HHS Public Access

Author manuscript

J Comput Aided Mol Des. Author manuscript; available in PMC 2020 December 01.

Published in final edited form as:

J Comput Aided Mol Des. 2019 December ; 33(12): 1095–1105. doi:10.1007/s10822-019-00247-3.

Exploring Fragment-Based Target-Specific Ranking Protocol with Machine Learning on Cathepsin S

Yuwei Yang[1], **Jianing Lu**[1], **Chao Yang**[1], **Yingkai Zhang**[1,2,*]

[1]Department of Chemistry, New York University, New York, NY, 10022, United States

[2]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

Abstract

Cathepsin S (CatS), a member of cysteine cathepsin proteases, has been well studied due to its significant role in many pathological processes, including arthritis, cancer and cardiovascular diseases. CatS inhibitors have been included in D3R-GC3 for both docking pose prediction and affinity ranking, and in D3R-GC4 for binding affinity ranking. The difficulties posed by CatS inhibitors in D3R mainly come from three aspects: large size, high flexibility and similar chemical structures. We have participated in GC4; our best submitted model, which employs a similarity-based alignment docking and Vina scoring protocol, yielded Kendall's τ of 0.23 for 459 binders in GC4. In our further explorations with machine learning, by curating a CatS specific training set, adopting a similarity-based constrained docking method as well as an arm-based fragmentation strategy which can describe large inhibitors in a locality-sensitive fashion, our best structure-based ranking protocol can achieve Kendall's τ of 0.52 for all binders in GC4. In this exploration process, we have demonstrated the importance of training data, docking approaches and fragmentation strategies in inhibitor-ranking protocol development with machine learning.

Keywords

Scoring function; virtual screening; fragmentation; docking; machine learning

I. INTRODUCTION

In computer-aided drug design, accurately ranking potential inhibitors for a specific target is critical in virtual screening and inhibitor optimization[1–13]. Typically, a structure-based computational protocol for ranking involves a number of different tasks, including data collection and preparation, receptor structure selection, inhibitor conformation generation, binding pose prediction, and binding affinity ranking[9]. The choice made for each task may influence its subsequent task, and eventually influence the overall performance of the ranking protocol. Besides conventional approaches, machine learning and deep learning algorithms have been widely applied in each stage of ranking protocol development[14–22]. An excellent platform that can facilitate the development and test of such integrated ranking protocols is Drug Design Data Resource (D3R), which aims to promote computer aided drug

*(Y.Z.) Yingkai.zhang@nyu.edu.

discovery (CADD) by providing large and high-quality protein-ligand binding data sets as well as organizing competitions to test CADD workflows regularly.

D3R Grand Challenges (GC) are competitions of pose and affinity or relative affinity prediction focusing on specific targets. In the past four years, D3R has released target-specific databases to evaluate docking and affinity ranking protocols in a blind test manner[23–25]. The availability of these high-quality experimental data greatly benefits CADD community, and simultaneously accelerates the development of new computational methods. Cathepsin S (CatS) is a member of cysteine cathepsin family, and is predominately found in antigen presenting cells, where it degrades MHC class II-associated invariant chain. Malfunction of CatS is related to hyper immune response, cancer, arthritis and etc.[26,27]. CatS has been used as a sub-challenge target in both D3R GC3 [24] and GC4: GC3 has evaluated both pose prediction and affinity ranking methods on CatS system; while in GC4, only affinity ranking is tested with a much larger data set of 459 CatS inhibitors.

We have participated in GC4 and submitted two models for the CatS sub-challenge. Both models used poses generated by the same similarity-based alignment docking protocol but different affinity ranking scoring function: one employed Vina score[28]; the other used a developing version of XGBoost-based scoring function, which incorporated features from a core-fragmentation method and was trained using a general training set covering protein-ligand complexes from multiple targets. However, test results are not satisfactory, and the better one using Vina score has a Kendall's τ of 0.23 for 459 binders in GC4. The difficulty of ranking CatS inhibitors may come from their large size, high flexibility and similar chemical structures: the average molecular weight of the CatS GC4 inhibitors is around 700 g/mol; the average number of rotatable bonds is around 8; all of the CatS GC4 inhibitors share the same tetrahydropyridinepyrazole derivative core, which is a well-established scaffold targeting CatS [29–31]. Meanwhile, considering that CatS has been studied for decades and it has a relatively large amount of binding affinity data which has not been used by us for developing those two submitted models, we have motivated to further explore target-specific ranking protocols with machine learning. In our further development, by curating a CatS specific training set and adopting a similarity-based constrained docking method as well as an arm-based fragmentation strategy which can describe large inhibitors in a locality-sensitive fashion, our best structure-based ranking protocol can achieve Kendall's τ of 0.52 for all binders in the GC4 test set. Our overall exploration process is summarized in Fig. 1 and is described in detail below.

II. METHODS

Our exploration of structure-based ranking protocols mainly focuses on three stages, namely training data collection, binding pose prediction and binding affinity ranking, and different approaches have been tested in each stage. The overview of our exploring process is summarized in Fig. 1.

II. 1. Data Collection:

To explore CatS-specific inhibitor ranking protocols, we have collected two types of data, namely general protein-ligand data and CatS-specific data. The general data set is based on

both PDBbind[32] and CSAR[33,34] data sets, which consist of both complex structures and affinities for a variety of protein targets.

The CatS-specific data set only includes crystal structures and binding affinities related to CatS inhibitors. Co-crystal structures are used to validate binding pose prediction methods, and binding affinities are used to develop affinity ranking methods. Considering that CatS GC4 inhibitors have the same tetrahydropyridinepyrazole derivative core, our data collection focuses on molecules containing this core. Thus, 22 co-crystal structures and 139 binding affinities are collected from GC3 released data[24]. Additional 291 binding affinities are collected from the ChEMBL database[35]. Since most molecules only have binding affinities, the CatS-inhibitor complex structures need to be generated using pose prediction methods. The minimal Tanimoto fingerprint[36] distances, calculated by RDKit[37], between the curated data sets and the CatS GC4 inhibitors are shown in Fig2.

II.2. Binding Pose Prediction

Structure-based ranking protocol depends on protein-ligand complex structures, but the large size and high flexibility of CatS inhibitors bring challenges for pose prediction due to the large conformational space they are associated with. GC3 CatS competition results indicated that the flipped binding mode of tetrahydropyridinepyrazole derivative core adds more difficulties for binding pose prediction (Fig. 3A)[24]. On the other hand, guidance from a similar co-crystal ligand, which includes both selecting a receptor conformation and using a co-crystal ligand as the docking template, can help with binding pose prediction even in those flipped core cases [38–42]. Thus, besides using the general conformer rigid docking protocol, we have explored two similarity-based docking approaches: similarity-based alignment docking and similarity-based constrained docking.

