Predicting Binding Affinity of CSAR Ligands Using Both Structure-Based and Ligand-Based Approaches

Denis Fourches,[†] Eugene Muratov,[†,‡] Feng Ding,[§] Nikolay V. Dokholyan,[§] and Alexander Tropsha*,[†]

[†]Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, United States

[‡]Laboratory of Theoretical Chemistry, Department of Molecular Structure, A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Odessa 65080, Ukraine

[§]Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599, United States

ABSTRACT: We report on the prediction accuracy of ligand-based (2D QSAR) and structure-based (MedusaDock) methods used both independently and in consensus for ranking the congeneric series of ligands binding to three protein targets (UK, ERK2, and CHK1) from the CSAR 2011 benchmark exercise. An ensemble of predictive QSAR models was developed using known binders of these three targets extracted from the publicly available ChEMBL database. Selected models were used to predict the binding affinity of CSAR compounds toward the corresponding targets and rank them accordingly; the overall ranking accuracy evaluated by Spearman correlation was as high as 0.78 for UK, 0.60 for ERK2, and 0.56 for CHK1, placing our predictions in the top 10% among all the participants. In parallel, MedusaDock, designed to predict reliable docking poses, was also used for ranking the CSAR ligands according to their docking scores; the resulting accuracy (Spearman correlation) for UK, ERK2, and CHK1 were 0.76, 0.31, and 0.26, respectively. In addition, performance of several consensus approaches combining MedusaDock- and QSAR-predicted ranks altogether has been explored; the best approach yielded Spearman correlation coefficients for UK, ERK2, and CHK1 of 0.82, 0.50, and 0.45, respectively. This study shows that (i) externally validated 2D QSAR models were capable of ranking CSAR ligands at least as accurately as more computationally intensive structure-based approaches used both by us and by other groups and (ii) ligand-based QSAR models can complement structure-based approaches by boosting the prediction performances when used in consensus.
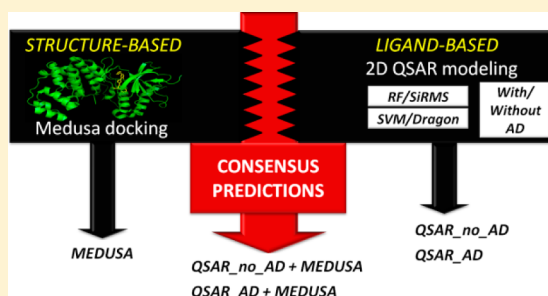
1. INTRODUCTION

The 2011 CSAR benchmark exercise provided the scientific community with the opportunity to evaluate and benchmark the reliability of the various computational approaches for predicting protein—ligand interactions. Four targets were considered: UK (urokinase), ERK2 (mitogen-activated protein kinase ERK2), CHK1 (checkpoint kinase 1), and LPXC (*Pseudomonas aeruginosa* UDP-3-O-acyl-GlcNAc deacetylase). The objectives for every participant were to (i) accurately predict the binding pose of each CSAR ligand and (ii) rank the series of CSAR ligands for different molecular targets according to an assessment of ligands' binding affinity for each target.

The first objective of the CSAR exercise was clearly thought as a "classical" benchmarking of molecular docking approaches for predicting native-like accurate binding poses of new ligands toward known targets. Meanwhile, our group especially welcomed the second objective as a unique opportunity to employ ligand-based approaches for ranking the CSAR ligands and compare their overall ranking reliability with that obtained by structure-based approaches used both by other participants and by our group. It is important to underline that CSAR organizers did not put any restrictions on the use of external

publicly available data, methods, and software.[1] The participants were even encouraged to use as many sources of any potentially useful information as possible. We also envisioned the possibility of employing both ligand-based[2] and structure-based[3,4] approaches and exploring some potential complementarities to rank CSAR ligands more accurately.

