JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

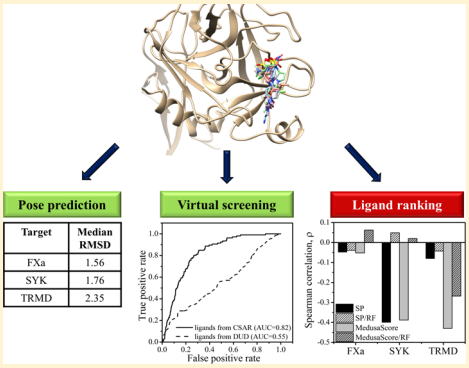
Docking and Scoring with Target-Specific Pose Classifier Succeeds in Native-Like Pose Identification But Not Binding Affinity Prediction in the CSAR 2014 Benchmark Exercise

Regina Politi,[†,‡] Marino Convertino,[‡] Konstantin Popov,[‡] Nikolay V. Dokholyan,*[,‡] and Alexander Tropsha*[,†]

[†]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, and [‡]Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States

Supporting Information

ABSTRACT: The CSAR 2014 exercise provided an important benchmark for testing current approaches for pose identification and ligand ranking using three X-ray characterized proteins: Factor Xa (FXa), Spleen Tyrosine Kinase (SYK), and tRNA Methyltransferase (TRMD). In Phase 1 of the exercise, we employed Glide and MedusaDock docking software, both individually and in combination, with the special target-specific pose classifier trained to discriminate native-like from decoy poses. All approaches succeeded in the accurate detection of native and native-like poses. We then used Glide SP and MedusaScore scoring functions individually and in combination with the pose-scoring approach to predict relative binding affinities of the congeneric series of ligands in Phase 2 of the exercise. Similar to other participants in the CSAR 2014 exercise, we found that our models showed modest prediction accuracy. Quantitative structure−activity relationship (QSAR) models developed for the FXa ligands using available bioactivity data from ChEMBL showed relatively low prediction accuracy for the CSAR 2014 ligands of the same target. Interestingly, QSAR models built with CSAR data only yielded Spearman correlation coefficients as high as $\rho = 0.69$ for FXa and $\rho = 0.79$ for SYK based on 5-fold cross-validation. Virtual screening of the DUD library using the FXa structure was successful in discriminating between active compounds and decoys in spite of poor ranking accuracy of the underlying scoring functions. Our results suggest that two of the three common tasks associated with molecular docking, i.e., native-like pose identification and virtual screening, but not binding affinity prediction, could be accomplished successfully for the CSAR 2014 challenge data set.



1. INTRODUCTION

Accurate prediction of binding affinity remains a challenging problem in drug design. When the three-dimensional structure of the target is available, binding affinity may be estimated by molecular docking, which includes two components: (i) generation of plausible poses of ligands within the binding pocket of the target and (ii) evaluation of those poses using one of many available scoring functions. Modern sampling algorithms often succeed in generating native like ligand poses, but scoring functions typically fail both to recognize correct binding poses and accurately predict ligand binding affinities.[1−3] The 2014 CSAR exercise provided the scientific community with yet another opportunity to evaluate and benchmark the reliability of various computational approaches for predicting protein−ligand interactions. The objectives of this latest challenge were to (i) identify the near-native pose within a set of docking decoys and (ii) dock and score congeneric series of ligands, starting from an X-ray characterized target protein structure and a list of SMILES for ligands. Prediction accuracy was evaluated by the Spearman correlation coefficient between predicted and actual ligand binding affinity.

Previously, we participated in the 2013 CSAR exercise[4] where we rank ordered a series of ten ligands according to their predicted binding affinity toward an engineered digoxigenin-binding protein,[5] and the predicted order was similar to the experimental one with the Spearman correlation coefficient of 0.75. The current study provided an opportunity to test our approach further for post processing of poses generated by conventional molecular docking approaches using larger set of ligands and three diverse protein systems. Protein targets included in the 2014 challenge were well studied, with a large amount of data on ligand binding affinity available in the ChEMBL database.[6] This availability afforded a special opportunity to employ ligand-based approaches such as QSAR modeling of previously reported binders for ranking

Special Issue: Community Structure Activity Resource (CSAR)

Received: December 25, 2015
Published: April 6, 2016

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

CSAR ligands and comparing structure based vs. ligand based methods for their prediction accuracy. It is important to underline that as in all previous cases, the CSAR organizers did not put any restrictions on the use of external publicly available data, methods, and software.[7] In fact, the participants were even encouraged to use as many sources of potentially useful information as possible.

The current exercise was based on the data donated by GlaxoSmithKline (GSK). Three sets of crystal structures of target proteins and their ligands were provided: Factor Xa (FXa), Spleen Tyrosine Kinase (SYK), and tRNA Methyltransferase (TRMD). For Phase 1, the challenge was to identify native-like poses for 22 ligands (3 for FXa, 5 for SYK, and 14 for TRMD) among the set of decoy poses. In Phase 2, we aimed to assess both the correct binding poses and relative ranking by predicted binding affinity of 163 FXa, 276 SYK, and 31 TRMD ligands to their respective protein targets. A unique consensus prediction approach that employed two different types of methods was used: (i) molecular docking that predicts the binding poses of CSAR ligands and ranks them according to their docking scores; (ii) cheminformatics-based classification models built with chemical descriptors of the protein−ligand interface[4,8] that can discriminate native-like poses from decoys based on descriptor values. Two docking/scoring methods, Glide[9−11] and MedusaDock,[12,13] were used and their results were compared. In addition, we have built QSAR models using ligand binding affinity data available in the ChEMBL database[6,14]and used these models to rank CSAR ligands. Finally, we have also built QSAR models using CSAR 2014 binding data only.

