

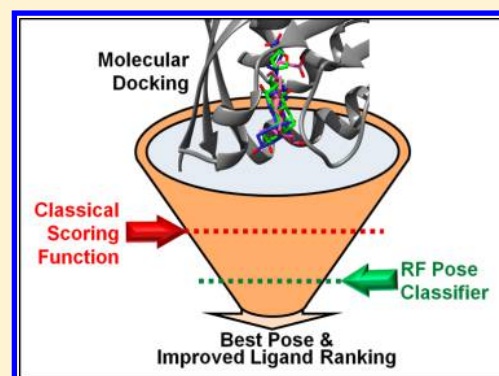
# Target-Specific Native/Decoy Pose Classifier Improves the Accuracy of Ligand Ranking in the CSAR 2013 Benchmark

Denis Fourches,<sup>†</sup> Regina Politi,<sup>†</sup> and Alexander Tropsha\*

Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States

## S Supporting Information

**ABSTRACT:** As part of the CSAR 2013 benchmark exercise, we have implemented a hybrid docking and scoring workflow to rank 10 steroid ligands of an engineered digoxigenin-binding protein. Schrödinger's Glide docking software was used to generate poses for each steroid ligand and rank them according to both standard docking precision (SP) and extra docking precision (XP) scoring functions. The unique component of our approach was the use of a target-specific pose classifier trained to discriminate nativelike from decoy poses. To build the classifier, a single cognate ligand with a known native pose (PDB code 4J8T) was docked multiple times into its target protein, and the generated poses were divided into two classes (nativelike and decoy) using a root-mean-square deviation threshold of 2 Å. All of the poses were characterized by the MCT-Tess descriptors of the protein–ligand interface, and random forest (RF) models were trained to discriminate the two classes of poses on the basis of their descriptors. The consensus pose classifier was then applied to the Glide-generated poses of each CSAR ligand in order to filter out those poses predicted as decoys and rerank the remaining ones using both XP and SP scoring functions. The best-scoring pose for each ligand following this filtering step was used for final ligand ranking. Overall, the ranking accuracy for the 10 ligands evaluated by the Spearman correlation coefficient was 0.64 for SP and 0.52 for XP but reached 0.75 for SP/RF consensus scoring (ranked third in the CSAR 2013 benchmark exercise). This study reconfirms that target-specific pose scoring models are capable of enhancing the reliability of structure-based molecular docking by discarding decoy poses.



## 1. INTRODUCTION

Virtual screening represents an inexpensive computational approach to identifying putative bioactive compounds within large chemical libraries.<sup>1</sup> When the three-dimensional structure of the target is available, screening workflows involve molecular docking approaches that (i) generate plausible poses of ligands within the binding pocket of the target and (ii) score those poses according to the predicted free energy of binding. However, scoring functions often fail to accurately forecast ligand binding affinities for their biological targets.<sup>2,3</sup> The 2013 CSAR exercise<sup>4</sup> provided the scientific community with another opportunity to evaluate and benchmark the reliability of various computational approaches for predicting protein–ligand interactions. As part of the exercise (Phase 3), the objectives for every participant were (i) to accurately predict nativelike poses of 10 potential steroid ligands toward an engineered digoxigenin-binding protein<sup>5</sup> and (ii) to rank this series of 10 ligands according to their predicted binding affinities.

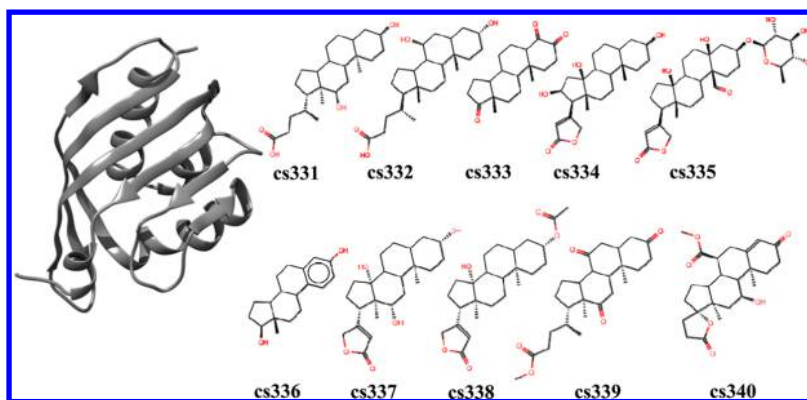
Although the 2013 CSAR exercise was unconventional because of the atypical nature of the biological target, we still considered this community exercise as a unique chance to objectively compare the overall ranking reliability of different structure-based approaches used by different participants. It is important to underline that as in the previous editions of the

benchmarking exercise, CSAR organizers<sup>6,7</sup> did not put any restrictions on the use of external publicly available data, methods, and software. The participants were even encouraged to use as many sources of any potentially useful information as possible.