In this study, we aimed at assessing and ranking the binding affinities of CSAR ligands using a unique consensus approach (Table 1) that employed two different types of methods: (i) Quantitative Structure—Activity Relationship (QSAR) models built with known inhibitors of UK, ERK2, and CHK1 using two-dimensional molecular descriptors and machine learning techniques, (ii) MedusaDock molecular docking that predicts both the binding poses of CSAR ligands and the corresponding molecular affinities. Beyond the comparison of the prediction power of QSAR models versus structure-based docking,[5] we pursued the idea of exploring the benefits of using ligand-based

Table 1. List of Approaches Used in This Study To Rank CSAR Ligands (see text for more details)

type	name	ID	description
Ligand-based (2D)	QSAR_no_AD	1	Consensus QSAR model averaging RF/SiRMS and SVM/Dragon predictions without applicability domain filtering
	QSAR_AD	2	Consensus QSAR model with applicability domain filtering
Structure-based (3D)	MEDUSA	3	Molecular docking
Consensus 2D/3D	QSAR_no_AD + MEDUSA	4	Consensus between models 1 and 3
	QSAR_AD + MEDUSA	5	Consensus between models 2 and 3

and structure-based approaches in a consensus way instead of contrasting them. Similar hybrid strategies have been rarely explored previously,[6,7] so we took this benchmarking exercise as an opportunity to test such methodology further.

The main goal of this study was to reliably assess the relative ranking of CSAR ligands by predicting their potency toward given kinase targets. To achieve this goal, we used cheminformatics approaches to (i) collect and curate chemical data extracted from ChEMBL related to binding toward UK and ERK2 and inhibition toward CHK1 and LPXC, (ii) develop statistically robust, validated, and externally predictive QSAR models to compute CSAR ligands' activities and rank them accordingly, and (iii) combine QSAR and structure-based docking predictions into consensus relative ranking lists. The results of our studies suggest that ligand-based QSAR approaches can perform similarly or even better than computationally expensive structure-based approaches. Moreover, we also show potential benefits coming from the synergistic use of both approaches as compared to single method predictions. These benefits mainly relate to the identification and subsequent overriding of activity cliffs (i.e., very similar compounds with dissimilar activities) by enriching the predictions from one structural space (2D or 3D) with the ones from another.

Here, we only report on the results of QSAR modeling and our consensus approach; all results and discussion related to the prediction of CSAR ligand binding poses (and their overall accuracy) by MedusaDock are published in a separate study.[8]

2. METHODS

2.1. Data Set Preparation. *2.1.1. Data Sources.* For each target, we extracted all known associated ligands from the ChEMBL version 12 database.[9] For the UK target (CHEMBL3286), a total of 828 binding affinity (K_i) values were retrieved. Approximately 1450 IC_{50} values were found for CHK1 (CHEMBL4630), whereas only 91 K_i values were retrieved for ERK2 (CHEMBL4040). For all three sets, we did not consider integrating experimental data coming from qHTS assays, mixing IC_{50} and K_i values, or adding data from other sources.

The fourth target of the CSAR benchmark exercise, LPXC, was excluded from our study due to the insufficient amount of data available for QSAR modeling. The LPXC-related set extracted from ChEMBL included 53 compounds with exact IC_{50} values. Among them, only 11 unique compounds had IC_{50} below 1 μM, and the overall distribution of pIC_{50} had a narrow range from 4.2 (inactive compounds) to 6.9 (weakly active compounds) with a strong distribution bias toward inactives.

Thus, QSAR analysis of this data set was not feasible; nevertheless, we have examined whether accurate predictions of LPXC binding affinity for CSAR ligands could be achieved based on global chemical similarity considerations. Indeed, among the 16 new ligands provided by the CSAR organizers, we found a few compounds highly similar to some of the 53 ChEMBL compounds using simple 2D similarity metric (Tanimoto coefficient threshold higher or equal than 0.85). Further examination showed that pairs of highly similar compounds had very similar binding affinities indeed. For instance, only one chlorine group differentiates CSAR_lpxc_11 (pIC_{50} = 4.7) from ChEMBL107127 (pIC_{50} = 4.66). CSAR_lpxc_14 (pIC_{50} = 5.6) is very similar to ChEMBL324440 (pIC_{50} = 5.80), CHEMBL104577 (pIC_{50} = 4.51), CHEMBL104043 (pIC_{50} = 5.92), and CHEMBL107004 (pIC_{50} = 5.26). CSAR_lpxc_11 (pIC_{50} = 4.7) is also very similar to CHEMBL104671 (pIC_{50} = 4.52). This analysis suggests that, even in the absence of sufficiently large amounts of data to enable QSAR modeling, it is still possible to obtain accurate prediction for a subset of LPXC ligands using cheminformatics techniques.

