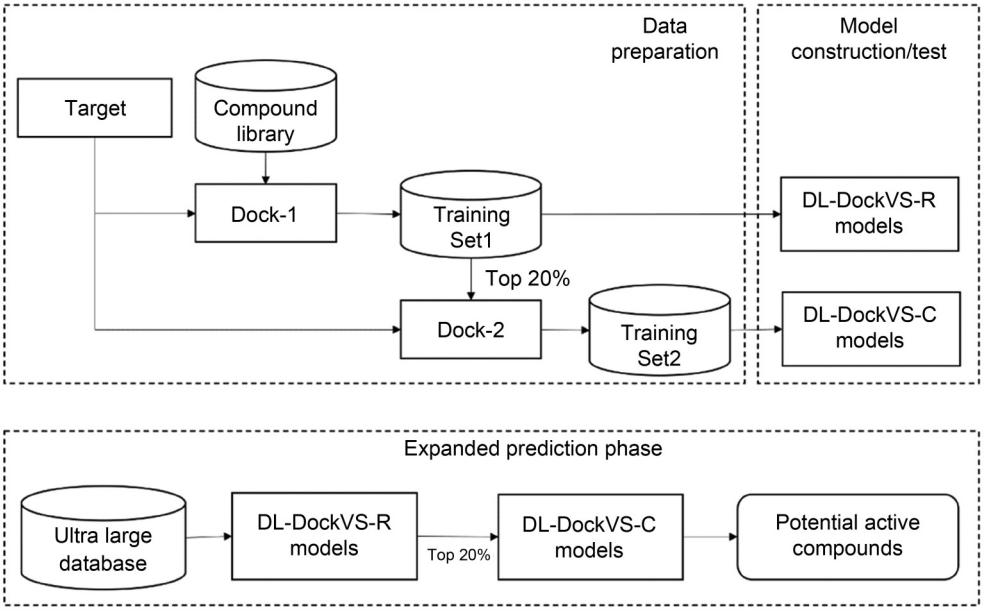

Figure 3. The workflow of DL-DockVS construction and usage.

compounds are used as negative samples. The regression models (DL-DockVS-R) and the classification models (DL-DockVS-C) are trained based on Training Set1 and Training Set2, respectively. After evaluation of the models, they can be used for ultra-large library filtering. During ultra-large library filtering, the DL-DockVS-R models are first used to obtain the compounds which would have scores with top 20% ranked values, and then these compounds are predicted by the DL-DockVS-C models. The compounds with positive labels can be used for further examination.

Molecular docking programs and parameter settings

We summarized available molecular docking programs in Supplementary Table S1. Due to intellectual property issues, two open-source and freely available docking programs, namely VinaLC [49,50] (*vina*-based implementation for large datasets) and rDock [16], were selected in this research. VinaLC is based on an experience scoring function developed by the MGL Tools laboratory. Compared to AutoDock 4.0, VinaLC improves the average accuracy of binding mode prediction and accelerates the search speed by using a simpler scoring function. rDock is a fast and versatile program that is used to dock molecules against both proteins and nucleic acids. Both programs can be installed and run on a computing cluster with a lot of CPU cores. This makes it possible for VS procedures to be done in a few days.

In our cases, the protein structures were preprocessed using ADTools, and the small molecules were prepared using OpenBabel [51]. For the parameters used in *vina*, the grid center was defined as the native ligand center in the crystal structure, a cubic grid was $20 \times 20 \times 20$. For rDock, the *rbcavity* program is used to determine a grid parameter from the native ligand with default parameters. VinaLC was selected as the first docking program (Docking-1), and rDock was selected as the second docking program (Docking-2). Both docking scores of Docking-1 and Docking-2 were saved as csv-format files (Additional file 3).