We showed that our approaches were successful in Phase 1, i.e., we have identified native-like poses as best scoring for all ligands across all Phase 1 data sets. However, similar to the absolute majority of all CSAR 2014 participants, both our docking models and ChEMBL-based QSAR models did not achieve statistically significant correlation between experimental IC_{50} and predicted affinities in Phase 2 of the 2014 Challenge, i.e., accurate ranking of ligands by their predicted binding affinity. QSAR models built with CSAR data only using Dragon descriptors reached Spearman correlation, ρ = 0.69 and R^2 = 0.52 for FXa and ρ = 0.79 and R^2 = 0.66 for SYK based on 5-fold cross-validation. We speculate that the low prediction accuracy in Phase 2 obtained by all participants of the challenge could be due to several confounding issues. In addition to limitations imposed by scoring functions itself[15] those may include missing domains in crystallographic structures (as was found for FXa); it is also possible that binding affinities provided by the organizers could have been measured in a protocol that was different from those used to generate data reported in ChEMBL; and finally, the results could be influenced by the existence of activity cliffs[16] in the data set. However, despite the low accuracy of binding affinity prediction, virtual screening using the FXa structure successfully discriminated binders from decoys. Overall, we have succeeded in the identification of correct binding pose and virtual screening but not in binding affinity prediction, which remains a major challenge for docking methods.

2. METHODS

2.1. Data Sets. In Phase 1 the organizers of the CSAR 2014 benchmark exercise provided the participants with 22 crystal structures from GSK: 3 for FXa, 5 for SYK, and 14 for TRMD. For each protein−ligand pair, protein coordinates were provided in a ".mol2" file, and the 200 ligand poses were provided in a "multi.mol2" file. Each crystal structure was set up for docking and scoring using the MOE 2011.10 (force field: MMFF94x with AM1-BCC charges for the ligands). Each ligand was removed from each protein−ligand complex, and 500 docked poses were generated with DOCK (version 6.5). A subset of 200 poses was chosen for each test set provided to the participants. One of the poses was within 1 Å RMSD of the actual crystal structure pose, and the other 199 were chosen systematically, guided by the diversity analysis in the MOE descriptor space.

The same three protein targets were provided in Phase 2 with larger lists of compounds (31 for TRMD, 276 for SYK, and 163 for FXa). Compounds for FXa were given in three separate SMILES files, which corresponded to respective data sets from GSK. Participants were asked to dock and score each set of small molecules as a separate data set labeled as set1 (FXA_set101), set2 (FXA_set398), and set3 (FXA_set401). There was some overlap in small molecules across all sets. We analyzed each of the sets separately and as one large set. For brevity, only the analysis of the large unified set is discussed here, but the conclusions are the same for each individual subset. To unify subsets into one large set the experimental binding affinities of overlapping compounds, provided by the organizers after the completion of the exercise, were averaged.

2.2. Molecular Docking. Two separate docking/scoring methods are reported here for Phase 1 and Phase 2 of this exercise: Glide[9−11] and MedusaDock.[12,13]

Phase 1 of the Exercise. The crystallographic structures of FXa, SYK, and TRMD provided by the organizers were preprocessed using the Protein Preparation wizard in the Schrodinger Suite (version 9.7, http://www.schrodinger.com/). All poses provided for small-molecule structures in complex with respective target proteins were scored in situ with either the Glide program using standard docking precision protocol (Glide SP) or MedusaScore,[17] a physical force field based scoring function. Explicit hydrogen atoms were added and ionizable compounds were converted to their most probable charged forms at pH 7.0 ± 2.0 using the LigPrep software available in Glide.

Phase 2 of the Exercise. The following crystallographic structures were used for docking (PDB codes are given): gtc401 was used for FXa set1 and gtc101 was used for FXa set2 and set3. SYK ligands were docked into each one of the crystallographic structures provided by the organizers at the end of Phase 1; TRMD_458 provided by the organizers at the end of Phase 1 was used for all the ligands of TRMD. The reason for using these specific crystallographic structures is described in section 2.4. These structures were preprocessed using the Protein Preparation wizard in the Schrodinger Suite (version 9.7, http://www.schrodinger.com/). The ligands were prepared as in Phase 1. When using Glide, conformational sampling was performed for all ligands using ConfGen.[18] The binding region for FXa and SYK was defined by a 10 Å × 10 Å × 10 Å grid box and for TRMD by 30 Å × 30 Å × 30 Å grid box centered on the active site of each target protein. A scaling factor of 0.8 was applied to the van der Waals radii. Default settings were used for all the remaining parameters. The top-100 poses were generated for each ligand and ligands were sorted according to the docking scores (termed SP) of the top pose.

The same crystallographic structures were used in docking calculations with MedusaDock,[12,13] an in-house developed tool

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

that simultaneously models the flexibility of both ligands and proteins. We ranked and identified binding poses of the investigated compounds as estimated by MedusaScore,[17] or by combining the binding energy values with a hierarchical cluster analysis of the geometry of binding of top-ranked ligand conformations. We performed 250 independent docking calculations for every compound docked in FXa, SYK, and TRMD. We collected the top-scored conformations (i.e., docking poses with Z-score lower than −2), the number of which varies from 2 to 442 depending on the provided molecules in the three data sets. The ensemble of docking solutions was clustered according to the root-mean-square deviation (RMSD) computed over the ligand's heavy atoms. The optimal number of highly populated clusters was identified by applying the average linkage method and the Kelley penalty index,[19] which minimizes the number of clusters and the spread of internal values in each cluster. The clustering level with the lowest Kelley penalty represents a condition where the clusters are highly populated and concurrently maintain the smallest internal spread of RMSD values (e.g., RMSD spread of the most populated clusters varies on average from 1.0 to 3.6 Å depending on the provided molecules in the three data sets). For each target, we adopted seven different metrics, summarized in Table 1, to choose the representative binding energy of ligands bound to the proteins: (i) LowEner; (ii) MPCcent; (iii) MPCLowEner; (iv) MPCAvEner; (v) MPCFreeEner; (vi) LowAvEnerClus; (vii) LowFreeEnerClus. The average MedusaScore energy in each cluster was calculated as

Table 1. Metrics Adopted to Choose the Representative Binding Energy of Ligands Bound to Each Target Protein[a]

code	metric
LowEner	absolute lowest binding energy
MPCcent	binding energy of the centroid of the most populated cluster of docking poses
MPCLowEner	lowest binding energy pose within the ensemble of solutions in the most populated cluster of docking poses
MPCAvEner	average binding energy of the most populated cluster of docking poses.
MPCFreeEner	effective free energy of the most populated cluster of docking poses
LowAvEnerClus	lowest average binding energy among the identified clusters of docking poses
LowFreeEnerClus	lowest effective free energy among the identified clusters of docking poses

[a]Binding energy is estimated by MedusaScore;[17] effective free energy is calculated from eq 3.

$$\langle E\rangle_c = \frac{\sum_i E_i}{n_c} \tag{1}$$

where, n_c is the cluster size, and E_c is the MedusaScore of pose i within a cluster.