2.1.2. Data Set Curation. The compounds retrieved from the ChEMBL database were preprocessed according to a set of guidelines for chemical data curation and standardization that our group published recently.[10] Briefly, after the removal of counterions, structures were standardized and converted into canonical tautomeric form with neutral representation and explicit hydrogens. As illustrated in Table 2, only 48 out of 91

Table 2. CSAR Targets, Ligands, and Related Compounds Found in ChEMBL Database

target	potency measured as	number of CSAR ligands to rank	number of ChEMBL compounds before curation	number of ChEMBL compounds after curation
extracellular signal-regulated kinase (ERK2)	pK_i	39	91	48
urokinase (UK)	pK_i	20	828	668
checkpoint kinase (CHK1)	pIC_{50}	45	1450	1215

compounds remained in the ERK2 data set after the curation steps including the deletion of stereoisomers and the compounds with uncertain and approximate K_i values. In the end, pK_i values for the 48 selected compounds were ranging from 4.60 to 8.70. Similarly, 668 compounds (out of 828 total) were still present in the UK set after curation, and their pK_i values ranged from 0.30 to 11. Lastly, 1215 out of 1450 compounds remained in the CHK1 data set with pIC_{50} values ranging between 3.68 and 10.

2.2. Molecular Descriptors. *2.2.1. Dragon Descriptors.* The following types of descriptors were generated using Dragon software (v.5.5, Talete SRL, Milan, Italy): 0D-constitutional descriptors (atom and group counts), 1D-functional groups, 1D-atom centered fragments, 2D-topological descriptors, 2D-walk and path counts, 2D-autocorrelations, 2D-connectivity indices, 2D-information indices, 2D-topological charge indices, 2D-eigenvalue-based indices, 2D-topological descriptors, 2D-edge adjacency indices, 2D-Burden eigenvalues, and various molecular properties such as octanol−water partition coefficient. Descriptors with low variance (standard deviation lower than 10^{-4}) or missing values were removed.

Furthermore, if the correlation coefficient between any two descriptors exceeded 95%, one of them was removed. The remaining descriptors were range-scaled, so that their values were within the interval [0, 1]. Definition and calculation procedures for Dragon descriptors and the related references are given in the Handbook of Molecular Descriptors.[11]

2.2.2. SiRMS Descriptors. HiT QSAR software[12] based on the Simplex representation of molecular structure (SiRMS)[13,14] was used for generating 2D Simplex descriptors, i.e., number of tetratomic fragments with fixed composition and topological structure. At the 2D level, the connectivity of atoms in simplex, atom type, and bond nature (single, double, triple, or aromatic) have been considered. SiRMS descriptors account not only for the atom type but also for other atomic characteristics that may impact biological activity of molecules, e.g., partial charge, lipophilicity, refraction, and atom ability for being a donor/acceptor in hydrogen-bond formation (H-bond). For atom characteristics with continuous values (charge, lipophilicity, refraction), the range was converted into several discrete groups. The atoms have been divided into four groups corresponding to their (i) partial charge $A \leq -0.05 < B \leq 0 < C \leq 0.05 < D$, (ii) lipophilicity $A \leq -0.5 < B \leq 0 < C \leq 0.5 < D$, and (iii) refraction $A \leq 1.5 < B \leq 3 < C \leq 8 < D$. For atomic H-bond characteristics, the atoms have been divided into three groups: A (acceptor of hydrogen in H-bond), D (donor of hydrogen in H-bond), and I (indifferent atom). The usage of sundry variants of differentiation of simplex vertexes (atoms) represents the principal feature of the SiRMS approach.[15] A detailed description of HiT QSAR based on SiRMS can be found elsewhere.[12−14] Constant, low-variance, and correlated (|R| ≥ 0.9) descriptors were excluded prior to modeling. Thus, descriptor pools of 435−889 Simplex descriptors (depending on the data set) were selected for the statistical processing.

2.3. QSAR Modeling. In this study, we developed a series of QSAR models following the workflow and other guidelines we published elsewhere.[10,16] The QSAR modeling workflow can be divided into three major steps:[2,16] (i) data curation, preparation, and analysis, (ii) model building, and (iii) model validation/selection. Here, we followed a 5-fold external cross-validation procedure. For each CSAR target, the full set of compounds with known experimental activity was randomly split into five modeling (80% of the full set) and external validation sets (remaining 20%). Models were built using the modeling set compounds only, and it is important to emphasize that the external set compounds were never taken into account to build and/or select the models. Briefly, each modeling set was split into many training and test sets for SVM method and a plethora of training and out-of-bag sets for the RF approach; then the models were built using the compounds belonging to each training set and applied to test set compounds for assessing their properties. Pearson's correlation coefficient (R^2), root mean square error (RMSE), and Spearman's rank correlation coefficient (ρ) were used to assess the prediction performances of developed models.