Deep learning model construction

An open source toolkit, Chemprop (github website (chemprop/chemprop)), is a deep learning based molecular property modeling and prediction toolkit [52]. It was used to develop our DL-DockVS models. The architecture is based on graph convolutional networks, which treats molecules as an attributed graph with node features (atoms) and edge features (bonds) for processing. This toolkit was used to develop DL-DockVS-R and DL-DockVS-C models with default hyperparameters. Besides, the setting parameters of *rdkit_2d_normalized* feature and *no_features_scaling* are added to the training process to improve the generalization capability of the models. Both the DL-DockVS-R and DL-DockVS-C models were trained and validated on 80% of the random split set and tested on the remaining 20% of the datasets. The hyperparameters, such as *hidden_size*, depth, dropout, and *ffn_num_layers*, were optimized by 10-fold cross validation. The

optimal hyperparameters were used to construct the DL-DockVS-R and DL-DockVS-C models.

Model evaluation metrics

For the regression tasks, the RMSE is calculated to evaluate the performance of models to reproduce the docking ranking. For the classification tasks, the metrics of AUC, Accuracy, Precision, Specificity, FPR (false positive rate), TPR, and F1-score are calculated to evaluate the model performance on external datasets. The formulas are as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (2\text{-}1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2\text{-}2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2\text{-}3)$$

$$FPR = \frac{FP}{FP + TN} \quad (2\text{-}4)$$

$$TPR = \frac{TP}{TP + FN} \quad (2\text{-}5)$$

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2\text{-}6)$$

The above formulas are the evaluation metrics of classification models. TP is true positive. TN is true negative. FP is false positive. FN is false negative. Accuracy is the percentage of correct predictions for the test data. Precision is defined as the fraction of correctly classified samples among all the data points that are predicted to belong to a certain class. Specificity is the proportion of the correctly predicted the negative samples. TPR is the fraction of correctly classified samples among all data points that are predicted to belong to a certain class. F1-score is the harmonic mean of Precision and Recall.

Datasets and their usages

In this study, several datasets were used for different purposes.

ChemDiv subset. For the compound library to be docked into each target, a ChemDiv library with 1.25 million purchasable compounds was clustered with a Tanimoto similarity threshold of 0.4 to yield a clustered subset of 287,216 compounds (named as the ChemDiv subset, and the SMILES-format files of the compounds are available in **Additional file 2**).

Test set from DUD-E dataset. The descriptions of the used target proteins are listed in Table 7. This table also includes the numbers of their experimentally validated active compounds as well as decoy compounds from the DUD-E database [53] and the evaluation of the crystal structure restoration ability of VinaLC and rDock. This table shows that all RMSD-rdock and RMSD-VinaLC are less than 2 Å. Therefore, VinaLC and rDock are suitable for these target systems. They are used as validation sets for DL-DockVS. All these compounds were also docked with two programs to corresponding targets, but were not used in the training process. Their SMILES strings are listed in Additional file 4.

Test set from ChEMBL dataset. In an external test stage, we tested DL-DockVS performance on two self-constructed datasets. To evaluate DL-DockVS-R models, a relatively small dataset was constructed by random selecting 500,000 compounds from the ChEMBL database (Their SMILES strings are listed in Additional file 5). To evaluate DL-DockVS-C performance, the compounds from the ChEMBL database are labeled as active or inactive according to the IC_{50} value of each target. Their SMILES strings are provided in Additional file 6. Therefore, the whole ChEMBL27

Table 7 Detailed information of the selected targets and their corresponding active/decoy compounds numbers

No.	Target	PDB-ID	Type	Actives	Decoys	RMSD-rDock	RMSD-VinaLC
1	ACE	3BKL	Protease	808	17,144	1.67	1.56
2	ADRB1	4BVN	GPCR	458	15,958	0.84	0.54
3	BRAF	3D4Q	Kinase	251	10,098	1.09	0.97
4	CDK2	1H00	Kinase	798	28,328	1.59	1.56
5	DRD3	3PBL	GPCR	877	34,188	0.65	1.15
6	DPP4	2I78	Protease	1,079	41,373	1.07	0.40
7	EGFR	2RGP	Kinase	832	35,442	1.20	1.89
8	JAK2	3LPB	Kinase	153	6590	1.82	1.20
9	LCK	2OF2	Kinase	683	27,856	0.97	1.13
10	VGFR2	2P2I	Kinase	620	25,280	0.75	0.64

database containing about 1.9 million molecules were predicted then filtered by the DL-DockVS-R models, and those filtered molecules would be used as the input to the DL-DockVS-C models.