In the present work, we employed a unique consensus prediction approach that incorporates two different types of methods: (i) molecular docking that predicts the binding poses of CSAR ligands and ranks them according to their docking scores and (ii) cheminformatics-inspired classification models built with chemical descriptors of the protein–ligand interface<sup>8–11</sup> to discriminate nativelike poses from decoys. Our results confirm that the prediction accuracy of virtual screening can be significantly boosted by using target-specific classification models to filter poses generated by conventional molecular docking approaches. Similar to studies reported previously by us<sup>10–12</sup> and others,<sup>13–17</sup> the approach employed herein exploits the synergies between independent approaches to scoring protein–ligand interactions in a consensus way as opposed to contrasting them. This study directly complements our previous investigation<sup>12</sup> integrating binding affinity

Received: August 27, 2014

Published: December 18, 2014



**Figure 1.** CSAR 2013 Phase 3 data set: structures of the engineered protein and the 10 steroid ligands.

predictions from ligand-based quantitative structure–activity relationship (QSAR) models and structure-based molecular docking.

## 2. METHODS

**2.1. Data Set.** The organizers of the CSAR 2013 benchmark exercise provided participants with one protein structure and a set of 10 potential ligands (see Figure 1). Interestingly, this protein was specifically designed<sup>5</sup> to bind the steroid digoxigenin (DIG). This protein design study was published recently and was *not* available at the time we conducted the study reported in this paper. After downloading the target protein structure from the CSAR Web site, we aligned this structure against known protein structures using the DALI Web server.<sup>18</sup> This analysis revealed that the CSAR protein shared 100% sequence identity with DIG10.2 protein (PDB code 4J8T, resolution 2.05 Å) and 95% sequence identity with DIG10.3 protein (PDB code 4J9A, resolution 3.20 Å). Both the DIG10.2 and DIG10.3 proteins were designed to bind DIG, and both protein structures were cocrystallized with DIG, which also appeared to be the CSAR ligand cs337 (see Figure 1). The remaining nine ligands were steroid analogues of DIG, with the ligand cs336 being  $\beta$ -estradiol. It is important to emphasize again that *no* information concerning the experimental binding affinities of those ligands toward either DIG10.2 or DIG10.3 was available at the time of the study.

**2.2. Molecular Docking.** The crystallographic structure of DIG10.2 was preprocessed using the Protein Preparation wizard in the Schrödinger Suite (version 9.3; see <http://www.schrodinger.com/>). In this study, all of the small-molecule structures were docked into the active site of the target protein by using Glide with standard docking precision (Glide SP) and extra docking precision (Glide XP) modes.<sup>19</sup> Explicit hydrogen atoms were added and ionizable compounds were converted to their most probable charged forms at pH 7.0  $\pm$  2.0 using LigPrep. Conformational sampling was performed for all of the ligands using ConfGen.<sup>20</sup> The binding region was defined by a 10 Å  $\times$  10 Å  $\times$  10 Å grid box centered on DIG. A scaling factor of 1.0 was applied to the van der Waals radii. Default settings were used for all of the remaining parameters. For each ligand, the top 100 generated poses were retained. The ligands were sorted according to the docking scores (termed SP and XP scores in this article, depending on the precision mode used) of the associated top-ranking poses.

**2.3. PL/MCT-Tess Descriptors.** Our group recently developed the PL/MCT-Tess approach to characterize protein–ligand interfaces.<sup>11</sup> PL/MCT-Tess descriptors employ

pairwise atomic potentials for the protein–ligand complexes (PL) based on maximal charge transfer (MCT).<sup>21</sup> The MCT characterizes the maximal electron flow between the donor and acceptor atoms at the protein–ligand interface. It is derived from conceptual density functional theory,<sup>21,22</sup> which provides a theoretical basis for calculating the PL/MCT-Tess descriptors. The values of the PL/MCT-Tess descriptors are calculated from the following equation:

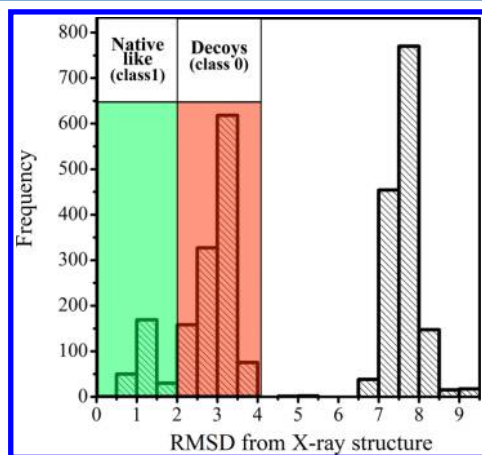
$$\text{PL/MCT-Tess}_m = \sum_{k=1}^n \sum_p^{1-3} \sum_l^{1-3} (\text{MCT}_p \cdot \text{MCT}_l / d_{pl})_k \quad (1)$$

where PL/MCT-Tess<sub>m</sub> is the potential of the *m*th 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<sub>pl</sub>* is the distance between a pair of protein and ligand atoms found in the same Delaunay tetrahedron.

For a given complex, Delaunay tessellation partitions the receptor–ligand (R–L) interface into an aggregate of space-filling, irregular tetrahedra in which both receptor and ligand atoms become vertices.<sup>11</sup> Four atom types (C, N, O, and S) are defined for receptor proteins and six atom types (C, N, O, S, P and halogens as X, and metals as M) for ligands. 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). For each quadruplet, we calculate the sum of MCT values of the composing atom vertices, and this sum represents the actual descriptor value. Considering all possible types of ligand–receptor quadruplets (R–R–L–L, R–L–L–L, and R–R–R–L) and all atom types, each pose is uniquely characterized by a vector of 554 MCT descriptor values (see our previous study<sup>11</sup> for additional details about the method).