Best models were identified and selected according to estimated R^2 values for test sets (SVM) or out-of-bag sets (RF). Then, selected models were applied to the external set compounds to predict their experimental properties. This overall procedure was repeated five times to ensure that every compound from the full set is present once (and only once) in the external test set. While compounds are present in the external test sets, they have never been used to derive, bias, or select the models; thus, the entire procedure gives more or less

fair estimation of the true predictivity of the models. In addition, 1000 rounds of Y-randomization were performed for each selected model in order to avoid chance correlations.

The model's applicability domain (AD) aims to determine whether the given model is capable of predicting the activity of a query compound within a reasonable error.[16] In this study, we defined the AD of SVM models as a threshold distance D_T between a query compound and its nearest neighbors in the training set. If the distance of the test compound from any of its k nearest neighbors in the training set exceeds the threshold, the prediction is considered unreliable. For RF models the AD was estimated using the local (Tree) approach that was described by Artemenko et al.[17]

2.4. Random Forest (RF). Random Forest models were constructed according to the original RF algorithm[18] implemented by Polischuk et al.[19] RF is an ensemble of single decision trees. Outputs of all trees are aggregated to obtain one final prediction. Each tree has been grown as follows: (i) A bootstrap sample was produced from the whole set of N compounds to form a training set for the current tree. Compounds that are not in this current tree training set are placed in the out-of-bag (OOB) set (OOB set size is ∼ N/3). (ii) The best split by the CART algorithm[20] among the m randomly selected descriptors from the entire pool in each node is chosen. The value of m is just one tuning parameter for which RF models are sensitive. (iii) Each tree is then grown to the largest possible extent. There is no pruning. Prediction of the out-of-bag set is made, but each tree predicts values only for compounds that are not included in the training set of that tree (for OOB set only). Because RF possesses its own reliable statistical characteristics (based on OOB set prediction) that could be used for validation and model selection, no cross-validation has been performed.[18] Thus, the final model is chosen by the lowest error for prediction of the OOB set and only after that resulting model was applied for blind prediction of external test set/fold compounds.

2.5. Support Vector Machines (SVM). The description of the original SVM algorithm can be found in many publications.[21] Briefly, molecular descriptors are first mapped onto a high dimensional feature space using various kernel functions. Then, SVM finds a separating hyperplane with the maximal margin in this high dimensional space in order to separate compounds with different activities. Models built with this machine learning technique allow the prediction of a target property using a set of descriptors solely calculated from the structure of a given compound. In this study, we used the WinSVM program developed in our group on the basis of the open-source libSVM package. The WinSVM program provides users with a graphical interface to prepare input data; splits data sets into training and test sets, sets up parameters for SVM grid calculations, including iterative and simultaneous grid optimization of SVM parameters; launches and follows calculation progress via a powerful graphical interface; selects models with the best prediction accuracy for both training and internal test sets; and applies them to the external evaluation set as an ensemble consensus model. The program also allows one to visualize molecular structures and various plots, making the use of SVM easier and more appropriate for QSAR modeling in order to obtain robust and predictive models and apply them to virtual libraries. WinSVM is freely available for academic laboratories from the following Web site: http://www.unc.edu/~fourches/.

2.6. MedusaDock. The MedusaDock software[22] was used to generate and score all ligand–receptor binding poses for the different CSAR targets. MedusaDock performs conformational sampling of both ligand and receptor side chains simultaneously and synergistically. Details of the docking method can be found elsewhere.[8,23] Briefly, a library of ligand rotamers is generated in a stochastic manner "on the fly". Ligand conformations are explored by random variation of ligands' rotatable angles and excluding those conformations that feature atomic clashes. The docking protocol involves two steps. First, a representative set of ligand conformations is generated by clustering the stochastic library of rotamers. Each representative conformation is rapidly fitted into a "smoothed" receptor pocket by disabling the van der Waals repulsion between the ligand and the receptor side chains and subsequent rigid-body docking. Second, fine docking is performed from each of the coarsely docked poses, where the binding pose is minimized by iterative repacking of the rotamers of ligand and receptor side chains as well as ligand rigid-body minimization. In the second fine-docking step, the van der Waals repulsions between ligand and receptor side chains are included. The MedusaScore scoring function was utilized to guide the docking.[23]