II.2.1 Structure Preparation—Both protein and ligand protonation states are calculated at pH = 5 to be the same as experimental conditions. Protein protonation state is calculated using PDB2PQR[43], and ligand protonation state is estimated using experimental pKa values of similar functional groups collected from SciFinder[44]. For each ligand, conformers are generated using RDKit[37,45] from its SMILES string and minimized using MMFF94 with water as solvent[37,46–51]. Docking pdbqt files for both the protein and the ligand are prepared using AutoDockTools[52].

II.2.2 General Docking Method: Conformer Rigid Docking—1000 conformers are generated for each ligand using RDKit[37,45], and docked onto 3IEJ, which is a crystal structure available before D3R GC3 with tetrahydropyridinepyrazole derivative core, using Vina rigid docking (performed by smina[53], a fork of Autodock Vina[28]). Vina score is used to select final poses. Symmetry-corrected RMSDs between the docked poses and the co-crystal ligand are calculated using ArbAlign[54].

II.2.3 CatS-Specific Docking Methods: Similarity-Based Docking—Two similarity-based docking methods are explored, namely similarity-based alignment docking and similarity-based constrained docking. In both methods, for ligand of interest, its most similar co-crystal ligand is picked as a docking reference and the similarity is measured

using fingerprint Tanimoto coefficient. The receptor of the reference ligand is used as docking receptor.

The major difference between these two methods lies in ligand conformer generation and initial positioning into the receptor binding site. In similarity-based alignment docking, maximum 1000 conformers are randomly generated and minimized, and then aligned to the reference ligand based on their maximum common substructure (MCS). While in similarity-based constrained docking, atoms in MCS are kept at the same positions as those in the reference ligand during conformer generation, and maximum 500 conformers are generated since part of the ligand is fixed. In both methods, only conformers within 100 kcal/mol from the lowest energy conformer are used for the subsequent docking procedure. Local minimization is performed to optimized conformers against the receptor using smina[53]. Since conformers are either aligned or constrained to the core of the reference ligand, their core conformation should be similar. Then we filter docking poses by their occupancies of pockets in the binding site. AlphaSpace[55] is used to detect and match pockets using different CatS structures and estimate pocket ligandability. Pocket analysis of all GC3 CatS co-crystal structures shows that highly ligandable pockets are well occupied by CatS inhibitors as shown in Fig. 3C. Thus, docking poses are ranked by their occupancies of highly ligandable pockets, and only the top 100 poses are kept. Among them, the pose with the most favorable Vina score is used as the final docking pose.

II.3. Binding Affinity Prediction

The last stage in a ranking protocol is binding affinity prediction. Fragmenting molecules can reduce molecular size and flexibility, but current available fragmentation methods[56–58] are not designed for scoring purpose. We have explored three fragmentation methods and used them to compare general and target-specific scoring functions by training on either the general data set or the CatS-specific data set with XGBoost[59,60], one of the most widely used machine learning algorithms.

II.3.1 Scoring Function Parameterized with General Protein-Ligand Data Set and Core-Based Fragmentation—The general protein-ligand data set contains complexes from multiple targets, and we used it to explore a core-based fragmentation approach, which separates ligands into a core and several side chains as illustrated in Fig. 4. The core is defined as the largest ring system in the molecule, and the rest is defined as side chains. The integrity of uncuttable motifs, such as rings, double bonds and triple bonds, is maintained in the fragmentation process. Features related to the entire ligand and the fragments are calculated. We have also tried 3 bridging-water features and 3 ligand stability features to take account of explicit water molecules in the binding site and ligand stability. CSAR[61,33] and part of PDBbind v2016[32] (structures released before 2015) are used in training, and the rest of PDBbind[32] is used as the validation set. The CatS GC3 affinity data set is used to fine-tune the model's performance on CatS system.

II.3.2 Target-Specific Scoring Function—We have explored three fragmentation approaches to develop CatS-specific scoring functions to rank CatS GC4 inhibitors, as illustrated in Fig. 4 and described in detail below.

- **Core-Based Fragmentation** is the same as used for the general protein-ligand data set, and it cuts molecules into a core and side chains. Features of the side chains are summed to get total side-chain features for scoring function development. Core-based fragmentation only depends on ligand chemical structure.
- **Arm-Based Fragmentation** cuts functional groups from the shared tetrahydropyridinepyrazole derivative core. It separates CatS inhibitors into a core and 4 arms, namely the left arm, the right arm, the lower arm, and the core modification. Due to the core-flipping effects (Fig. 3A), the lower arm can be switched with the left arm. The core flipping status is predicted to be the same as inhibitor's reference co-crystal ligand's core status. Since the core structure is the same, only the arm features are used in scoring function development. Arm-based fragmentation depends on ligand chemical structure and core-flipping effect prediction.
- **Pocket-Based Fragmentation** separates a ligand by its occupied pockets. By employing the AlphaSpace[55] pocket matching algorithm, pockets on different crystal structures of CatS can be clustered together by their relative positions on the receptor surface. To reduce the number of fragments, matched pockets within 8 Å are combined to form matched pocket communities. Atoms occupying the same pocket community as well as their associated uncuttable motifs are grouped into one fragment. The rest of atoms are considered as linker atoms. Fragments in the same matched pocket community are considered as matched fragments, which occupy the similar region on the protein surface. Pocket-based fragmentation needs protein-ligand complex structures, which depends on pose prediction accuracy.

In target-specific scoring function development, XGBoost[59,60] algorithm is used for parameterization. The CatS ChEMBL affinity data set is used as the training set and the CatS GC3 affinity data set is used as the validation set. Besides the Vina score, 10 features used by our previously developed Δ_{vina} RF_{20}[21], are employed to describe the contribution of each fragment. Among them, five are fragment descriptors and the other five are interaction descriptors. Additionally, each fragment's occupied AlphaSpace[55] volume is used as a feature for model training. All features are summarized in Table 1. Before testing on the CatS GC4 data set, to make the maximum use of all available data, the best model is retrained using the combination of the CatS ChEMBL and the CatS GC3 affinity data sets, and then applied to the CatS GC4 data set to get the final performance.

II.4 Performance Evaluation

RMSE refers to the root-mean-squared error between the predicted pIC50 and the experimentally measured pIC50, and it is used to evaluate the affinity prediction accuracy. Kendall's $\boldsymbol{\tau}$ and Spearman's $\boldsymbol{\rho}$ are used to evaluate the ranking performance.

$$\text{Kendall's } \boldsymbol{\tau} = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{n(n-1)/2}$$

$$\text{Spearman's } \boldsymbol{\rho} = \frac{cov(r_P, r_O)}{\sigma_P \sigma_O}$$

where $cov(r_P, r_O)$ is the covariance of the predicted ranking and the experimental ranking and $\boldsymbol{\sigma}$ is the standard deviation of the rank variables.