**2.4. Generation of the Modeling Set of Nativelike and Decoy Poses.** The Schrödinger suite was used to generate an ensemble of nativelike and decoy poses for the DIG cognate ligand using the 4J8T (DIG10.2 protein) PDB structure. As many as 1428 and 530 poses with root-mean-square deviations (RMSDs) smaller than 4 Å from the native pose were generated using Glide SP and Glide XP, respectively, on the basis of the default parameters. When compared and represented as a heat map (see Figure S1 in the Supporting Information), these two sets of poses appeared relatively similar as evaluated by the distribution of their pairwise RMSDs: the

vast majority of poses from one group have a similar pose with  $\text{RMSD} < 1 \text{ \AA}$  in the other group. For the purpose of modeling, we defined an RMSD threshold of  $2 \text{ \AA}$  to discriminate native-like poses from decoys. This threshold was consistent with the gap in the distribution plot of poses generated by redocking of the cognate ligand (Figure 2). Furthermore, poses



**Figure 2.** Frequency distribution of poses vs RMSD. Poses were generated by redocking ligand cs337 (cognate ligand). RMSDs were computed against the native pose of the cognate ligand from the crystal structure (PDB code 4J8T). Fully flipped poses with  $\text{RMSD} > 4 \text{ \AA}$  are shown in gray.

with RMSD larger than  $4 \text{ \AA}$  were not taken into account since they were found to be fully flipped compared with native-like poses. For SP, 260 poses with RMSD smaller than  $2 \text{ \AA}$  were considered as native-like (class 1), whereas 284 randomly chosen poses with RMSD larger than  $2 \text{ \AA}$  were considered as decoys (class 0). Similarly for XP, 117 poses were assigned to class 1 and 100 poses to class 0.

**2.5. Pose Classifier.** In this study, we used a classical QSAR-like modeling approach<sup>23</sup> to build models that evaluate the likelihood of a ligand pose to be native-like. Thus, the problem of discriminating native-like poses from decoys can be addressed 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 has published elsewhere.<sup>24,25</sup> Three steps<sup>24,26</sup> can be defined: (1) data curation/preparation/analysis (selection of poses and descriptors), (2) model building, and (3) model validation/selection. Here we applied a fivefold external cross-validation procedure: the full set of poses was split randomly into five groups of equal size and constant balance of classes. Four different groups were systematically used as the modeling set (80% of the full set), and the remaining group was employed as an external validation set (remaining 20%). This procedure was repeated five times with each group used as the external validation set once, allowing cumulative statistics to be computed for the whole set of poses. Random forest (RF) implemented in R 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 to build or select the models, this condition being critical to ensure the rigor of the external validation

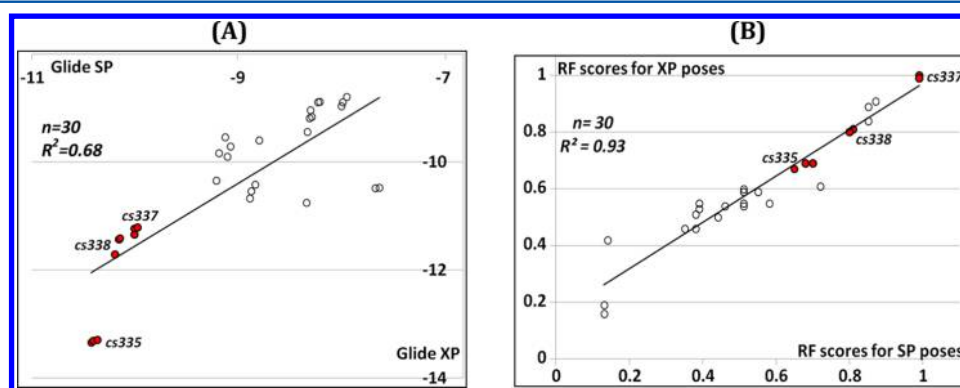
procedure.<sup>27</sup> Additionally, each modeling set was split into multiple internal training and test sets. Models were built using poses belonging to each training set and applied to internal test set poses for assessing their nativelikeness.

The best models were identified and selected according to their prediction performances on the respective training and internal test sets using a threshold ( $\geq 0.7$ ) of specificity, sensitivity, balanced accuracy (BA), and the Spearman's rank correlation coefficient. Then all of the 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 was present once (and only once) in the external test set. It is important to emphasize that in this protocol, ligand poses in the external test sets were never used to derive, bias, or select the models; thus, the entire procedure gives a fair estimation of the true external predictivity of the models. In addition, Y-randomization (randomization of class assignments for each ligand) was performed for each selected model in order to ensure that there was no chance correlation between actual and predicted classes of ligands. It is expected that models obtained for the training set with randomized responses should have significantly lower predictivity compared with 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 of the training/test data set divisions. All validated models were then stored and used in an ensemble for predicting the nativelikeness of any external pose.