3. RESULTS AND DISCUSSION

3.1. Presence of CSAR Duplicates in ChEMBL Modeling Set. First, for each target, we used ISIDA/duplicates software to search for structural duplicates between CSAR ligands and the compounds retrieved from the ChEMBL database. We were not expecting to find any duplicate compounds assuming that none of the "blind set" CSAR ligands were supposed to be in the public domain already. Surprisingly, we identified several CSAR ligands that were indeed present in the ChEMBL sets with known experimental affinities for the targets of interest. Results are summarized in Table 3.
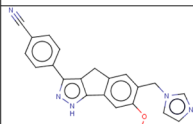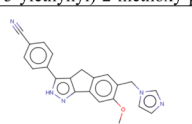
In total, six duplicates were found in the CHK1 data set: 5 out of 6 compounds were CHK1 inhibitors with pIC_{50} ranging from 7.60 (ChEMBL401274) to 8.80 (ChEMBL245796). The sixth compound (ChEMBL396034) was annotated as being inactive (pIC_{50} = 4.77). Only one duplicated structure was identified for both the UK data set (CSAR_UK_18 with ChEMBL319264) and the ERK2 data set (CSAR_ERK2_30 with ChEMBL220320).

When submitting our prediction results to CSAR organizers, we enclosed this list of structural duplicates and a letter underlining that (i) some CSAR compounds and their supposedly unknown experimental activities were indeed present in the public domain and thus could potentially bias the results of the overall benchmark, (ii) simple methods such as similarity search can easily identify them, and (iii) our group honestly acknowledged that we knew the experimental values for those duplicate compounds and thus we advocated for the removal of those compounds from the CSAR ligand set in order to calculate unbiased statistics between the different participants. Although these ligands were present in the training sets we used to develop models, our group submitted only the values obtained from the 5-fold external cross-validation when these compounds were blindly predicted.

Later when the experimental activities of all CSAR ligands were revealed, we indeed observed their perfect agreement with the values retrieved from the ChEMBL database (Table 3).

3.2. Prediction Performance of QSAR Models. QSAR modeling results are given in Table 4. Models built using the

Table 3. List of CSAR Compounds with Identified Structural Duplicates Retrieved in the ChEMBL Database and Their Reported Activities

CSAR compounds (experimental activity)	ChEMBL compounds identical to CSAR compounds
CSAR_chk1_1 *(CSAR pIC$_{50}$=7.60)*	**CHEMBL401274; pIC$_{50}$=7.60** 4-(6,7-dimethoxy-2,4-dihydro-indeno[1,2-c]pyrazol-3-yl)-phenol;
CSAR_chk1_3 *(CSAR pIC$_{50}$=8.30)*	**CHEMBL248396; pIC$_{50}$=8.30** 4-(6,7-dimethoxy-2,4-dihydro-indeno[1,2-c]pyrazol-3-ylethynyl)-2-methoxy-phenol;
CSAR_chk1_4 *(CSAR pIC$_{50}$ = Not Reported)*	**CHEMBL247396; pIC$_{50}$=8.70** 4-(6-imidazol-1-ylmethyl-7-methoxy-2,4-dihydro-indeno[1,2-c]pyrazol-3-yl)-benzonitrile;
CSAR_chk1_6 *(CSAR pIC$_{50}$=8.80)*	**CHEMBL245796; pIC$_{50}$=8.80** 4'-(6,7-dimethoxy-1,4-dihydro-indeno[1,2-c]pyrazol-3-yl)-biphenyl-4-ol;
CSAR_chk1_13 *(CSAR pIC$_{50}$=7.64)*	**CHEMBL248010; pIC$_{50}$=7.64** 4'-{6-[2-(5-ethyl-pyridin-2-yl)-ethoxy]-7-methoxy-2,4-dihydro-indeno[1,2-c]pyrazol-3-yl}-biphenyl-4-ol;
CSAR_chk1_20 *(CSAR pIC50=4.80)*	**CHEMBL396034; pIC$_{50}$=4.77** 8-chloro-5,10-dihydro-dibenzo[b,e][1,4]diazepin-11-one;
CSAR_uk_18 *(CSAR pKi=6.30)*	**CHEMBL319264; pKi=6.35** 8-Amino-naphthalene-2-carboxamidine;
CSAR_erk2_30 *(CSAR pKi=7.10)*	**CHEMBL220320; pKi=7.07** N-benzyl-4-(4-(3-chlorophenyl)-1H-pyrazol-3-yl)-1H-pyrrole-2-carboxamide;

SiRMS fragment descriptors and Random Forests afforded reasonable prediction performances evaluated by Spearman rank correlation ρ ranging from 0.78 (CHK1) to 0.85 (UK). When considering models' applicability domains, the reliability of RF predictions increased (up to $\rho = 0.89$ for UK) but the coverage decreased, i.e., about 25% of the compounds had to be excluded due to the model applicability domain.