III. RESULTS AND DISCUSSION

III.1. Binding Pose prediction

Binding pose prediction is an essential step in structure-based ranking protocols, but the large size and high flexibility of CatS inhibitors together with their core-flipping effects (Fig. 3A) bring difficulties in the binding pose prediction. We have explored similarity-based alignment docking and similarity-based constrained docking to tackle these challenges. Both methods are based on the hypothesis that similar ligands should bind similarly on a specific receptor, and thus they use the compound's most similar co-crystal ligand as the reference in the docking process. Binding mode analysis of CatS co-crystal structures indicates that ligand similarity can distinguish different core-flipping conformations as shown in Fig. 3B. Therefore, adopting a similar co-crystal ligand as docking reference helps with core-flipping effects prediction.

As shown in Fig. 5, similarity-based docking methods significantly improve the docking accuracy comparing to a conventional conformer rigid docking. Similarity-based constrained docking has the best performance, which has a medium RMSD of 2.2 Å and an average RMSD of 2.3 Å. On the other hand, conformer rigid docking has a median RMSD of 6.4 Å if only the top Vina score poses are used. Extending the limit to top 10 Vina score poses and selecting the lowest RMSD pose for docking performance evaluation ends up with a median RMSD of 4.2 Å, which is still not close to the near native poses (RMSD ≤ 2 Å). Meanwhile, similarity-based alignment docking has a median RMSD of 3.0 Å and an average RMSD of 3.8 Å, but has an obvious outlier with RMSD of 12.3 Å.

In D3R GC4 CatS sub-challenge, our submitted models employ Vina score and the general core-based fragmentation score using similarity-based alignment docking poses. When applying these general scoring methods on similarity-based constrained docking poses, their affinity ranking performances are improved as shown in Table 2. Although previous GC3 results suggest that the docking accuracy may not be correlated with the affinity ranking[24], our comparison here demonstrates that affinity ranking can be improved by enhancing docking accuracy. Therefore, our further target-specific scoring function exploration uses similarity-based constrained docking poses.

III.2. Binding Affinity Ranking:

In order to further improve the binding affinity ranking, we have explored three fragmentation approaches, and the results are summarized in Table 3. The best model is the CatS arm-based scoring function trained on the CatS ChEMBL and the CatS GC3 data sets with Kendall's $\boldsymbol{\tau}$ of 0.52 and Spearman's $\boldsymbol{r}$ of 0.71. This is significantly better than Vina,

which has Kendall's tau of 0.29 calculated with similarity-based constrained docking poses (Table 2). Furthermore, comparing to our best submitted model using Vina score with similarity-based alignment docking poses, this new best model has improved both ranking and scoring performances, as shown in Fig. 6.

The improved performance may come from two aspects: the training data selection and the fragmentation idea. General scoring functions, which are developed using general data sets (Vina is derived using PDBbind data), do not perform as well as CatS-specific scoring functions on the CatS GC4 test set, except in the case of the CatS non-fragment scoring functions trained only using the CatS ChEMBL affinity data. After adding more related data, the CatS GC3 affinity data, into the training set, the performances of all CatS-specific scoring functions get further boosted to surpass general scoring functions. Additionally, we have tried training the core-based scoring function using either the general data or the target-specific data. Training with the CatS-specific data achieves better ranking performance on the test set. These results indicate that having more related data in the training set generally can improve the affinity ranking performance.

The fragment-based CatS-specific scoring functions always outperform the non-fragment-based one on the test set, although the latter uses the same descriptors but on the ligand level. Even the core-based scoring function, which captures the least locality information, performs better than the non-fragment-based scoring function. The results indicate that locality-based fragmentation methods contribute to the binding affinity ranking. Within fragment-based scoring functions, the arm-based and the pocket-based scoring functions achieve better performance than the core-based scoring function, which could result from the relative more locality information they captured. Interestingly, the arm-based scoring function always performs better than the pocket-based one on the test set, regardless of how much CatS data presents in the training set. Although, pocket-based fragmentation captures more locality information, it also has stronger dependency on the docking accuracy. When docking poses are inaccurate, fragments can be assigned to the wrong pockets, which would negatively affect the training process by grouping fragments occupying different localities together. The arm-based fragmentation, with less dependency on the pose prediction and relatively more locality information, outperforms other fragment-based scoring functions.

Even though the target-specific training set with arm-based fragmentation shows the best performance, they are critically dependent on available data. One common pre-requisite for target-specific ranking protocol development is to have sufficient experimental data. For the arm-based fragmentation, it also requires a shared core. Many target-specific inhibitors are developed from a lead compound, which can be considered as a conserved core, and arm-based fragmentation should be applicable to these cases. Nevertheless, the arm-based fragmentation defines each arm by their attaching positions on the core structure, if the core changes, definitions of the arms are not consistent, which can generate misleading locality information. Both core-based and pocket-based fragmentation methods can be applied to inhibitors without a shared core. The core-based fragmentation can be generalizable to general data set, but it only distinguishes core binding sites from the rest. The pocket-based fragmentation can be applied to a specific target or a group of similar targets but has higher dependency on the accuracy of docking poses, which limits its performance to some extent.

CONCLUSION

When exploring CatS-specific ranking protocols, a common trend we have seen is that tailored data with tailored docking and scoring methods usually leads to better performance. The reason why general docking methods do not have good performance on CatS system could come from the large size and high flexibility of CatS ligands. In this work, we have explored the similarity-based docking approaches for the pose prediction task and the target-specific fragment-based scoring functions for the affinity ranking task. Our results show that the similarity-based constrained docking achieves the best docking accuracy with a medium RMSD of 2.2 Å and an average RMSD of 2.3 Å on the CatS GC3 data set, and better docking accuracy often can lead to better ranking performance. Our explored fragmentation approaches that break ligands into pieces can reduce the size and flexibility of each fragment, while account for local structural differences regardless of high similarities among CatS ligands. Combining with the target-specific data set, the similarity-based constrained docking method and the XGBoost[59,60] algorithm, all fragment-based scoring functions can outperform the general and non-fragment-based scoring functions. Among all models that we explored, the arm-based scoring function performs the best with Kendall's τ of 0.52. Comparing to our best submitted model, whose Kendall's τ is only 0.23, we have made significant improvement by selecting related training set, improving the pose prediction accuracy and employing fragmentation ideas in target-specific scoring function development. Such a strategy may also be helpful for other protein targets with relatively large and flexible ligands.

ACKNOWLEDGMENTS

We would like to acknowledge the support by NIH (R35-GM127040, R01GM073943 and R01GM120736) and computing resources provided by NYU-ITS.