**2.6. Combining Docking and Pose Classifier Scores for Pose and Ligand Ranking.** For any ligand–protein pose, the ensemble of selected RF models outputs a continuous, averaged consensus score (RF score) ranging from 0 (pose predicted to be a decoy by all models) to 1 (pose predicted to be native-like by all models). In other words, the consensus RF score takes a value between 0 and 1 depending on the ratio of models predicting a pose as native-like or non-native. 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 thus be used to rank those poses on the basis of their decreasing RF-evaluated likelihood of being native-like. This treatment is particularly interesting as it offers an additional approach for quantitative ranking of poses on top of the scoring function from the conventional molecular docking program. For each CSAR ligand, the whole set of Glide-generated poses 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 three poses on the basis of predicted RF scores were chosen for each ligand. In this study, all of the poses resulted from docking using Glide SP and Glide XP modes,<sup>19</sup> so they were characterized by SP and XP Glide scores, respectively. Out of the top three poses on the basis of RF scores, the pose associated with the best Glide docking score (SP score for the poses created with Glide SP and XP score for the poses created with Glide XP) was further selected and used to rank the 10 CSAR ligands. This pose-selection protocol based on progressive use of RF and Glide scores resulted in a new ligand ranking that was different from the original ranking based on Glide scoring only. The rank obtained using this workflow was ultimately compared to the experimental rank resulting from measured binding affinities.

Table 1. CSAR Ligand Ranking According to SP and XP Docking Scores

compound	SP scoring function			XP scoring function		
	docking scores	ranking based on the best pose	ranking based on average score of the three best poses	docking scores	ranking based on the best pose	ranking based on average score of the three best poses
cs331	−9.10	5	5	−9.90	7	7
	−9.07			−9.71		
	−8.79			−9.59		
cs332	−9.21	4	4	−10.34	6	6
	−9.18			−9.83		
	−9.12			−9.54		
cs333	−7.99	10	9	−8.96	10	10
	−7.98			−8.89		
	−7.94			−8.79		
cs334	−8.88	6	6	−10.67	5	5
	−8.87			−10.54		
	−8.83			−10.41		
cs335	−10.42	1	1	−13.34	1	1
	−10.40			−13.31		
	−10.36			−13.29		
cs336	−8.28	9	8	−9.16	9	9
	−8.22			−8.89		
	−8.20			−8.88		
cs337	−10.00	3	3	−11.34	3	3
	−10.00			−11.23		
	−9.97			−11.21		
cs338	−10.19	2	2	−11.71	2	2
	−10.15			−11.43		
	−10.14			−11.41		
cs339	−8.33	7	10	−10.75	4	4
	−7.66			−10.48		
	−7.62			−10.47		
cs340	−8.32	8	7	−9.44	8	8
	−8.30			−9.18		
	−8.29			−9.04		



**Figure 3.** Correlations between (A) XP and SP Glide docking scores obtained for the 10 CSAR ligands (top three poses per ligand as given in Table 1) and (B) RF/MCT-Tess scores obtained for XP and SP poses (top three poses per ligand as given in Table 2).

### 3. RESULTS AND DISCUSSION

**3.1. Molecular Docking of CSAR Compounds.** All 10 CSAR compounds were docked using both Glide SP and Glide XP and ranked according to their respective Glide docking scores. Table 1 summarizes the docking scores obtained for the three top-ranking poses of each ligand. With SP, the best-scored pose (SP score = −10.42) was retrieved for the ligand cs335, whereas the worst pose (SP score = −7.62) was associated with ligand cs339. With the XP scoring function, the best pose (XP score = −13.34) was also retrieved for ligand

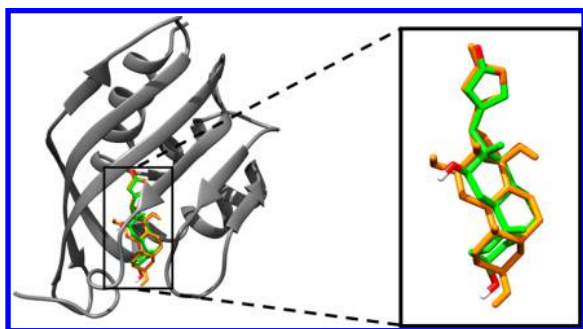
cs335, whereas ligand cs333 obtained the worst score (XP score = −8.79).

Overall, the correlation between the SP and XP scores for the 30 best ligand poses was found to be as high as  $R^2 = 0.67$  (Figure 3A). Similarly, the Spearman rank correlation coefficient between SP and XP scores for top-scoring poses was also very high ( $\rho = 0.71$ ). Importantly, ligand cs335 showed the best docking score using both SP and XP scoring functions and was thus ranked number 1. Ligand cs333 was ranked 10th.



**3.2. Prediction Performances of the RF Classification Models.** RF models based on MCT-Tess descriptors led to very high prediction performances (BA = 0.99; Figure S2 in the Supporting Information) in discriminating nativelike poses from decoys as evaluated by an external fivefold cross-validation (see Methods). The predictivity of these models is indicative of the clear separation between nativelike and decoy poses as shown in Figure 2.

Following the same idea and protocols developed in our earlier studies,<sup>11</sup> we attempted to filter out Glide-generated poses predicted to be decoys by the RF/MCT-Tess classification model. Thus, the RF model was used to identify and eliminate non-native poses for all 10 CSAR compounds. Figure 4 shows the best pose obtained for ligand cs337



**Figure 4.** Best pose obtained for ligand cs337 (represented in orange) on the basis of RF score, superimposed with the cognate ligand from the crystallographic structure 4J8T (represented in green). This pose was generated by regular SP docking of the compound into the engineered protein provided by the CSAR organizers. The SP docking score and RF score for the pose are given in Table 2. The RMSD between the docking pose and the native crystallographic structure is below 1 Å.