The prediction power of SVM models based on Dragon descriptors was slightly lower than that of RF models with ρ ranging from 0.77 (CHK1) to 0.84 (UK). In particular, ERK2 predictions were less accurate with R^2 going down from 0.71 (RF) to 0.62 (SVM). With applicability domain, ranking accuracy of SVM models ranged from $\rho = 0.72$ (ERK2) to 0.86 (UK).

In addition to the individual RF and SVM models, we also explored the predictive power of the simple consensus prediction where activities for external compounds were predicted as simple arithmetic means of predictions made with RF and SVM models. Obtained results showed that in most cases, with or without taking into account the models' applicability domain, the consensus model was consistently achieving higher reliability compared to any of the individual QSAR models. For instance, the modeling results obtained for UK were as follows. Without applicability domain filtering, RF models afforded very good performance ($\rho = 0.85$, $R^2 = 0.69$) as well as SVM models ($\rho = 0.84$, $R^2 = 0.68$). The consensus model improved the accuracy reaching up to $\rho = 0.87$ and even $\rho = 0.88$ taking into account the applicability domain. Importantly, the coverage of the consensus is significantly boosted from 71% to 75% (individual SVM and RF models) up to 88%. This result means that more compounds were predicted correctly compared to individual QSAR models.

3.3. Application of QSAR Models to CSAR Ligands. The results of activity prediction for CSAR ligands are given in

Table 4. Statistical Characteristics of QSAR Models for CSAR Data Sets Assessed by 5-fold External Validation

method	descriptor	target	no AD			with AD			
			R^2	RMSE	Spearman ρ	R^2	RMSE	Spearman ρ	coverage
RF	SiRMS	CHK1 ($n = 1215$)	0.64	0.77	0.78	0.72	0.66	0.85	75%
SVM	Dragon	CHK1 ($n = 1215$)	0.64	0.76	0.77	0.65	0.73	0.81	72%
QSAR_CONSENSUS		CHK1 ($n = 1215$)	0.67	0.74	0.79	0.68	0.71	0.83	87%
RF	SiRMS	ERK2 ($n = 48$)	0.71	0.62	0.80	0.69	0.56	0.79	75%
SVM	Dragon	ERK2 ($n = 48$)	0.62	0.68	0.77	0.59	0.65	0.72	73%
QSAR_CONSENSUS		ERK2 ($n = 48$)	0.69	0.62	0.79	0.67	0.59	0.76	90%
RF	SiRMS	UK ($n = 668$)	0.69	0.66	0.85	0.77	0.53	0.89	75%
SVM	Dragon	UK ($n = 668$)	0.68	0.72	0.84	0.70	0.64	0.86	71%
QSAR_CONSENSUS		UK ($n = 668$)	0.71	0.68	0.87	0.73	0.61	0.88	88%

Table 5. To match the ranking metric used by the organizers, ranking performance was evaluated using the Spearman

Table 5. Spearman Correlation Coefficient (ρ) between Experimental and Predicted Ranks of CSAR Ligands

type	ranking methods	ρ		
		UK ($n = 20$)	ERK2 ($n = 39$)	CHK1 ($n = 45$)
2D	QSAR_no_AD	0.77	0.60	0.55
	QSAR_AD	0.78	0.59	0.55
3D	MEDUSA	0.76	0.31	0.26
2D/3D	QSAR_no_AD + MEDUSA	0.79	0.48	0.45
	QSAR_AD + MEDUSA	0.82	0.50	0.45

correlation coefficient ρ expressing the ranking accuracy of ligands by comparing the ligands' rank orders based on model's predicted potency (pKi or pIC$_{50}$ depending on the target) with the actual experimental rank provided by the CSAR organizers.[1] As discussed below, model predictive accuracy was evaluated both for all ligands as was stipulated by the CSAR challenge organizers as well for ligands found within the AD of QSAR models only to follow our standard modeling workflow (see Methods).