For computing the effective free energy we consider each cluster as a canonical ensemble of states (i.e., docking poses) described by the partition function $Z = \sum_i e^{-\beta E_i}$.

The probability of a pose with energy E_i in each cluster is given by

$$P_i = \frac{1}{Z} e^{-\beta E_i} \tag{2}$$

Hence, the effective free energy of each cluster can be written as

$$F_c = \frac{\sum_i E_i e^{-\beta E_i}}{\sum_i e^{-\beta E_i}} - k_B T \ln(n_c) \tag{3}$$

where, β is the reciprocal of $k_B T$, ∼0.6 kcal/mol, which corresponds to thermal fluctuation energy at room temperature (298 K).

The first term of the equation represents the average energy $\langle E\rangle = \sum_i P_i E_i$ of a cluster, while the second term denotes the combinatorial entropic contribution depending on the number of poses in the cluster, n_c.

2.3. PL/MCT-Tess Descriptors. Previously, one of our groups developed the PL/MCT-Tess approach to characterize protein−ligand interfaces.[20] Briefly, when applied to different poses generated by molecular docking, Delaunay tessellation[21] partitions the protein−ligand interface into an aggregate of space-filling, irregular tetrahedra, where both protein and ligand atoms are vertices.[20] Each Delaunay quadruplet is characterized by its unique four-atom composition, which defines the descriptor type (certainly, the same four-body compositions may occur in different or even, the same protein−ligand interfaces). Furthermore, for each quadruplet we calculate the sum of MCT values (vide infra) of the composing atom-vertices and this sum represents the actual descriptor value. Overall, each pose is uniquely characterized by a vector of 554 MCT descriptor values.

PL/MCT-Tess descriptors employ pairwise atomic potentials for the protein−ligand complexes (PL) based on maximal charge transfer (MCT).[22] The MCT characterizes the maximal electron flow between the donor and acceptor atoms at the protein−ligand interface. It is derived from the conceptual DFT,[22,23] which provides a theoretical basis for calculating the PL/MCT-Tess descriptors. The values of PL/MCT-Tess descriptors are calculated from the following equation:

$$\text{PL/MCT-Tess}_m = \sum_{k=1}^{n} \sum_{p}^{1-3} \sum_{l}^{1-3} (\text{MCT}_p \text{MCT}_l / d_{pl})_k \tag{4}$$

where PL/MCT-Tess$_m$ is the potential of the mth tetrahedron type defined by its four-atom composition (i.e., individual descriptor type); n is the number of occurrences of this tetrahedron type in a given protein−ligand complex; p is the index of protein vertex-atoms, l is the index of ligand vertex-atoms, and d_{pl} is the distance between a pair of protein and ligand atoms found in the same Delaunay tetrahedron.

2.4. Generation of the Target Specific Modeling Set of Native-Like and Decoy Poses. *Phase I.* In our recent studies, we have developed an approach to pose scoring that could classify poses in a distribution obtained with docking as native-like or non-native using a classifier developed for a known protein−ligand complex.[4] The classifier is developed by generating a sample of poses using any docking software, assigning each pose to a native-like or non-native class based on their RMSD from the known native structure, and using a machine learning approach such as random forest (RF) and PL/MCT-Tess descriptors of each pose to build a model. The code to build a classifier using RF is implemented within the freely available ChemBench server (https://chembench.mml.unc.edu/). To develop such a classifier here, the Schrodinger suite was used to generate an ensemble of native-like and decoy poses for a single ligand of each target with known X-ray characterized structure available from the Protein Data Bank

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

(PDB).[24] From our previous experience[4] we learned that crystallographic structures whose cognate ligand shares the highest similarity with the ligands in the external data set should be used for developing models with the highest external predictive accuracy. Thus, Tanimoto similarity coefficient (Tc) was calculated between all available crystallographic structures from the PDB and ligands provided for Phase 1 to identify a known protein for generating the training set. Tc was calculated using ISIDA fragments[25] implemented in ISIDA program (http://infochim.u-strasbg.fr/spip.php?rubrique49). Two ligands provided for FXa, gtc398, and gtc401 shared high similarity (Tc = 0.93 for gtc398 and 0.91 for gtc401) with the ligand from a crystallographic structure with the PDB code 2J4I. However, the third ligand provided, gtc101, shared high similarity (Tc = 0.95) with another crystallographic structure with the PDB code 2WYG. It should be noted that while two separate models were created to identify native-like poses for these ligands in Phase 1, we discuss here only application of the model created based on crystallographic structure with PDB code 2J4I (see Results and Discussion for details).

Ligands provided for SYK showed relatively low similarity with the ligands in PDB. The highest similarity (Tc = 0.6) was found with the ligand in the crystallographic structure with PDB code 4FZ6. For ligands of TRMD crystallographic structure with the PDB code 4MCC was used to build the model even though the Tc was very low (~0.5).

Phase 2. Most of the ligands provided for FXa as part of set1 were either similar to the ligand from crystallographic structure with PDB code 2CJI or to the ligand from Phase 1, complex coded as gtc401. Both ligands were docked into the crystallographic structure, gtc401, provided by the organizers at the end of Phase 1. Ligands of two other sets provided for FXa were either similar to the ligand from crystallographic structure with PDB code 2UWP or to the ligand from Phase 1, complex labeled as gtc101. These ligands were docked into the crystallographic structure provided by the organizers at the end of Phase 1 for gtc101. All Phase 2 ligands of SYK were highly similar to the ligands provided in Phase 1. Thus, all ligands of Phase 1 were docked into their respective structure provided by the organizers to build the model. Ligands provided for TRMD in Phase 2 had high similarity with several ligands from Phase 1 (gtc451, gtc452, gtc453, gtc457, gtc460, gtc464, gtc465). To create the model, these ligands from Phase 1 were docked into the crystallographic structure with the highest resolution released by the CSAR organizers for this target after the completion of Phase 1, TRMD_458.