References

1. Lavecchia A, Di Giovanni C (2013) Virtual Screening Strategies in Drug Discovery: A Critical Review. Curr Med Chem 20 (23):2839–2860. [PubMed: 23651302]
2. Schneider G (2010) Virtual screening: an endless staircase? Nat Rev Drug Discov 9 (4):273–276. [PubMed: 20357802]
3. Ashtawy HM, Mahapatra NR (2012) A Comparative Assessment of Ranking Accuracies of Conventional and Machine-Learning-Based Scoring Functions for Protein-Ligand Binding Affinity Prediction. IEEE/ACM Trans Comput Biol Bioinform 9 (5):1301–1313. [PubMed: 22411892]
4. Kim R, Skolnick J (2008) Assessment of programs for ligand binding affinity prediction. J Comput Chem 29 (8):1316–1331. [PubMed: 18172838]
5. Stjernschantz E, Oostenbrink C (2010) Improved Ligand-Protein Binding Affinity Predictions Using Multiple Binding Modes. Biophys J 98 (11):2682–2691. [PubMed: 20513413]
6. Su MY, Yang QF, Du Y, Feng GQ, Liu ZH, Li Y, Wang RX (2019) Comparative Assessment of Scoring Functions: The CASF-2016 Update. J Chem Inf Model 59 (2):895–913. [PubMed: 30481020]
7. Li Y, Liu ZH, Li J, Han L, Liu J, Zhao ZX, Wang RX (2014) Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. J Chem Inf Model 54 (6): 1700–1716. [PubMed: 24716849]
8. Li Y, Han L, Liu ZH, Wang RX (2014) Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. J Chem Inf Model 54 (6):1717–1736. [PubMed: 24708446]

9. Cheng TJ, Li QL, Zhou ZG, Wang YL, Bryant SH (2012) Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. AAPS J 14 (1):133–141. [PubMed: 22281989]
10. Cheng TJ, Li X, Li Y, Liu ZH, Wang RX (2009) Comparative Assessment of Scoring Functions on a Diverse Test Set. J Chem Inf Model 49 (4):1079–1093. [PubMed: 19358517]
11. Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM (2013) Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0-A Public Library of Challenging Docking Benchmark Sets. J Chem Inf Model 53 (6):1447–1462. [PubMed: 23705874]
12. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. J Med Chem 55 (14):6582–6594. [PubMed: 22716043]
13. Rohrer SG, Baumann K (2009) Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. J Chem Inf Model 49 (2):169–184. [PubMed: 19434821]
14. Amini A, Shrimpton PJ, Muggleton SH, Sternberg MJE (2007) A general approach for developing system-specific functions to score protein-ligand docked complexes using support vector inductive logic programming. Proteins 69 (4):823–831. [PubMed: 17910057]
15. Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics 26 (9):1169–1175. [PubMed: 20236947]
16. Durrant JD, McCammon JA (2011) NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function. J Chem Inf Model 51 (11):2897–2903. [PubMed: 22017367]
17. Kinnings SL, Liu NN, Tonge PJ, Jackson RM, Xie L, Bourne PE (2011) A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing. J Chem Inf Model 51 (2):408–419. [PubMed: 21291174]
18. Li LW, Khanna M, Jo IH, Wang F, Ashpole NM, Hudmon A, Meroueh SO (2011) Target-Specific Support Vector Machine Scoring in Structure-Based Virtual Screening: Computational Validation, On Vitro Testing in Kinases, and Effects on Lung Cancer Cell Proliferation. J Chem Inf Model 51 (4):755–759. [PubMed: 21438548]
19. Zilian D, Sotriffer CA (2013) SFCscore(RF): A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. J Chem Inf Model 53 (8):1923–1933. [PubMed: 23705795]
20. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR (2017) Protein-Ligand Scoring with Convolutional Neural Networks. J Chem Inf Model 57 (4):942–957. [PubMed: 28368587]
21. Wang C, Zhang YK (2017) Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions using Random Forest. J Comput Chem 38 (3):169–177. [PubMed: 27859414]
22. Jimenez J, Skalic M, Martinez-Rosell G, De Fabritiis G (2018) K-DEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J Chem Inf Model 58 (2): 287–296. [PubMed: 29309725]
23. Gathiaka S, Liu S, Chiu M, Yang HW, Stuckey JA, Kang YN, Delproposto J, Kubish G, Dunbar JB, Carlson HA, Burley SK, Walters WP, Amaro RE, Feher VA, Gilson MK (2016) D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. J Comput Aided Mol Des 30 (9):651–668. [PubMed: 27696240]
24. Gaieb Z, Parks CD, Chiu M, Yang HW, Shao CH, Walters WP, Lambert MH, Nevins N, Bembenek SD, Ameriks MK, Mirzadegan T, Burley SK, Amaro RE, Gilson MK (2019) D3R Grand Challenge 3: blind prediction of protein-ligand poses and affinity rankings. J Comput Aided Mol Des 33 (1):1–18. [PubMed: 30632055]
25. Gaieb Z, Liu S, Gathiaka S, Chiu M, Yang HW, Shao CH, Feher VA, Walters WP, Kuhn B, Rudolph MG, Burley SK, Gilson MK, Amaro RE (2018) D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. J Comput Aided Mol Des 32 (1):1–20. [PubMed: 29204945]
26. Turk V, Stoka V, Vasiljeva O, Renko M, Sun T, Turk B, Turk D (2012) Cysteine cathepsins: From structure, function and regulation to new frontiers. Biochim Biophys Acta 1824 (1):68–88. [PubMed: 22024571]