When predicting all ligands in the absence of any AD, the QSAR models afforded relatively high accuracy for UK ligands ($\rho = 0.77$, $n = 20$) and lower accuracies for ERK2 ($\rho = 0.60$, $n = 39$) and CHK1 ($\rho = 0.55$, $n = 45$) ligands (Table 5; QSAR_no_AD). Independently, we have employed an ad hoc scheme to make predictions for all compounds when using the respective AD thresholds for individual RF and SVM models. Thus, for compounds found either within or outside of the AD of the individual models, the predicted activities were averaged, whereas for compounds found within the AD of only one model, the activity predicted by that model was used. As shown in Table 5, the prediction accuracy for this QSAR_AD model was similar to that for the QSAR_no_AD model.

In addition, we also made predictions for ligands within the models' AD only, i.e., with reduced coverage of the CSAR data sets. Indeed, many CSAR ligands were found to be outside of the respective AD of either SVM or RF models. Certain ligands were even outside of the AD of both models: 14 compounds for UK, 8 compounds for ERK2, and 13 compounds for CHK1. Only six out of 20 UK ligands were found to be within the AD

of one of the models making it nonsensible to evaluate model prediction accuracy in this case. After removing compounds outside of the AD, the Spearman correlation coefficients between experimental and predicted ranks (QSAR_AD model) for the remaining CSAR compounds increased to 0.64 for both ERK2 ($n = 31$, coverage = 79.5%) and CHK1 ($n = 32$, coverage = 71.1%) data sets as compared to 0.59 and 0.55, respectively, when all compounds were considered (Table 5). Thus, the effect of AD on prediction accuracy of QSAR models is data set dependent. In one of the considered cases (UK), the default AD appears over-restrictive, whereas in two other cases, the use of AD slightly improves model accuracy but at the expense of reduced data coverage, which is typical for QSAR-based predictions.

Second, we analyzed the results obtained by using the MedusaDock scores for ranking the CSAR ligands. Medusa-Dock was as accurate as QSAR models for UK. However, the QSAR models were found to have twice as high predictive power than MedusaDock for ERK2 ($\rho = 0.60$ versus $\rho = 0.31$) and CHK1 ($\rho = 0.55$ versus $\rho = 0.26$).

Overall, we could make the following observations: (i) The "true" accuracy of QSAR models and MedusaDock for ranking CSAR ligands is slightly (UK) or significantly (CHK1) worse than the one found at the modeling and validation stages. (ii) Ligand-based QSAR models performed better than computationally expensive molecular docking. (iii) The QSAR models' applicability domains in their current form do not significantly improve the overall prediction accuracy for the remaining compounds.

3.4. Consensus Scoring Using QSAR Predictions and MedusaDock. As part of the exercise, we considered another type of consensus model including the predictions coming from both QSAR models and Medusa docking. Ranks for CSAR ligands predicted by the QSAR models (e.g., QSAR_no_AD) were added to the ranks predicted by Medusa docking. Then, the ligands were reranked based on these summed QSAR/Medusa ranks. The overall results are shown in Table 5.

The accuracy of QSAR/Medusa consensus predictions was higher than the accuracy reached by Medusa predictions for all three targets. This remark is particularly true for CHK1 and ERK2 for which the QSAR/Medusa consensus model was found to be almost twice more predictive than Medusa alone; in the case of ERK2 for instance, $\rho = 0.48$ compared to $\rho = 0.31$, respectively.

Also we noticed that the QSAR/Medusa consensus predictions afforded higher ranking accuracy than individual QSAR models and MedusaDock only in the case of UK ($\rho =$

0.79–0.82 versus ρ = 0.76–0.78). Because of the higher ranking accuracy obtained by QSAR models over Medusa for both ERK2 and CHK1, this result was expected.

3.5. Half Success or Half Failure? The analysis of the results revealed the overall reliability of QSAR models to rank CSAR ligands from the most active to the most inactive, especially for UK. However, there is a significant portion of ligands that have been mispredicted by both QSAR and docking. Among them, some compounds predicted to be active were confirmed as being weak active or inactive. In this section, we are giving some examples and some clues to improve our current approach based on what we learned in this exercise.

First, the results tend to contradict the general principle commonly trusted by the QSAR community posing that the bigger the modeling set is, the more predictive the model will be. In this CSAR benchmark, our largest modeling set included 1215 compounds for the CHK1 target. Although QSAR models developed using this large data set afforded reasonable prediction power in the 5-fold external cross-validation procedure (Table 4), the set of 45 CSAR ligands tested toward CHK1 was the most difficult to annotate as shown by the results: Spearman ρ = 0.55 for QSAR models as compared to ρ = 0.59–0.78 for UK and ERK2; ρ = 0.26 for docking as compared to ρ = 0.31–0.76 for UK and ERK2. Besides the ranking accuracy per se, the consensus QSAR model was indeed able to correctly predict 17 out of 21 CHK1 actives and 10 out of 24 inactive compounds but missed 14 false positives and 4 false negatives (sensitivity = 0.81, specificity = 0.42, and balanced accuracy = 0.61, considering the activity threshold of pIC_{50} = 7). Out of these 18 mispredicted compounds, we should underline that eight compounds have their experimental pIC_{50} ranging from 6.5 to 7.7, which is very close to the activity threshold we used to separate active from inactive compounds.