All ligands were docked into the corresponding PDB structures described above using Glide SP with default parameters. Following the previously developed methodology,[4,8,20] only poses with RMSD smaller than 4 Å from the native pose were used for modeling. Subsequent to this initial pose filtering, we defined a RMSD threshold of 2 Å to discriminate native-like poses from decoys: poses with RMSD smaller than 2 Å were considered as native-like (class 1) whereas poses with RMSD larger than 2 Å (but smaller than 4 Å) were considered as decoys (class 0).

2.5. Pose Classifier. In this study, we used a quantitative structure−activity relationship (QSAR) like approach to build models that evaluate the likelihood of a ligand pose to be either native-like or a decoy. Thus, the problem of separating native-like poses from decoys can be approached by developing binary classification models, in which each pose is characterized by multiple descriptors of the protein−ligand interface (i.e., PL/MCT-Tess descriptors in this study).

We followed the model building and validation workflow and other guidelines our group published previously.[26,27] The workflow includes three steps:[27,28] (1) data curation/preparation (selection of poses and descriptors), (2) model building, and (3) model validation. Here, we applied a 5-fold cross-validation procedure. The full set of poses was randomly split into five groups of roughly equal size. Four different groups were systematically used as modeling sets (80% of the full set) and the remaining group was employed as an external validation set (remaining 20%). This procedure was repeated five times so that each group formed the external validation set once and that cumulative statistics could be computed for the whole set of poses. Random Forest (RF) implemented in R (https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm) was used for model development. Binary classification models were built using modeling set poses only; it is important to emphasize that poses in the external set were never taken into account neither to build nor select models, this condition being critical to ensure the rigor of the external validation procedure.[29] Additionally, each modeling set was split into multiple internal training and test sets; models were built using poses belonging to each training set and applied to internal test set poses for assessing their native-likeness.

The best training set models were identified and selected according to their prediction performances on the respective test sets as measured by specificity, sensitivity, balanced accuracy (BA), and the Spearman's rank correlation coefficient. Then, all selected models were applied jointly to the external set poses. Again, this procedure was repeated five times to ensure that every pose from the full set is present once (and only once) in the external test set. In addition, Y-randomization (randomization of classes) was performed for each selected model in order to ensure that there was no chance correlation. It is expected that models obtained for the training set with randomized responses should have significantly lower predictivity as compared to models built using the original training set. If this condition is not satisfied, those models should be discarded. In this study, Y-randomization was applied to all training/test data set divisions. All validated models were stored and used in an ensemble for predicting the native-likeness of any pose of any ligand binding to the same protein.

2.6. Combining Docking and Pose Classifier Scores for Pose and Ligand Ranking. For any given ligand−protein pose, the ensemble of selected RF models outputs a continuous consensus score (RF score) ranging from 0 (pose predicted to be decoy by all models) to 1 (pose predicted to be native-like by all models). When there is a disagreement between those individual RF models, the consensus RF score can thus take any value between 0 and 1. When computed for a set of poses, RF scores can be used to rank those poses based on their decreasing RF-evaluated likelihood of being native-like. This treatment is particularly interesting since it offers an additional approach for quantitative ranking of poses on top of scoring using a function from the actual molecular docking program.

For each CSAR ligand, the whole set of poses (either provided by the organizers or generated with either Glide or MedusaDock) were ranked according to the native-like score predicted by the RF classification models (RF score): the closer the consensus score is to 1, the more likely the pose is native-like. The top pose based on the predicted RF score was chosen

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

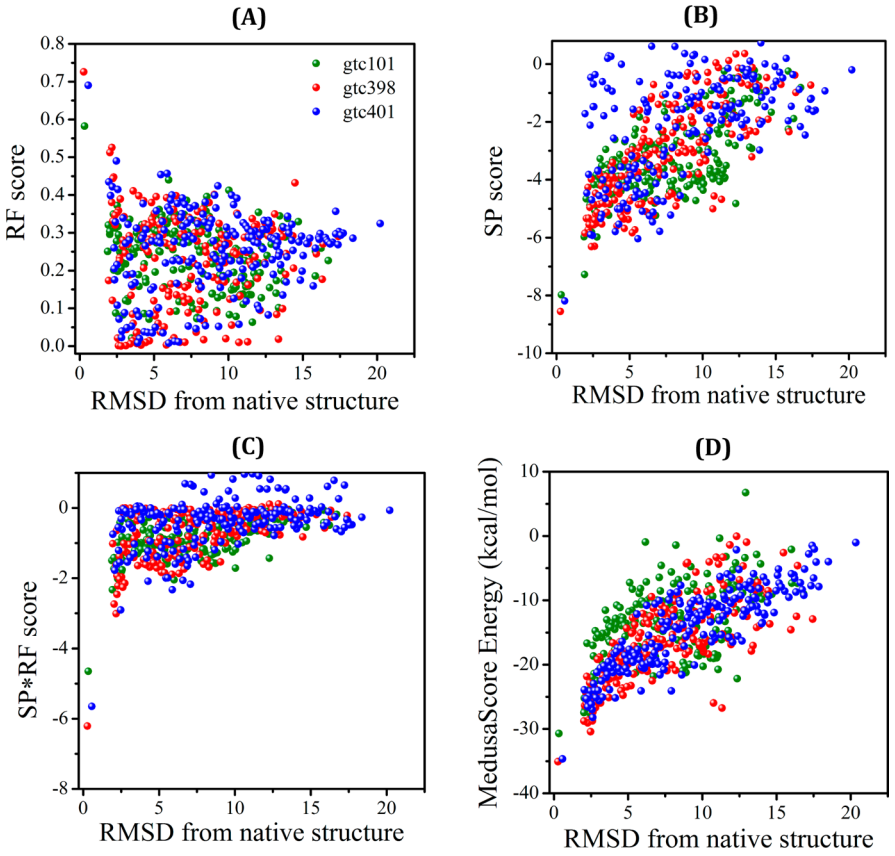


Figure 1. Scores for 200 poses for each of the three FXa ligands as a function of RMSD: (A) RF scores; (B) Glide SP docking scores; (C) product of Glide SP and RF scores; (D) MedusaScore energies.[17]

for each ligand. The poses submitted in Phase 1 were either ranked by RF score, or by docking score using standard docking precision (Glide SP). As it has been shown in the literature that consensus scoring improves the results of docking (e.g., Charifson et.al.,[30]) we have also explored the product of Glide SP and RF scores. In Phase 2, in addition to RF score all poses were characterized by either SP Glide scores or MedusaScore depending on the method used for docking. The top ranked poses of each ligand based on RF scores were further selected and used for ranking the CSAR ligands based on their corresponding docking score. These combined Glide SP and RF (SP/RF), or MedusaScore and RF (MedusaScore/RF) scores were used to generate new ligand ranking which was different from the original ranking based on docking scores only. The rank received using this workflow was ultimately compared to experimental rank obtained from measured binding affinities. To match the metric used by the organizers, ranking performance was evaluated using the Spearman correlation coefficient expressing the ranking accuracy of ligands by comparing the ligands' rank orders based on model's predicted potency (pIC50) with the actual experimental rank provided by the CSAR organizers. In addition, square of the correlation coefficient,[31] R^2, between the observed and predicted values was used to assess models accuracy.