27. Wilkinson RDA, Williams R, Scott CJ, Burden RE (2015) Cathepsin S: therapeutic, diagnostic, and prognostic potential. Biol Chem 396 (8):867–882. [PubMed: 25872877]
28. Trott O, Olson AJ (2010) Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. J Comput Chem 31 (2):455–461. [PubMed: 19499576]
29. Ameriks MK, Axe FU, Bembenek SD, Edwards JP, Gu Y, Karlsson L, Randal M, Sun SQ, Thurmond RL, Zhu J (2009) Pyrazole-based cathepsin S inhibitors with arylalkynes as P1 binding elements. Bioorg Med Chem Lett 19 (21):6131–6134. [PubMed: 19773165]
30. Thurmond RL, Sun SQ, Sehon CA, Baker SM, Cai H, Gu Y, Jiang W, Riley JP, Williams KN, Edwards JP, Karlsson L (2004) Identification of a potent and selective noncovalent cathepsin S inhibitor. J Pharmacol Exp Ther 308 (1):268–276. [PubMed: 14566006]
31. Wiener DK, Lee-Dutra A, Bembenek S, Nguyen S, Thurmond RL, Sun S, Karlsson L, Grice CA, Jones TK, Edwards JP (2010) Thioether acetamides as P3 binding elements for tetrahydropyrido-pyrazole cathepsin S inhibitors. Bioorg Med Chem Lett 20 (7):2379–2382. [PubMed: 20188543]
32. Liu ZH, Su MY, Han L, Liu J, Yang QF, Li Y, Wang RX (2017) Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. Acc Chem Res 50 (2):302–309. [PubMed: 28182403]
33. Dunbar JB, Smith RD, Yang CY, Ung PMU, Lexa KW, Khazanov NA, Stuckey JA, Wang SM, Carlson HA (2011) CSAR Benchmark Exercise of 2010: Selection of the Protein-Ligand Complexes (vol 51, pg 2036, 2011). J Chem Inf Model 51 (9):2146–2146.
34. Huang SY, Zou XQ (2011) Scoring and Lessons Learned with the CSAR Benchmark Using an Improved Iterative Knowledge-Based Scoring Function. J Chem Inf Model 51 (9):2097–2106. [PubMed: 21830787]
35. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) The ChEMBL database in 2017. Nucleic Acids Res 45 (D1):D945–D954. [PubMed: 27899562]
36. Rogers D, Hahn M (2010) Extended-Connectivity Fingerprints. J Chem Inf Model 50 (5):742–754. [PubMed: 20426451]
37. RDKit: Open-source cheminformatics; http://www.rdkit.org.
38. Koukos PI, Xue LC, Bonvin A (2019) Protein-ligand pose and affinity prediction: Lessons from D3R Grand Challenge 3. J Comput Aided Mol Des 33 (1):83–91. [PubMed: 30128928]
39. Kumar A, Zhang KYJ (2019) Shape similarity guided pose prediction: lessons from D3R Grand Challenge 3. J Comput Aided Mol Des 33 (1):47–59. [PubMed: 30084081]
40. Lam PCH, Abagyan R, Totrov M (2019) Hybrid receptor structure/ligand-based docking and activity prediction in ICM: development and evaluation in D3R Grand Challenge 3. J Comput Aided Mol Des 33 (1):35–46. [PubMed: 30094533]
41. Nguyen DD, Cang ZX, Wu KD, Wang ML, Cao Y, Wei GW (2019) Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. J Comput Aided Mol Des 33 (1):71–82. [PubMed: 30116918]
42. Ignatov M, Liu C, Alekseenko A, Sun ZYZ, Padhorny D, Kotelnikov S, Kazennov A, Grebenkin I, Kholodov Y, Kolosvari I, Perez A, Dill K, Kozakov D (2019) Monte Carlo on the manifold and MD refinement for binding pose prediction of protein-ligand complexes: 2017 D3R Grand Challenge. J Comput Aided Mol Des 33 (1):119–127. [PubMed: 30421350]
43. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res 35:W522–W525. [PubMed: 17488841]
44. SciFinder; https://scifinder.cas.org/scifinder/.
45. Riniker S, Landrum GA (2015) Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. J Chem Inf Model 55 (12):2562–2574. [PubMed: 26575315]
46. Halgren TA (1996) Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 17 (5–6):490–519.
47. Halgren TA (1996) Merck molecular force field .2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem 17 (5–6):520–552.

48. Halgren TA (1996) Merck molecular force field .3. Molecular geometries and vibrational frequencies for MMFF94. J Comput Chem 17 (5–6):553–586.
49. Halgren TA (1996) Merck molecular force field .5. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J Comput Chem 17 (5–6):616–641.
50. Halgren TA, Nachbar RB (1996) Merck molecular force field .4. Conformational energies and geometries for MMFF94. J Comput Chem 17 (5–6):587–615.
51. Tosco P, Stiefl N, Landrum G (2014) Bringing the MMFF force field to the RDKit: implementation and validation. J Cheminformatics 6.
52. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. J Comput Chem 30 (16):2785–2791. [PubMed: 19399780]
53. Koes DR, Baumgartner MP, Camacho CJ (2013) Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. J Chem Inf Model 53 (8):1893–1904. [PubMed: 23379370]
54. Temelso B, Mabey JM, Kubota T, Appiah-Padi N, Shields GC (2017) ArbAlign: A Tool for Optimal Alignment of Arbitrarily Ordered Isomers Using the Kuhn-Munkres Algorithm. J Chem Inf Model 57 (5):1045–1054. [PubMed: 28398732]
55. Rooklin D, Wang C, Katigbak J, Arora PS, Zhang YK (2015) Alpha Space: Fragment-Centric Topographical Mapping To Target Protein-Protein Interaction Interfaces. J Chem Inf Model 55 (8): 1585–1599. [PubMed: 26225450]
56. Liu TR, Naderi M, Alvin C, Mukhopadhyay S, Brylinski M (2017) Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag. J Chem Inf Model 57 (4):627–631. [PubMed: 28346786]
57. Murray CW, Rees DC (2009) The rise of fragment-based drug discovery. Nat Chem 1 (3):187–192. [PubMed: 21378847]
58. Lewell XQ, Judd DB, Watson SP, Hann MM (1998) RECAP - Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. J Chem Inf Comput Sci 38 (3):511–522. [PubMed: 9611787]
59. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining:785–794.
60. XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754.
61. Dunbar JB, Smith RD, Damm-Ganamet KL, Ahmed A, Esposito EX, Delproposto J, Chinnaswamy K, Kang YN, Kubish G, Gestwicki JE, Stuckey JA, Carlson HA (2013) CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. J Chem Inf Model 53 (8):1842–1852. [PubMed: 23617227]
62. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semiempirical free energy force field with charge-based desolvation. J Comput Chem 28 (6):1145–1152. [PubMed: 17274016]