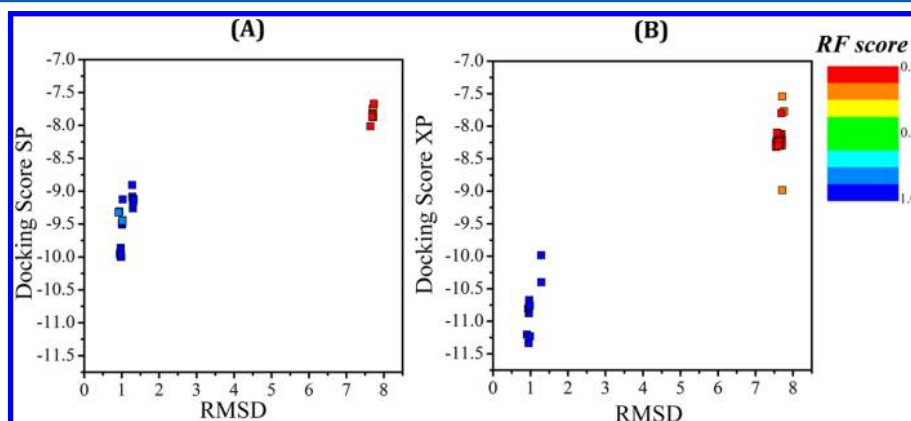
according to the RF score. The RMSD of this pose was below 1 Å compared with the native pose of the ligand in the 4J8T PDB structure. Furthermore, we plotted the docking scores as a function of RMSD for all of the poses of ligand cs337 generated with Glide (Figure 5). For this particular ligand, nativelike poses are perfectly discriminated from decoys using both the SP and XP scoring functions as well as using our binary classification model (Figure 5).

For each CSAR ligand, the three poses with the highest RF scores were retained (see Table 2). The closer to 1 the

**Table 2.** CSAR Ligand Scores According to SP/RF and XP/RF Scoring Approaches

compound	SP docking scores	RF scores for SP poses	XP docking scores	RF scores for XP poses
cs331	−8.49	0.44	−8.30	0.50
	−6.99	0.38	−8.35	0.46
	−7.21	0.35	−8.37	0.46
cs332	−7.63	0.51	−8.87	0.55
	−7.92	0.51	−6.80	0.54
	−8.32	0.46	−8.77	0.54
cs333	−6.93	0.72	−8.67	0.61
	−7.53	0.55	−8.93	0.59
	−7.58	0.39	−8.85	0.55
cs334	−8.88	0.87	−8.18	0.91
	−8.51	0.85	−7.80	0.89
	−8.53	0.85	−8.11	0.84
cs335	−8.97	0.70	−10.59	0.69
	−9.85	0.68	−10.53	0.69
	−9.12	0.65	−12.16	0.67
cs336	−8.28	0.51	−8.90	0.60
	−8.19	0.51	−8.89	0.59
	−8.22	0.51	−8.92	0.59
cs337	−9.13	0.99	−10.40	1.00
	−9.95	0.99	−10.88	0.99
	−9.86	0.99	−9.99	0.99
cs338	−9.95	0.81	−11.71	0.81
	−10.06	0.81	−11.43	0.81
	−10.04	0.80	−11.07	0.80
cs339	−8.33	0.58	−10.18	0.55
	−7.60	0.39	−10.75	0.53
	−7.66	0.38	−10.05	0.51
cs340	−8.29	0.14	−9.44	0.42
	−8.26	0.13	−9.10	0.19
	−8.30	0.13	−8.15	0.16

consensus RF score predicted by the ensemble of pose-classifying models was, the higher was the probability that this pose was nativelike. Interestingly, there was no clear correlation between RF scores and Glide docking scores: the correlation



**Figure 5.** Distribution of binding poses generated for ligand cs337. (A) Pose distribution based on RMSD values vs Glide SP docking scores. (B) Pose distribution based on RMSD values vs Glide XP docking scores. Data points are colored according to their RF scores (the closer to 1, the more likely to be nativelike).

was not insignificant for SP ( $R^2 = 0.39$ ) but was as low as  $R^2 = 0.12$  for XP. On the contrary, the pairwise correlation between RF scores obtained for SP and XP poses was very high ( $R^2 = 0.93$ ; see Figure 3B). Of note was also the low correlation between SP and XP scores for filtered-out poses ( $R^2 = 0.36$ ; data not shown).

Finally, poses predicted to be nativelike by the classifier for the 10 ligands were reranked on the basis of their docking score with either the Glide SP or Glide XP scoring function, keeping in mind that higher ranks correspond to lower binding affinities. We submitted the best poses and the predicted ranks of the 10 ligands computed using all four approaches (SP, SP/RF, XP, XP/RF) to the CSAR organizers for the evaluation of this blind prediction for novel compounds.

**3.3. Comparison between Experimental and Predicted Ranks for the CSAR Ligands.** After the CSAR exercise was over, the organizers released the experimental data related to the 10 ligands (the crystallographic structures of the protein–ligand complexes were still not available at the time this paper was written, so we cannot discuss the pose accuracy).

The predicted and experimental ranks for the CSAR ligands are recapitulated in Table 3. Overall, the best correlation with

SP/RF (rank = 8) and incorrectly assessed by XP (rank = 4) and XP/RF (rank = 4).