3. RESULTS AND DISCUSSION

3.1. Developing Target-Specific RF Classification Models. Following the approach implemented in our previous work,[4] we built target specific pose classification models to filter out poses predicted to be decoys. The RF pose classification models developed for three targets, FXa, SYK, and TRMD of the 2014 CSAR exercise separately were used to identify and eliminate non-native poses for all CSAR compounds of Phase 1 and 2. It was previously noticed that to achieve reliable prediction performance of the classification models, ligands of high similarity to those in the new data set should be used to build the model.[4] Thus, despite having the same targets in both Phases 1 and 2, different crystallographic structures were used to build the classification models (section 2.2 in Methods for details). Figure S1 in the Supporting Information shows that RF models developed for each target based on MCT-Tess descriptors led to very high prediction performances in discriminating native-like poses from decoys as evaluated by 5-fold cross-validation.

The RF pose classification models developed for Phase 1 were applied to identify the native-like pose among 200 poses provided by the CSAR organizers. Similarly, models developed for the use in Phase 2 were applied as part of the protocol to rank ligands provided for Phase 2 of the exercise. As explained in the Methods section, the closer the consensus RF score predicted by the ensemble of pose-classifying models to 1, the higher the probability that this pose is indeed native-like.

We submitted the best poses and the predicted ranks for all the ligands for three targets of interest; FXa, SYK, and TRMD, computed using SP/RF approach (see Methods for details) to the CSAR organizers for the evaluation of this blind prediction for novel compounds. The combination of MedusaDock with RF is analyzed here but was not employed as part of our submission for the Challenge.

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

3.2. Scoring of CSAR Compounds to Identify the Native Pose—Phase 1 of the CSAR Exercise. All pregenerated CSAR decoys of each ligand were scored in their corresponding protein using Random Forest (RF), Glide SP, product of Glide SP with RF (SP*RF), and MedusaDock. Figures 1, S2, and S3 show different types of scores used for protein−ligand complexes plotted as a function of RMSD from native pose released by CSAR organizers after the completion of the challenge. Overall, all scoring functions used for ligands of FXa and SYK successfully identified the native-like poses among 200 decoys provided for each ligand. Not all respective poses of TRMD ligands were correctly recognized with the RF score as the native like pose (Figure S3). This can be a result of very low similarity (Tc $\leq$ 0.5) found between the ligands in the CSAR 2014 set and the ligand used to build a classification model. Figure S4 in the Supporting Information provided by the organizers shows that only 7 out of 52 submitted methods were able to correctly identify native-like pose for all 22 TRMD's complexes.

For the results herein, crystallographic structure with PDB code 2J4I was used to build the classification model for all three ligands of FXa. Tanimoto similarity coefficients (Tc) between the ligand in the crystallographic structure and CSAR ligands were 0.93 for gtc398, 0.91 for gtc401, and 0.35 for gtc101. However, at the time of submission, the ligand from crystallographic structure with PDB code 2WYG (Tc = 0.95) was used to build the model for the ligand gtc101. For this particular ligand we could identify the native like pose among the three top scored when the other two highly scored poses deviated from the native like pose by 1.89 and 1.93 Å (data not shown). The limitation of the approach is that it classifies poses as native like when the RMSD from the native is below 2 Å. Since similar descriptor profiles are created for very similar poses with RMSD below or very close to 2 Å it will be hard to discriminate between poses with such a small deviation from the native one.

From the information provided later by the organizers diversity analysis guided the choice of this particular 200 poses out of 500 docked poses generated with DOCK (version 6.5). Data shown in Figure 1 indicates that among 200 poses provided by the organizers for each ligand, very few (if any) had RMSD within 2 Å from the respective native poses. Indeed, the absolute majority of poses were distributed fairly evenly within RMSD range of 2−20 Å. One may argue that this pose prefiltering, and especially, exclusion of poses with relatively low deviation from native, may have facilitated achieving success in this exercise.

3.3. Docking and Scoring Congeneric Series of Ligands—Phase 2 of the CSAR Exercise. All CSAR ligands were docked using both Glide SP and MedusaDock and ranked according to their respective docking score values. Several clustering methods were used to rank compound poses generated by MedusaDock. Figure S5 in the Supporting Information summarizes prediction accuracy as a result of different clustering methods. However, none of the methods showed significant improvement in the prediction performance as compared to others. Thus, we further used only one of the clustering methods which was based on the MPCcent (section 2.2) following previously shown success of this metric.[12,32] In addition, following the same idea and protocols developed in our earlier studies,[4,20] we filtered out docking poses predicted to be decoys by the RF/MCT-tess classification model described in Methods. The classification models developed for each of the three targets led to very high prediction performance (BA equals 1 for FXa and SYK and 0.99 for TRMD; cf. Figure S1b in the Supporting Information) in discriminating native-like poses from decoys as evaluated by 5-fold cross-validation (see the Methods section). Each RF model was applied to identify and eliminate non-native poses generated for the respective targets by either one of docking approaches used.