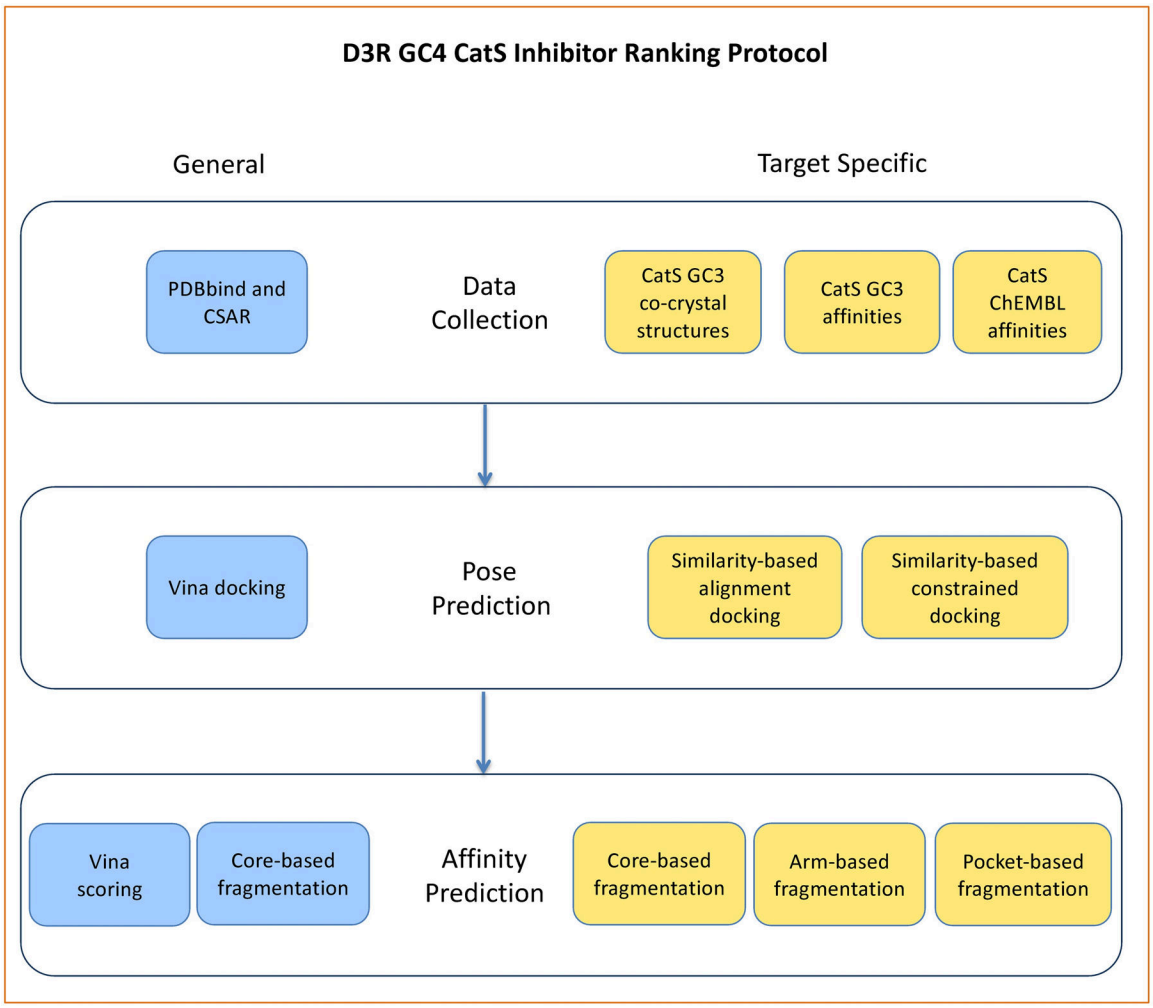

Figure 1.
Ranking protocol development for GC4 CatS sub-challenge.

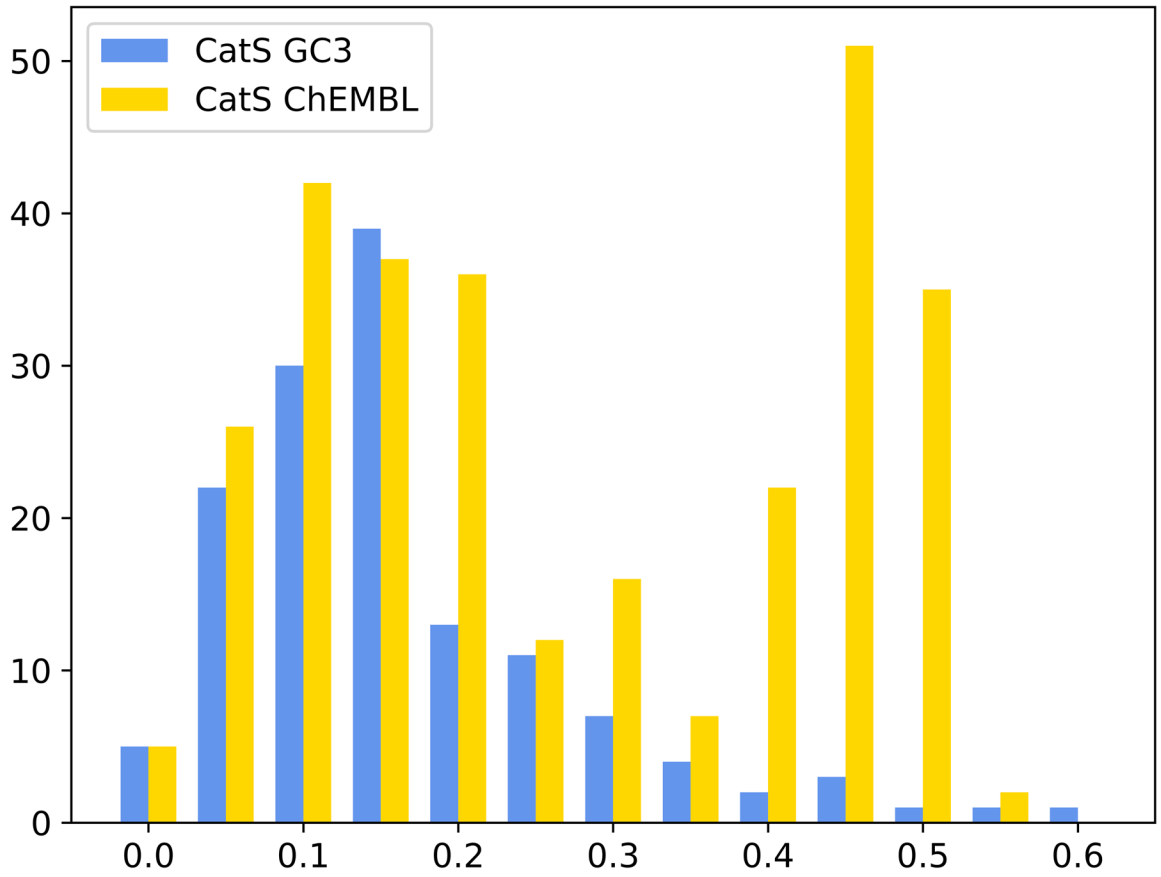

Figure 2.
The minimal Tanimoto fingerprint distances between the CatS GC4 compounds and the collected compounds. CatS GC3 and CatS ChEMBL data sets maintain high similarity to the blind test set, the CatS GC4 data set, with maximum Tanimoto distance to be 0.6.

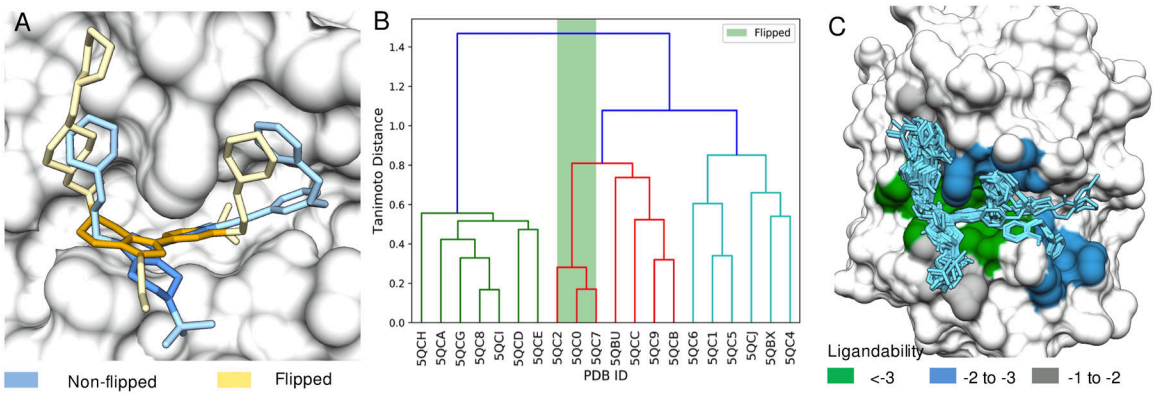

Figure 3.