Test set from ZINC. The ZINC-15 dataset was used to illustrate the speed of DL-DockVS on a large dataset for target BARF as an example.

DATA AVAILABILITY

The Python codes are available at github website (OpenIIPharma/mlddm). The MOL2 and CSV files of the clustered compounds from ChemDiv are available in Additional file 2. Docking scores of Training Set1 and the following Training Set2 for each target were saved as csv files and provided in Additional file 3. The SMILES, MOL2, SDF of DUD-E compounds and PDB of receptors used for validation are provided in Additional file 4. The SMILES of 500,000 compounds randomly selected from the ChEMBL database are provided in Additional file 5. The SMILES of compounds with activities from the ChEMBL database for each target are provided in Additional file 6. All these Additional files can be downloaded at zenodo website (5665378).

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.15302/J-QB-022-0321.

ACKNOWLEDGEMENTS

We thanked Dr. Jianfeng Pei for valuable discussions related to deep learning and computer-aided drug discovery. Part of the computation and analysis were performed on the High Performance Computing Platform of the Peking-Tsinghua Center for Life Sciences, Peking University. We also thanked Dr. Zihao Shen, Dr. Fangjin Chen, Ms. Ting Fang for their help in the support of computational resources. This work is supported by the funding from Infinite Intelligence Pharma Ltd.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Qin Xie, Wei Ma, Jianhang Zhang, Shiliang Li, Xiaobing Deng, Youjun Xu and Weilin Zhang declare that they have no conflict of interest or financial conflicts to disclose.

All procedures performed in studies involving animals were in accordance with the ethical standards of the institution or practice at which the studies were conducted, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

OPEN ACCESS

REFERENCES

1. Wouters, O. J., McKee, M. and Luyten, J. (2020) Estimated research and development investment needed to bring a new medicine to market, 2009-2018. JAMA, 323, 844−853
2. Ban, F., Dalal, K., Li, H., LeBlanc, E., Rennie, P. S. and Cherkasov, A. (2017) Best practices of computer-aided drug discovery: lessons learned from the development of a preclinical candidate for prostate cancer with a new mechanism of action. J. Chem. Inf. Model., 57, 1018−1028
3. Kurcinski, M., Pawel Ciemny, M., Oleniecki, T., Kuriata, A., Badaczewska-Dawid, A. E., Kolinski, A. and Kmiecik, S. (2019) CABS-dock standalone: a toolbox for flexible protein-peptide docking. Bioinformatics, 35, 4170−4172
4. Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A. and Kmiecik, S. (2015) CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. Nucleic Acids Res., 43, W419−W424
5. Tsuji, M., Shudo, K. and Kagechika, H., (2017) Identifying the receptor subtype selectivity of retinoid X and retinoic acid receptors via quantum mechanics. FEBS Open Bio, 7, 391–396
6. Grosdidier, A., Zoete, V. and Michielin, O. (2007) EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. Proteins, 67, 1010−1025
7. Campagna-Slater, V., Pottel, J., Therrien, E., Cantin, L. D. and Moitessier, N. (2012) Development of a computational tool to rival experts in the prediction of sites of metabolism of xenobiotics by p450s. J. Chem. Inf. Model., 52, 2471−2483
8. Lee, H., Heo, L., Lee, M. S. and Seok, C. (2015) GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. Nucleic Acids Res., 43, W431−W435
9. Shin, W. H., Lee, G. R., Heo, L., Lee, H., and Seok, C. (2014) Prediction of protein structure and interaction by galaxy protein modeling programs. On the website of researchgate
10. van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., van Dijk, M., de Vries, S. J. and Bonvin, A. M. J. J. (2016) The haddock2.2 webserver: user-friendly integrative modeling of biomolecular complexes. J. Mol. Biol., 428, 720−725
11. Dominguez, C., Boelens, R. and Bonvin, A. M. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc., 125, 1731−1737
12. Jiménez-García, B., Roel-Touris, J., Romero-Durana, M., Vidal, M., Jiménez-González, D. and Fernández-Recio, J. (2018) LightDock: a new multi-scale approach to protein-protein docking. Bioinformatics, 34, 49−55