After the submission of predictions was closed, CSAR organizers released the experimental data on the binding affinities of ligands as well as binding poses for a few ligands of each target (5 for FXa, 9 for SYK and 31 for TRMD). The results suggested that most groups, including our own, achieved relatively high accuracy in predicting the correct ligand binding poses as shown in Figure S6a (the median RMSD values are 1.76, 1.56, and 2.35 for SYK, FXa, and TRMD, respectively), which is considered as a prerequisite for accurate prediction of binding affinities. However, no statistically significant correlation was found between the experimental and predicted affinities by all methods used (Figures S6b and S7). To examine if the latter result could be a consequence of inaccurate docking, we manually created binding poses for a set of 24 compounds from Phase 2 with Tanimoto similarity $\geq$0.95 to the reference ligands used in Phase 1 (gtc101, gtc398, and gtc401) using the following procedure. Starting from the bound conformations of these ligands, we manually introduced specific chemical moieties present in the highly similar analogs, and performed a minimization of each ligand-target complex followed by the estimation of the binding energy of each ligand bound conformation. Also, this alternative procedure resulted in very low correlation between MedusaScore and the provided experimental activity. The Spearman correlation was as low as ρ= −0.36 for these 24 compounds. Figure 2 shows

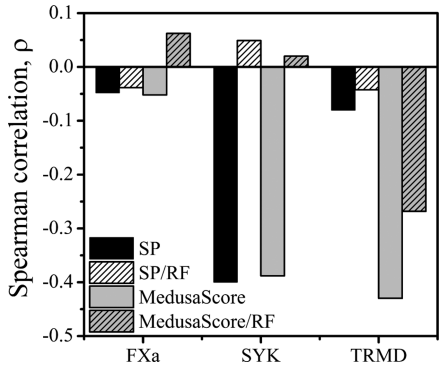


Figure 2. Prediction accuracy, evaluated by the Spearman correlation coefficient, obtained for all the ligands of Phase 2. Results of four approaches are shown for each of three targets: Glide SP, SP/RF, MedusaScore, and MedusaScore/RF.

that our approach, successfully applied to the CSAR 2013 data set,[4] only slightly improved the prediction accuracy of Glide SP and MedusaScore scoring functions in the current exercise. The largest improvement was found for SYK where the Spearman correlation has changed from −0.39 to 0.049 once the RF classification was applied to the poses docked and scored by Glide SP and from −0.34 to 0.02 once the classification model was applied to the poses docked and scored by MedusaDock and MedusaScore, respectively. We should keep in mind, however, that in the previous studies we have also seen an improvement toward better accuracy. In this case, the initial

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

accuracy was already pretty low implying that either poses were far from native or experimental affinities were measured poorly; in any case, one should not expect that a pose scoring function could substantially improve a poor correlation.

In the CSAR 2013 study, we concluded that pose classification models that were specifically and rigorously trained for a given target using its cognate ligand can enable a substantial improvement in the overall hit recovery rate.[4] However, we also showed that the approach is successful for filtering out the ligand poses when conventional scoring functions scored well poses that were predicted to be non-native-like according to the classification model. When there are many ligands inaccurately scored by the conventional scoring function this approach will not be expected to improve the accuracy.

3.4. Analysis of Low Correlation. We believe that low correlation obtained can be a result of multiple parameters that make ranking and affinity prediction of congeneric series of ligands challenging. A recent study compared the accuracy of 8 docking programs and 16 scoring functions[33] for predicting experimental binding affinity against six unrelated protein targets including FXa. The authors note that the nature of the active site of the proteins, the choice of scoring functions, and the set of ligands used for comparisons all affected the performance in scoring and ranking compounds,[33] and none of the scoring function or programs used was equally effective for all targets tested. Based on the summary provided by CSAR organizers for all groups that participated in the competition, the FXa data set was found to be the most challenging. Out of 28 scoring methods used for this particular data set none was able to rank ligands with $\rho \geq 0.5$. It should be noted that the active site of the FXa is largely solvent-exposed and the complementarity appears poor. In addition, all FXa crystal structures are missing several domains, and the effect on ligand binding is unclear. It is possible that some of these domains provide parts of the pocket for the inhibitors or affect the electrostatics or conformational behavior of the catalytic domain. All of the above make the structure-based approach challenging.

3.5. Improving the Binding Affinity Prediction Accuracy Using Ligand Based Method? We have previously used ligand-based QSAR modeling to complement structure-based approach to boost the prediction accuracy.[34] In order to possibly benefit from using ligand-based approaches in assessing binding affinities, we extracted all known ligands for each of the CSAR targets from the ChEMBL database.[6] The application of ligand-based approaches for FXa is challenging as a large fraction of ligands for this target are known to be activity cliffs.[16] We wanted to refine the available data set from ChEMBL by retaining only compounds that were more likely to be measured in the same or very similar experimental conditions. An additional criterion was to keep only those ligands that were reported in publications with 20 and more compounds tested. In this way, data set of 1683 compounds with their corresponding pIC_{50} values ranging from 2.9 to 10.7 were used for QSAR modeling for FXa. The model accuracy based on 5-fold cross-validation (Table 2) was relatively high ($R^2 = 0.67$ and $\rho = 0.81$). However, when making predictions for all FXa ligands provided by the CSAR organizers, the QSAR models developed with ChEMBL ligands afforded very low prediction performance ($R^2 = 0.007$ and $\rho = -0.005$; cf. Table 2).

Table 2. Statistical Characteristics for QSAR Models of FXa Ligands and for Model-Based Predictions

	modeling data sets[a]			external data set of CSAR ligands	
	ChEMBL	refined ChEMBL (subset of compounds most similar to CSAR ligands)	CSAR ligands	predicted using ChEMBL model	predicted using refined ChEMBL model
R^2	0.67	0.64	0.52	0.007	0.004
Spearman, ρ	0.81	0.75	0.69	−0.005	0.077

[a]Assessed by 5-fold cross-validation.