Illustration of similarity-based docking. A. CatS tetrahydropyridinepyrazole derivative core has two binding modes, flipped and non-flipped. B. Hierarchical clustering of CatS co-crystal ligands shows that similar ligands have similar binding modes. Thus, using the most similar co-crystal ligand as the docking reference can help predict the binding mode. C. All CatS co-crystal ligands bind to the similar region on the receptor surface, which can be predicted by AlphaSpace[55] pocket ligandability. Similarity-based docking first brings poses to the binding site by employing the docking reference, and then uses pocket occupancies of highly ligandable pockets to initially filter binding poses.

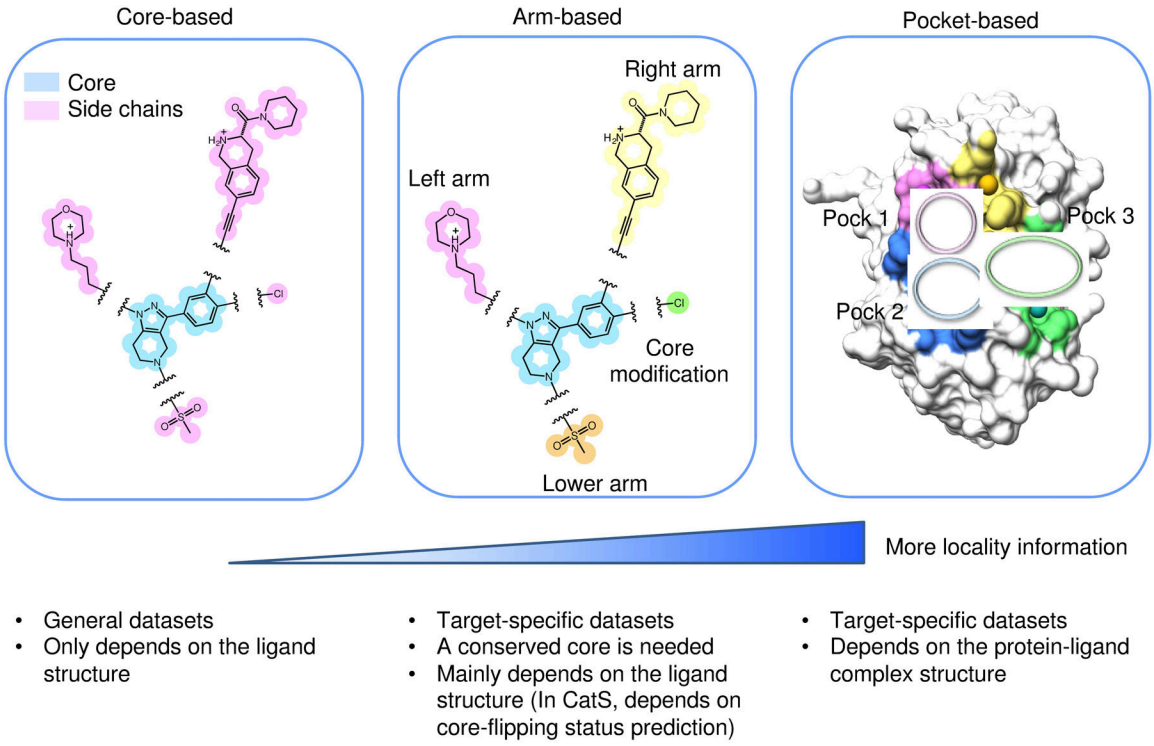

Figure 4.

Illustration of three fragmentation methods: 1. core-based fragmentation, in which each ligand is divided into one core and several side chains; 2. arm-based fragmentation, in which each distinct arm is truncated from the shared core structure and labeled by its attachment point; 3. pocket-based fragmentation, in which fragments are assigned and labeled based on their occupied pockets.

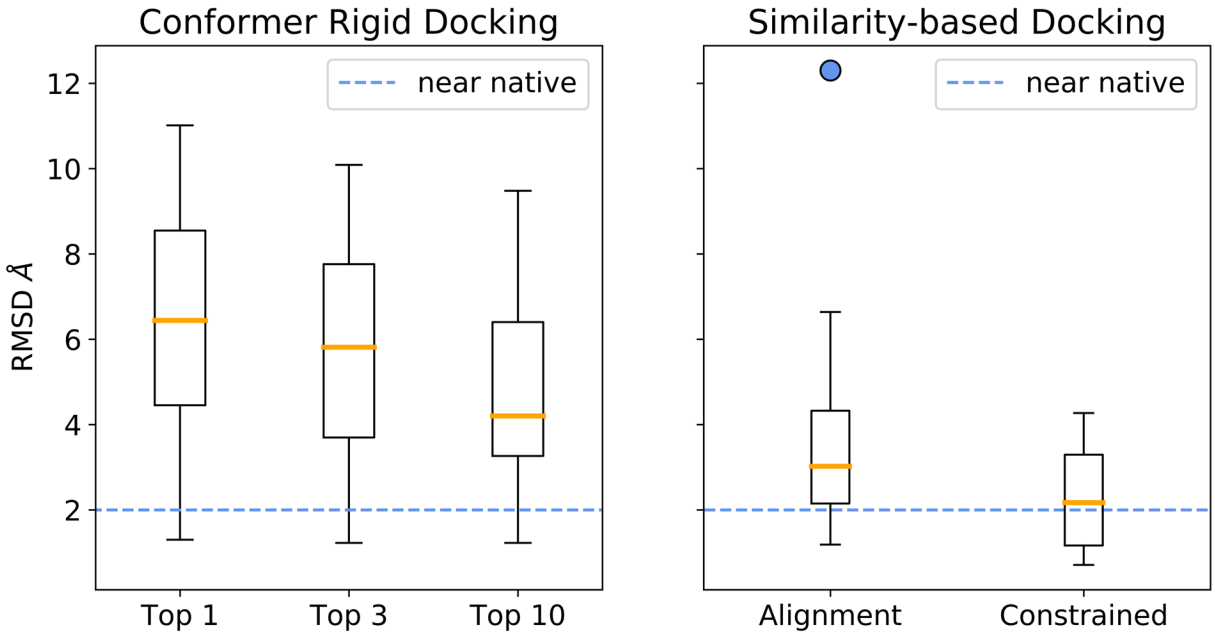

Figure 5.

Docking performance. Two similarity-based docking methods show better performances than the conformer rigid docking. Similarity-based constrained docking performs better than similarity-based alignment docking for having smaller average and median RMSDs and no obvious outliers.