**3.4. Success or Failure?** Our SP/RF model with its good prediction performances ( $\rho = 0.75$ ,  $R^2 = 0.75$ ) was ranked third out of 31 submissions for this CSAR 2013 Phase 3 Community Challenge (cf. csardock.org). Although the comparison is based on 10 ligands only, it is still interesting to note that out of 31 submissions, only five afforded very good prediction performances (when measured as experimental vs predicted responses with a minimum threshold of  $R^2 \geq 0.70$ ). Thus, we shall discuss whether our results should be regarded as fairly successful or as a failure in the CSAR 2013 benchmark exercise.

All of the active compounds were correctly identified and ranked accordingly by all four different scoring functions employed in our study. This is indeed an important result confirming that molecular docking can be employed as a reliable tool to prioritize compounds for experimental confirmatory testing, especially when these compounds belong to a congeneric set. In addition, our results show 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. This is a direct consequence of filtering out ligand poses predicted to be non-nativelike according to the classification model and yet scored relatively high using conventional scoring functions. As a result, pose classifiers can complement the traditional scoring functions used in any molecular docking software. The RF/MCT classifier described in this study can be considered as target-specific since it was specifically trained and validated to discriminate nativelike poses from a set of decoys for that particular target, as opposed to physics-based scoring functions usually fitted on a panel of diverse protein–ligand complexes. Moreover, a given RF/MCT classifier is likely to be binding-mode-specific, especially for such an engineered protein–steroid complex.

However, we should also emphasize that our target-specific RF models failed to identify nativelike poses for certain CSAR ligands. In some cases, the best-ranked pose did not change as a result of filtering on the basis of the RF classification model. For instance, the application of RF/MCT-Tess filtering to SP poses led to an improvement in the ranking accuracy for cs331, cs332, cs334, cs335, cs336, cs337, and cs338, whereas it decreased the ranking accuracy for cs339 and cs340. For XP, one should also note the significant loss of ranking accuracy for cs334 and cs340.

To test whether the chemical similarity between different CSAR ligands and the cognate ligand used to develop the RF models was influential for the prediction accuracy, we computed the pairwise Tanimoto similarity coefficient ( $T_c$ ) between cs337 and the other CSAR ligands (the data are given in Table S1 in the Supporting Information). We found no direct correlation between the ligands' two-dimensional similarity with cs337 and the reliability of their ranking. Some of the compounds with high similarity to the cognate ligand cs337 were correctly ranked by the different scoring approaches, such as cs334 ( $T_c = 0.99$ ), cs338 ( $T_c = 0.98$ ), cs331 ( $T_c = 0.94$ ), cs332 ( $T_c = 0.94$ ), and cs335 ( $T_c = 0.94$ ), whereas cs340 ( $T_c = 0.95$ ) was not.

Two of these compounds, cs331 and cs332, were found to be highly similar to cs337 ( $T_c = 0.94$ ). However, the best-scoring poses of these two ligands had RF scores as low as 0.44 and 0.51, respectively, meaning that among all of the poses generated for these compounds, none was actually predicted

**Table 3. Predicted and Experimental Ranks for the 10 CSAR Ligands**

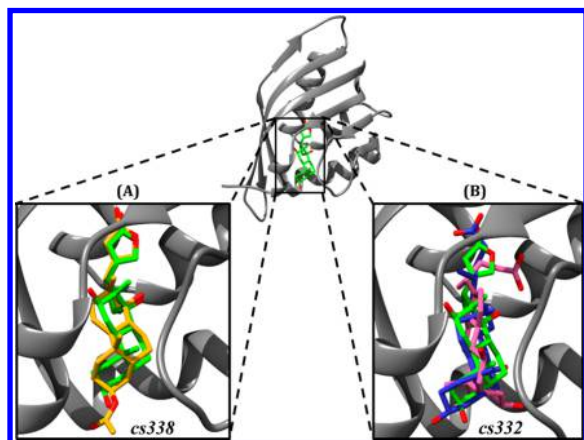
compound	predicted ranks using				experimental <sup>a</sup>	
	XP	XP/RF	SP	SP/RF	rank	$pK_d$
cs331	7	8	5	9	7	4.35
cs332	6	9	4	7	8	4.34
cs333	10	7	10	10	6	4.51
cs334	5	10	6	4	4	5.49
cs335	1	2	1	3	3	5.51
cs336	9	5	8	6	5	4.72
cs337	3	3	3	2	2	6.32
cs338	2	1	2	1	1	6.66
cs339	4	4	9	8	10	4.10
cs340	8	6	7	5	9	4.27
Spearman $\rho$	0.52	0.48	0.64	0.75		

<sup>a</sup>Obtained from the CSAR organizers after the end of the competition and the full publication of the results among all participants.

experimental data for the 10 ligands (Spearman  $\rho = 0.75$ ) was achieved by SP/RF (i.e., poses created by Glide SP, filtered by RF/MCT-Tess models, and reranked on the basis of the Glide SP scoring function). It is interesting to note that (i) the three other approaches afforded lower but still reasonable prediction performances ( $\rho = 0.48$ – $0.64$ ); (ii) SP ( $\rho = 0.64$ ) was slightly more accurate than XP ( $\rho = 0.52$ ); and (iii) the RF/MCT-Tess filtering helped significantly to improve the discrimination between nativelike and decoy poses generated by SP ( $\rho$  value increasing from 0.64 to 0.75) compared with XP ( $\rho$  value actually decreasing slightly from 0.52 to 0.48).