Keeping in mind the notion of high proportion of activity cliffs in this data set[16] we decided to refine the data set used for QSAR modeling even more by selecting, from the ChEMBL database, 100 compounds with the highest pairwise Tanimoto similarity coefficient (Tc) to molecules in the respective CSAR data set. Tc between this and CSAR data set revealed that even highly similar compounds with Tc ≥ 0.90 may differ in activity by 2 orders of magnitude (Figure 3). Moreover, despite overall

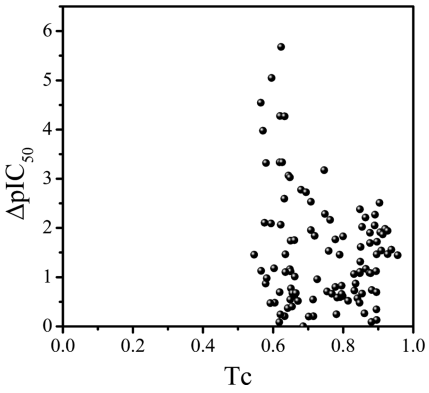


Figure 3. Difference in reported activity computed between FXa ligands provided by the CSAR organizers and those in the ChEMBL data set as a function of pairwise Tanimoto similarity coefficients (Tc).

high concordance ($R^2 = 0.95$) between same FXa ligands measured in two, independent and slightly different assays, some of these ligands provided by the organizers had different activity. For example, GTC000006A in set 1 had $pIC_{50} = 9.11$ and in set 2 $pIC_{50} = 8.22$; similarly, GTC000044A in set 1 had $pIC_{50} = 7.35$ and in set 2 $pIC_{50} = 6.39$.

Despite relatively high modeling accuracy shown in Table 2 ($R^2 = 0.64$ and $\rho = 0.75$) for the refined ChEMBL data set, the prediction performance of models for the CSAR compounds was still as low as $\rho = 0.077$ and $R^2 = 0.004$. Interestingly, once the CSAR data set itself was used to build the model, its accuracy based on 5-fold cross-validation reached $\rho = 0.69$ and $R^2 = 0.52$ for FXa and $\rho = 0.79$ and $R^2 = 0.66$ for SYK.

The results of QSAR modeling of previously existing data, the low prediction performance of QSAR models built with ChEMBL data for the CSAR data set and the low accuracy of QSAR models built for the CSAR data set itself may indicate that the activity measurements for compounds provided by the CSAR organizers may have been generated in a way different from those reported in ChEMBL making models developed for either data set inapplicable to another data set.

3.6. Structure Based Virtual Screening to Test Whether Known Active Compounds Could Be Identi-

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

fied. We used decoys and ligands compiled for the FXa in the Directory of Useful Decoys (DUD)[35] to test the ability of the Glide software to identify active compounds provided by the organizers within a large chemical library. All compounds (106 CSAR unique ligands, 146 DUD ligands and 5745 presumed decoys listed in DUD for FXa) were docked to the protein conformation that had the highest resolution (1.61 Å) among all the structures provided by the CSAR organizers, i.e., gtc101. The AUC was calculated to establish the discriminatory power of structure-based virtual screening of FXa ligands provided by the organizers and compared to the AUC values calculated for DUD ligands of the same target protein. Figure 4 shows that we

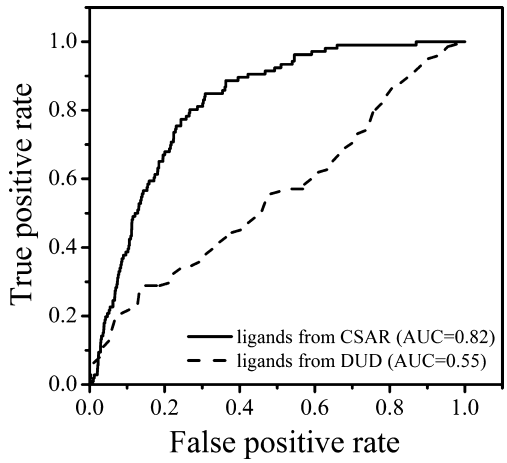


Figure 4. Receiver operating characteristic (ROC) plot for recovering FXa ligands by virtual screening of the DUD benchmarking database. The solid line corresponds to FXa ligands provided by the CSAR organizers, and the dashed line corresponds to FXa ligands from the DUD database.

could successfully differentiate between active compounds and presumed binding decoys (taken from DUD) to separate active vs inactive molecules achieving AUC of 0.82 as opposed to AUC of 0.55 for FXa ligands in DUD. Virtual screening study identified several compounds among presumed decoys that potentially could have an anticoagulation activity and should be further tested experimentally, with the ZINC-ID's of ZINC00590094, ZINC03993392, ZINC01485557, and ZINC03825744. The SP scores for these compounds were −11.68, −11.32, −11.08, and −11.01 whereas the best SP score of known active compounds, for GTC000081A, was −10.06. We then examined a reasonable concern that the results of virtual screening may not be meaningful because we used DUD decoys for DUD actives that could be structurally different from CSAR actives. To address this concern, we have employed the Automated Decoy generation method from the DUD enhanced (DUD-E) available online (http://decoys.docking.org) to construct decoys for CSAR ligands. We found that the results of this additional virtual screening exercise were indistinguishable from those generated with DUD decoys (Figure S8 in the Supporting Information). Thus, we have established that despite the inability of the underlying scoring functions to accurately rank compounds based on their experimental binding affinity we could successfully discriminate between active compounds and presumed decoys in virtual screening.

4. CONCLUSIONS

The most recent CSAR challenge dealt with scoring a set of ligand poses to identify the native-like pose among those provided by the organizers (Phase 1) and submit binding affinity predictions for ligands of three protein targets (Phase 2). This study provided an important opportunity for further testing of our hybrid docking/pose classification/rescoring workflow for ligand ranking. In this study, we employed Glide docking and MedusaDock software in combination with a pose-scoring approach. Each of these methods independently allowed identifying best pose (Phase 1) and consensus allowed slightly improved results. Results obtained with MedusaDock were compared to a similar approach performed independently that relies on docking to an ensemble of targets.[36] We found that simpler single-target docking affords almost similar results, however in some cases slight improvement was shown for docking to an ensemble of structures.[36] It is a subject of further investigation to understand in what cases docking to an ensemble provides significant improvement.

In addition, QSAR models were developed with publically available experimental data extracted from the ChEMBL database. Despite significant improvement achieved by our hybrid method previously[4] in this particular CSAR benchmark, we noticed only marginal improvement in the prediction accuracy using this approach. Moreover, based on the results of all the participating groups provided by the organizers none of the reported methods was effective for all targets in predicting experimental activities. This fact raises a concern regarding quality of experimental data released at the end of the challenge. In an effort to explore the source of poor outcomes we manually introduced into ligands from existing crystallographic structures the specific chemical moieties present in the highly similar analogs, and performed a minimization of each ligand−target complex followed by the estimation of the binding energy of each ligand bound conformation. We concluded that currently available scoring functions cannot rank order ligands based on their binding affinity. However, at the same time, encouraging results of virtual screening performed for one of the targets, FXa, showed that even though the ranking accuracy of the scoring functions remains challenging it still can discriminate between active compounds and decoys.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00751.