13. Meier, R., Pippel, M., Brandt, F., Sippl, W., and Baldauf, C. (2010) Paradocks: a framework for molecular docking with population-based metaheuristics. J. Chem. Inf. Model., 50, 879−889
14. Pei, J., Wang, Q., Liu, Z., Li, Q., Yang, K. and Lai, L. (2006) PSI-DOCK: towards highly efficient and accurate flexible ligand docking. Proteins, 62, 934−946
15. McMartin, C. and Bohacek, R. S. (1997) QXP: powerful, rapid computer algorithms for structure-based drug design. J. Comput. Aided Mol. Des., 11, 333−344
16. Ruiz-Carmona, S., Alvarez-Garcia, D., Foloppe, N., Garmendia-Doval, A. B., Juhos, S., Schmidtke, P., Barril, X., Hubbard, R. E. and Morley, S. D. (2014) rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLOS Comput. Biol., 10, e1003571−e1003578
17. Morley, S. D. and Afshar, M. (2004) Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock. J. Comput. Aided Mol. Des., 18, 189−208
18. Majeux, N., Apostolakis, M. S. J., Ehrhardt, C. and Caflisch, A. (1999) Exhaustive docking of molecular fragments on protein binding sites with electrostatic solvation. Proteins, 37, 88−105
19. Koes, D. R., Baumgartner, M. P. and Camacho, C. J. (2013) Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J. Chem. Inf. Model., 53, 1893−1904
20. Onawole, A. T., Kolapo, T. U., Sulaiman, K. O. and Adegoke, R. O. (2018) Structure based virtual screening of the Ebola virus trimeric glycoprotein using consensus scoring. Comput. Biol. Chem., 72, 170−180
21. Feher, M. (2006) Consensus scoring for protein-ligand interactions. Drug Discov. Today, 11, 421−428
22. Mavrogeni, M.E., Pronios, F., Zareifi, D., Vasilakaki, S., Lozach, O., Alexopoulos, L., Meijer, L., Myrianthopoulos, V. and Mikros, E. (2018) A facile consensus ranking approach enhances virtual screening robustness and identifies a cell-active DYRK1α inhibitor. Future Med. Chem., 10, 2411−2430
23. Houston, D. R. and Walkinshaw, M. D. (2013) Consensus docking: improving the reliability of docking in a virtual screening context. J. Chem. Inf. Model., 53, 384−390
24. Berenger, F., Vu, O. and Meiler, J. (2017) Consensus queries in ligand-based virtual screening experiments. J. Cheminform., 9, 60
25. Masters, L., Eagon, S. and Heying, M. (2020) Evaluation of consensus scoring methods for AutoDock Vina, smina and idock. J. Mol. Graph. Model., 96, 107532
26. Onawole, A. T., Sulaiman, K. O., Adegoke, R. O., and Kolapo, T. U. (2017) Identification of potential inhibitors against the Zika virus using consensus scoring. J. Mole. Graphi., 73, 54–61
27. Wang, R. and Wang, S. (2001) How does consensus scoring work for virtual library screening? An idealized computer experiment. J. Chem. Inf. Comput. Sci., 41, 1422–1426
28. Yang, J. M., Chen, Y. F., Shen, T. W., Kristal, B. S. and Hsu, D. F. (2005) Consensus scoring criteria for improving enrichment in virtual screening. J. Chem. Inf. Model., 45, 1134−1146
29. Clark, R. D., Strizhev, A., Leonard, J. M., Blake, J. F. and Matthew, J. B. (2002) Consensus scoring for ligand/protein interactions. J. Mol. Graph. Model., 20, 281−295
30. Liu, S., Fu, R., Zhou, L. H. and Chen, S. P. (2012) Application of consensus scoring and principal component analysis for virtual screening against β-secretase (BACE-1). PLoS One, 7, e38086
31. Paul, N. and Rognan, D. (2002) ConsDock: a new program for the consensus analysis of protein-ligand interactions. Proteins, 47, 521−533
32. Gorgulla, C., Boeszoermenyi, A., Wang, Z. F., Fischer, P. D., Coote, P. W., Padmanabha Das, K. M., Malets, Y. S., Radchenko, D. S., Moroz, Y. S., Scott, D. A., *et al*. (2020) An open-source drug discovery platform enables ultra-large virtual screens. Nature, 580, 663−668
33. Sterling, T. and Irwin, J. J. (2015) Zinc 15—ligand discovery for everyone. J. Chem. Inf. Model., 55, 2324−2337
34. Irwin, J. J. and Shoichet, B. K. (2005) ZINC—a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model., 45, 177−182
35. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. and Coleman, R. G. (2012) ZINC: a free tool to discover chemistry for biology. J. Chem. Inf. Model., 52, 1757−1768
36. Capuccini, M., Ahmed, L., Schaal, W., Laure, E. and Spjuth, O. (2017) Large-scale virtual screening on public cloud resources with Apache Spark. J. Cheminform., 9, 15
37. Gentile, F., Agrawal, V., Hsing, M., Ton, A. T., Ban, F., Norinder, U., Gleave, M. E. and Cherkasov, A. (2020) Deep docking: a deep learning platform for augmentation of structure based drug discovery. ACS Cent. Sci., 6, 939−949
38. Gentile, F., Yaacoub, J. C., Gleave, J., Fernandez, M., Ton, A.-T., Ban, F., Stern, A. and Cherkasov, A. (2022) Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. Nat. Protoc., 17, 672−697
39. Berenger, F., Kumar, A., Zhang, K. Y. J. and Yamanishi, Y. (2021) Lean-docking: exploiting ligands' predicted docking scores to accelerate molecular docking. J. Chem. Inf. Model., 61, 2341−2352
40. Sadybekov, A. A., Brouillette, R. L., Marin, E., Sadybekov, A. V., Luginina, A., Gusach, A., Mishin, A., Besserer-Offroy, É., Longpré, J. M., Borshchevskiy, V., *et al*. (2020) Structure-based virtual screening of ultra-large library yields potent antagonists for a lipid gpcr. Biomolecules, 10, 1634
41. Soleimany, A., Amini, A., Goldman, S., Rus, D., Bhatia, S. and Coley, C. (2021) Evidential deep learning for guided molecular property prediction and discovery. ACS Cent. Sci., 8, 1356−1367
42. Yang, Y., Yao, K., Repasky, M. P., Leswing, K., Abel, R., Shoichet, B. K. and Jerome, S. V. (2021) Efficient exploration of chemical space with docking and deep learning. J. Chem. Theory Comput., 17, 7106−7119
43. Graff, D. E., Shakhnovich, E. I. and Coley, C. W. (2021) Accelerating high-throughput virtual screening through molecular pool-based active learning. Chem. Sci. (Camb.), 12, 7866−7881
44. Graff, D. E., Aldeghi, M., Morrone, J. A., Jordan, K. E., Pyzer-