The results also showed that the four most active compounds ( $pK_d \geq 5.50$ ) were ranked as the top four ligands regardless of the scoring approach. For instance, compound cs338 ( $pK_d = 6.66$ , experimental rank = 1) was ranked first by SP/RF and XP/RF and ranked second by both SP and XP. However, cs335 ( $pK_d = 5.51$ , experimental rank = 3) was correctly predicted by SP/RF (rank = 3) but its predicted rank was overestimated by SP (rank = 1), XP (rank = 1), and XP/RF (rank = 2). The most inactive compound among the 10 CSAR ligands was cs339 ( $pK_d = 4.10$ ); it was correctly predicted by SP (rank = 9) and

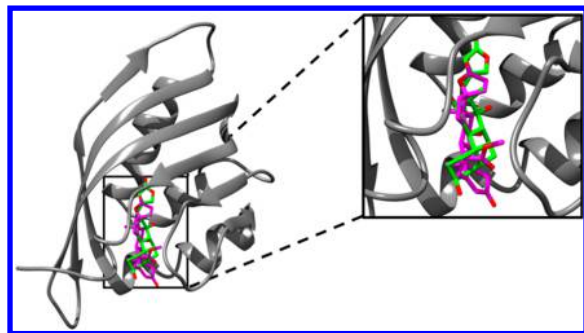
to be nativelike (which is also in line with their weak  $pK_d < 4.5$ ). Figure 6 shows two poses for cs332: the best pose on the



**Figure 6.** Visualization of the cognate ligand superimposed with the best poses obtained for the ligands cs332 and cs338. (A) Cognate ligand (represented in green) and the best pose for cs338 (represented in yellow) after using the RF pose classification model for filtering. (B) cognate ligand and two poses for cs332: the best pose selected solely on the basis of the SP scoring function (pink) and best pose after using the RF pose classification model for filtering (blue).

basis of the SP scoring function (SP score =  $-9.21$  and RF score =  $0.16$ ) and the best pose after the RF filtering (SP score =  $-7.63$  and RF score =  $0.51$ ). Although the pose selected solely on the basis of the SP score superimposes with cs337 reasonably well, this pose was filtered out by the RF classification model, correctly pushing this compound from fourth to seventh place (experimental rank = 8). This case illustrates quite well how and why the RF model improved the ranking accuracy of CSAR ligands: indeed, the cs332 pose with the best SP score is actually predicted to be a decoy according to the RF model, thereby facilitating the identification of false positives. In the cases of ligands cs331 and cs332, the poses remaining after the RF-based filtering have less favorable docking scores than cs337, as confirmed by their respective experimental activities. Figure 6 also shows the best pose obtained for cs338, which is extremely well superimposed with cs337 and has good docking and RF scores.

Interestingly, despite being highly similar to cs337 ( $T_c = 0.95$ ), the ligand cs340 (Figure 7) was ranked fifth on the basis of the SP/MCT-Tess method, whereas its experimental rank was only ninth. The best RF score obtained for selected SP



**Figure 7.** Cognate ligand cs337 (represented in green) superimposed with the best pose obtained for cs340 after applying the RF classification model (represented in magenta).

poses of cs340 was  $0.14$  ( $0.42$  for XP). Again, among all of the cs340 poses generated by Glide, none seemed to be nativelike on the basis of RF scores. With a  $pK_d$  of  $4.27$ , cs340 was found to be the second most inactive compound.

In general, ranking non-native poses better than native poses is actually an undesirable feature of any scoring function. Therefore, our RF model is trained to remove ligand poses predicted to be non-native. If those poses score better than the native ones, then their removal by our classification method would improve the results of the conventional scoring function (as we observed in case of SP). If the native poses are scored poorly and non-native poses incidentally give more accurate rank than the native poses, then the results of our filtering could actually decrease the accuracy of scoring (as we observed in the case of XP).

Therefore we should expect that our filtering approach would not improve the results obtained with a poorly performing scoring function that does not produce accurate ranking for both nativelike and non-native poses. However, our method could improve the outcome of a scoring function that on occasion scores non-native poses better than native ones while accurately scoring nativelike poses. Thus, the overall limitation is not primarily that of our method but rather that of the scoring function(s) employed in the study.

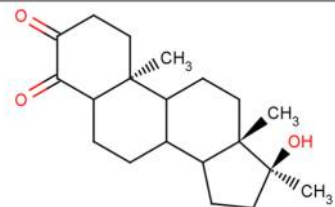
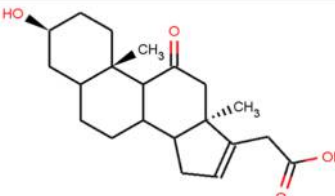
**3.5. Virtual Screening of ZINC.** In order to identify additional DIG analogues with some potential affinity for the engineered protein, we searched the ZINC database (<http://zinc.docking.org/>) comprising  $\sim 2.4$  million druglike compounds. Specifically, we searched for steroid compounds with a high pairwise structural similarity to DIG. We identified a set of 5386 compounds with  $T_c \geq 0.9$  using the ISIDA/Similarity program.<sup>28</sup> These compounds were docked using the Glide SP scoring function, and only those poses with RF scores higher than  $0.7$  were retained. The chemical structures of two of these compounds, ZINC04655335 and ZINC71770245, are shown in Table 4; their SP scores are  $-8.83$  and  $-8.71$ , respectively, and their RF score is  $0.88$ . The best SP docking score for any ZINC compound was found to be equal to  $-10.7$  (ZINC38557509), however the RF score of this pose was very low ( $0.57$ ). This virtual screening study was not expected by the CSAR 2013 organizers; however, we included it in our work to illustrate the utility of our hybrid docking approach for virtual screening of large chemical libraries to identify novel putative binders.