Figure S1. Statistical parameters obtained for the native/decoy classification models. Figure S2. Scores for 200 poses for each of the five SYK ligands as a function of RMSD. Figure S3. Scores for 200 poses for each of the 14 TRMD ligands as a function of RMSD. Figure S4. Summary for the results of Phase 1 provided by the organizers. Figure S5. Prediction accuracy as a result of different clustering methods. Figure S6. Summary for the results of Phase 2 provided by the organizers. Figure S7. Spearman correlation (ρ) provided by the organizers for all ligands in Phase 2. Figure S8. ROC plot for recovering FXa ligands by virtual screening of the DUD-E benchmarking database (PDF)

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

AUTHOR INFORMATION

Corresponding Authors

*Mailing address: 120 Mason Farm Road, CB no. 7260 Genetic Medicine, Chapel Hill, NC, 27599. Phone: +1 919 843-2513. E-mail: dokh@unc.edu (N.V.D.)

*Mailing address: 301 Pharmacy Lane, Chapel Hill, NC, 27599. Phone: +1 919 966-2955. E-mail: alex_tropsha@unc.edu (A.T.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge the financial support from NSF (ABI 1147145) to A.T. and from NIH (R01GM080742 and R01AI102732) to N.V.D. Schrodinger is acknowledged for their technical support. We are grateful to Chemaxon for providing us with a copy of their software.

REFERENCES

(1) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(2) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.

(3) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717−1736.

(4) Fourches, D.; Politi, R.; Tropsha, A. Target-Specific Native/Decoy Pose Classifier Improves the Accuracy of Ligand Ranking in the CSAR 2013 Benchmark. *J. Chem. Inf. Model.* **2015**, *55*, 63−71.

(5) Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Nelson, J. W.; Schena, A.; Jankowski, W.; Kalodimos, C. G.; Johnsson, K.; Stoddard, B. L.; Baker, D. Computational Design of Ligand-Binding Proteins With High Affinity and Selectivity. *Nature* **2013**, *501*, 212−216.

(6) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: an Update. *Nucleic Acids Res.* **2014**, *42*, D1083−90.

(7) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B.; Stuckey, J. A.; Carlson, H. A. CSAR Benchmark Exercise 2011−2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *J. Chem. Inf. Model.* **2013**, *53*, 1853−1870.

(8) Hsieh, J.-H.; Yin, S.; Wang, X. S.; Liu, S.; Dokholyan, N. V.; Tropsha, A. Cheminformatics Meets Molecular Mechanics: A Combined Application of Knowledge-Based Pose Scoring and Physical Force Field-Based Hit Scoring Functions Improves the Accuracy of Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52*, 16−28.

(9) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(10) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(11) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.

(12) Ding, F.; Dokholyan, N. V. Incorporating Backbone Flexibility in MedusaDock Improves Ligand-Binding Pose Prediction in the CSAR2011 Docking Benchmark. *J. Chem. Inf. Model.* **2013**, *53*, 1871−1879.

(13) Ding, F.; Yin, S.; Dokholyan, N. V. Rapid Flexible Docking Using a Stochastic Rotamer Library of Ligands. *J. Chem. Inf. Model.* **2010**, *50*, 1623−1632.

(14) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−7.

(15) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.

(16) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.

(17) Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V. MedusaScore: An Accurate Force-Field Based Scoring Function for Virtual Drug Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1656−1662.

(18) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534−546.

(19) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An Automated Approach for Clustering an Ensemble of NMR-derived Protein Structures Into Conformationally Related Subfamilies. *Protein Eng., Des. Sel.* **1996**, *9*, 1063−1065.

(20) Hsieh, J.-H.; Yin, S.; Liu, S.; Sedykh, A.; Dokholyan, N. V.; Tropsha, A. Combined Application of Cheminformatics- and Physical Force Field-Based Scoring Functions Improves Binding Affinity Prediction for CSAR Data Sets. *J. Chem. Inf. Model.* **2011**, *51*, 2027−2035.

(21) Singh, R. K.; Tropsha, A.; Vaisman, I. I. Delaunay Tessellation of Proteins: Four Body Nearest-Neighbor Propensities of Amino Acid Residues. *J. Comput. Biol.* **1996**, *3*, 213−221.

(22) Parr, R. G.; Szentpály, L. v.; Liu, S. Electrophilicity Index. *J. Am. Chem. Soc.* **1999**, *121*, 1922−1924.

(23) Liu, S.-B. Conceptual Density Functional Theory and Some Recent Developments. *Acta Phys. Chim. Sin.* **2009**, *25*, 590−600.

(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(25) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855−868.

(26) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On The Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−204.

(27) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494−3504.

(28) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.

(29) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−76.

(30) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(31) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R(2): Simple, Unambiguous Assessment of The Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316−1322.

(32) Martin, L. J.; Piltonen, M. H.; Gauthier, J.; Convertino, M.; Acland, E. L.; Dokholyan, N. V.; Mogil, J. S.; Diatchenko, L.; Maixner, W. Differences in the Antinociceptive Effects and Binding Properties

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041

of Propranolol and Bupranolol Enantiomers. *J. Pain* **2015**, *16*, 1321−1333.
(33) Xu, W.; Lucke, A. J.; Fairlie, D. P. Comparing Sixteen Scoring Functions for Predicting Biological Activities of Ligands for Protein Targets. *J. Mol. Graphics Modell.* **2015**, *57*, 76−88.
(34) Fourches, D.; Muratov, E.; Ding, F.; Dokholyan, N. V.; Tropsha, A. Predicting Binding Affinity of CSAR Ligands Using Both Structure-Based and Ligand-Based Approaches. *J. Chem. Inf. Model.* **2013**, *53*, 1915−1922.
(35) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(36) Nedumpully-Govindan, P.; Jemec, D. B.; Ding, F. CSAR Benchmark of Flexible MedusaDock in Affinity Prediction and Nativelike Binding Pose Selection. *J. Chem. Inf. Model.* **2015**, DOI: 10.1021/acs.jcim.5b00303.

DOI: 10.1021/acs.jcim.5b00751
J. Chem. Inf. Model. 2016, 56, 1032−1041