Knapp, E. O. and Coley, C. W. (2022) Self-focusing virtual screening with active design space pruning. J. Chem. Inf. Model., 62, 3854−3862

45. Shen, C., Ding, J., Wang, Z., Cao, D., Ding, X. and Hou, T. (2019) From machine learning to deep learning: advances in scoring functions for protein–ligand docking. Wiley Interdiscip. Rev. Comput. Mol. Sci., 10, e1429
46. Li, H., Sze, K. H., Lu, G. and Ballester, P. J. (2020) Machine-learning scoring functions for structure-based drug lead optimization. Wiley Interdiscip. Rev. Comput. Mol. Sci., 10, e1465
47. Yang, J., Shen, C. and Huang, N. (2020) Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. Front. Pharmacol., 11, 69
48. Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J. and Sayle, R. A. (2020) Zinc20—a free ultralarge-scale chemical database for ligand discovery. J. Chem. Inf. Model., 60, 6065−6073
49. Trott, O. and Olson, A. J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem., 31, 455−461
50. Zhang, X., Wong, S. E. and Lightstone, F. C. (2013) Message passing interface and multithreading hybrid for parallel molecular docking of large databases on petascale high performance computing machines. J. Comput. Chem., 34, 915−927
51. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. and Hutchison, G. R. (2011) Open Babel: an open chemical toolbox. J. Cheminform., 3, 33
52. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., *et al*. (2019) Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model., 59, 3370−3388
53. Mysinger, M. M., Carchia, M., Irwin, J. J. and Shoichet, B. K. (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J. Med. Chem., 55, 6582−6594