## 4. CONCLUSIONS

The CSAR benchmark is a great initiative that enables the unbiased comparison and benchmarking of scoring functions and docking protocols. The main goal of this study was to reliably assess the relative ranking of CSAR ligands by predicting their potency toward an engineered target protein. This exercise was interesting, compared with traditional docking/scoring problems, because by design no data were available to train ligand-based QSAR models and/or recalibrate the scoring function. We have established that our hybrid docking/pose classification/rescoring workflow afforded a significant improvement in terms of ligand ranking accuracy over the use of conventional docking alone. This result was mainly achieved by an efficient filtering of ligand poses predicted to be decoys: this filtering eliminated best-scoring poses for some ligands, changing their relative rankings on the basis of their respective docking scores. We showed that all of the active compounds were correctly identified and ranked by



Table 4. Structures of Two Potential DIG Binders Identified in the ZINC Database

ID	Structure	SP Docking Score
ZINC04655335		-8.83
ZINC71770245		-8.71

our approach. Overall, the ranking accuracy for the 10 ligands evaluated by the Spearman correlation coefficient was as high as 0.75 for the SP/MCT-Tess consensus scoring (third-best in the CSAR competition). In previous studies using similar protocols, we found that (i) filtering poses ranked by the Medusa docking program<sup>29</sup> improved the hit rate in virtual screening<sup>10,11</sup> and (ii) ligand-based and structure-based ranking approaches could be combined to improve the prediction performances of ligand ranking.<sup>12</sup> In the present study, we used Glide docking software in combination with the same pose-scoring approach as before. Thus, our results reconfirm that integrating a pose-filtering step based on a target-specific pose classifier can be used as a general tool to improve the accuracy of molecular docking methods relying on conventional scoring functions.

## ■ ASSOCIATED CONTENT

### ■ Supporting Information

Pairwise Tanimoto similarity coefficients computed between ligand cs337 and the other nine CSAR ligands using ISIDA/Similarity software and ISIDA fragmental descriptors (Table S1); pairwise RMSD values between poses generated by Glide XP (*X* axis) and Glide SP (*Y* axis) ordered on the basis of their respective RMSDs to the cognate ligand (PDB code 4J8T) (Figure S1); and statistical parameters obtained for the native/decoy classification models (poses are discriminated using a RMSD threshold of 2 Å), with external prediction performances (sensitivity, specificity, and CCR) estimated according to a fivefold external cross-validation procedure with/without taking into account the models' applicability domains (Figure S2). This material is available free of charge via the Internet at <http://pubs.acs.org>.

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: alex\_tropsha@unc.edu.

### Author Contributions

†D.F. and R.P. contributed equally to this work.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors acknowledge the financial support from NSF (ABI 1147145). D.F. also thanks UNC-Chapel Hill for the Junior

Faculty Development Award. Schrödinger is thankfully acknowledged for their technical support. Chemaxon is also acknowledged for providing us with JChem and the Marvin suite.

## ■ REFERENCES

- (1) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862–865.
- (2) 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.
- (3) 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.
- (4) Irwin, J. J. Community Benchmarks for Virtual Screening. *J. Comput.-Aided. Mol. Des.* **2008**, *22*, 193–199.
- (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) 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.
- (7) Dunbar, J. B.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y.-N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.; Carlson, H. A. CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *J. Chem. Inf. Model.* **2013**, *53*, 1842–1852.
- (8) Da, C.; Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* **2014**, *54*, 2555–2561.
- (9) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.
- (10) 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.
- (11) 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.

(12) 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.

(13) Chupakhin, V.; Marcou, G.; Baskin, I.; Varnek, A.; Rognan, D. Predicting Ligand Binding Modes from Neural Networks Trained on Protein–Ligand Interaction Fingerprints. *J. Chem. Inf. Model.* **2013**, *53*, 763–772.

(14) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein–Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* **2014**, *54*, 944–955.

(15) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein–Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169–1175.

(16) Zilian, D.; Sottriffer, C. A. SFCscore(RF): A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923–1933.

(17) Sulimov, A. V.; Kutov, D. C.; Oferkin, I. V.; Katkova, E. V.; Sulimov, V. B. Application of the Docking Program SOL for CSAR Benchmark. *J. Chem. Inf. Model.* **2013**, *53*, 1946–1956.

(18) Holm, L.; Rosenstrom, P. Dali Server: Conservation Mapping in 3D. *Nucleic Acids Res.* **2010**, *38*, W545–W549.

(19) 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.

(20) 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.

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

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

(23) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.

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

(25) 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*, 1189–1204.

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

(27) Golbraikh, A.; Tropsha, A. Beware of q<sup>2</sup>! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

(28) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA—Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided. Drug Des.* **2008**, *4*, 191–198.

(29) 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.