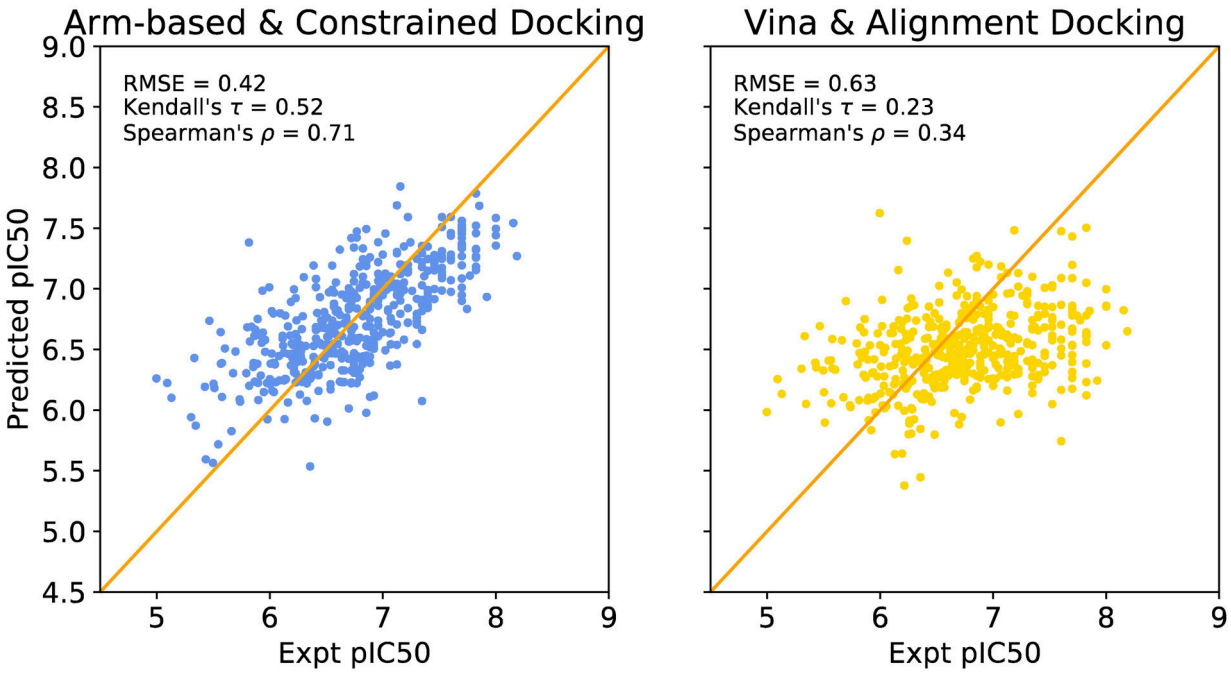

Figure 6.

Scoring performance of the best newly developed model (Arm-based & Constrained docking) and best submitted model (Vina & Alignment Docking).

Table 1.

Features Used in Target-Specific Fragmentation-Based Scoring Function

No.	Feature Description	Feature Type
1	Vina score	Ligand
2	Occupied AlphaSpace volume	Fragment
	Autodock Vina Interaction Terms	
3	$\text{Non\_hydrophobic}\left(a_1,a_2,d\right)=\begin{cases}0, & a_1 \text{ or } a_2 \text{ is hydrophobic}\\ 1, & d_{\text{diff}}\left(a_1,a_2\right)<0.5\\ 0, & d_{\text{diff}}\left(a_1,a_2\right)\geq 1.5\\ 1.5-d_{\text{diff}}\left(a_1,a_2\right), & \text{otherwise}\end{cases}$	Fragment
4	$\text{Non\_hydrophobic}\left(a_1,a_2,d\right)=\begin{cases}0, & a_1 \text{ or } a_2 \text{ do not form hydrogen bond}\\ 1, & d_{\text{diff}}\left(a_1,a_2\right)<-0.7\\ 0, & d_{\text{diff}}\left(a_1,a_2\right)\geq 0.4\\ \frac{d_{\text{diff}}\left(a_1,a_2\right)-0.4}{-1.1}, & \text{otherwise}\end{cases}$	Fragment
5	$\text{Solvation}\left(a_1,a_2,d\right)=\left[\left(\text{ASP}a_1+QASP\times\left\lvert q_{a_1}\right\rvert\right)V_{a_2}+\left(\text{ASP}_{a_2}+QASP\times\left\lvert q_{a_2}\right\rvert\right)V_{a_1}\right]e^{-\left(\frac{d}{7.2}\right)^2}$	Fragment
6–7	$\text{Electrostatic}\left(a_1,a_2,d\right)=\frac{q_{a_1}\times q_{a_2}}{d^x},\qquad x=1 \text{ or } 2$	Fragment
	Autodock Vina Fragment Description Terms	
8	Number of heavy atoms	Fragment
9	Number of hydrophobic atoms	Fragment
10	Number of torsion	Fragment
11	Number of rotors	Fragment
12	Ligand length	Fragment

Note: d is the distance between two atoms, a_1 and a_2. d_{diff} is the surface distance calculated by $d_{diff} = d - R(a_1) - R(a_2)$, where $R(a_1)$ and $R(a_2)$ are the van der Waals radius of atoms a_1 and a_2 [28]. q is the atomic charge and V is the atomic volume. ASP and $QASP$ refer to the atomic solvation parameter and the charge-based solvation parameter, respectively[62].

Table 2.

General Scoring Function Performance Using Similarity-Based Alignment Docking and Constrained Docking Poses. It should be noted that if using ligand-only molecular weight for the affinity prediction, it has Kendall's τ of 0.26 and Spearsman's ρ of 0.37 for GC4.

	Similarity-based Alignment Docking			Similarity-based Constrained Docking		
	GC4 Performance			GC4 Performance		
	RMSE	Kendall	Spearman	RMSE	Kendall	Spearman
Vina	0.63*	0.23*	0.34*	0.67	0.29	0.42
General Core-based	0.72*	0.21*	0.30*	0.66	0.25	0.37

*Performance of submitted models

Table3.

CatS-Specific Scoring Function Performance Using Similarity-Based Constrained Docking Poses. It should be noted that if using ligand-only molecular weight for the affinity prediction, it has Kendall's τ of 0.26 and Spearsman's ρ of 0.37 for the test set (CatS GC4).

Training Set	**CatS ChEMBL**						**CatS ChEMBL and CatS GC3**		
Evaluation	**Validataion Set CatS GC3**			**Test Set CatS GC4**			**Test Set CatS GC4**		
	RMSE	**Kendall**	**Spearman**	**RMSE**	**Kendall**	**Spearman**	**RMSE**	**Kendall**	**Spearman**
CatS Core-based	0.66	0.25	0.36	0.56	0.30	0.43	0.50	0.41	0.58
CatS Arm-based	0.65	0.34	0.48	0.50	0.42	0.59	**0.42**	**0.52**	**0.71**
CatS Pocket-based	0.66	0.35	0.49	0.52	0.37	0.52	0.48	0.43	0.61
CatS Non-Fragment	0.66	0.26	0.37	0.59	0.24	0.35	0.53	0.35	0.49