The smallest modeling set (ERK2) included 48 compounds only. Nevertheless, QSAR models built for this small modeling set afforded relatively good prediction performance at the 5-fold cross-validation stage (R^2 = 0.69, RMSE = 0.62, and ρ = 0.79) and reasonable reliability on CSAR ligands (ρ = 0.60). As illustrated in Figure 1, the balanced accuracy is reaching 0.77 for the 39 CSAR ligands tested toward ERK2 using an affinity threshold of pK_i = 7 to distinguish active from inactive
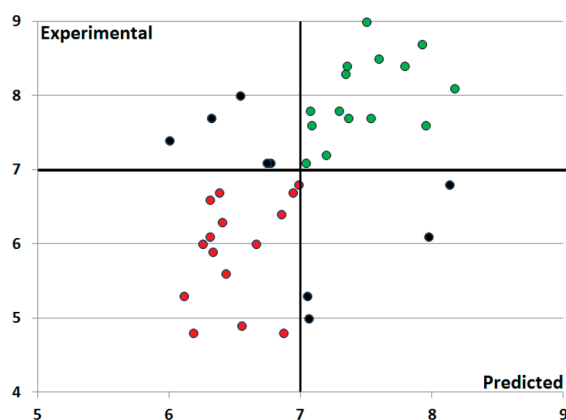

Figure 1. Experimental versus predicted binding affinities (pK_i) for ERK2 CSAR ligands (n = 39) based on QSAR_no_AD model's predictions. Correctly predicted ERK2 binders ($pK_i \geq 7$) are colored in green, whereas correctly predicted nonbinders are colored in red (balanced accuracy = 0.77). Mispredicted compounds are colored in black.

compounds. These results demonstrate once again the importance of the cross-validation procedures in the QSAR modeling workflow but also the fact that such procedures must involve the building and selection of QSAR models using the modeling sets only and a truly external validation with the test sets.

Second, the overall accuracy of 2D QSAR models was affected by the presence of large activity cliffs in both the modeling and external sets of ligands. To illustrate this point, let us consider again the example of ERK2 ligands and more precisely the CSAR_ERK2_1 compound. As shown in Figure 2, very similar structures to CSAR_ERK2_1 found in our modeling set are annotated as strong binders with pK_i equal to 8.4 and above. It is thus not surprising that our QSAR models computed CSAR_ERK2_1's affinity toward ERK2 to be approximately pK_i = 6.9. However, the experimental binding affinity has been determined to be 4.8 only. This perfectly corresponds to the case of activity cliffs.[24]

In Figure 2, we showed CSAR_ERK2_1 as well as two other compounds, CSAR_ERK2_6 and CSAR_ERK2_9, that have not been assessed correctly by our QSAR models. Despite the fact that the QSAR model succeeded to correctly predict the increasing activity trend CSAR_ERK2_1 (pK_i = 4.8) < CSAR_ERK2_6 (pK_i = 6.1) < CSAR_ERK2_9 (pK_i = 6.8), the model did calculate their binding affinities with $\Delta pK_{i,\text{pred-exp}}$ > 1.5 log units. Interestingly, the docking score obtained for CSAR_ERK2_1 is relatively high (−42.2), meaning that the binding of the compound is predicted to be unfavorable. To provide the necessary context, MedusaDock scores ranged from −59.2 (very favorable docking) to −36.4 (unfavorable docking) for the ERK2 CSAR ligands. As a result, this observation opens the way for defining new strategies to calculate consensus predictions between QSAR models and docking scores as well as identifying potential activity cliffs such as CSAR_ERK2_1. Simply summing up the ranks from QSAR models and docking scores as we did in this exercise does not seem to be the optimal workflow. On the contrary, using docking score thresholds to automatically discard some compounds from the predicted actives is more likely to avoid the prediction of false-positives such as CSAR_ERK2_1.

Third, based on the results presented in this study, there are some additional evidence how to complement structure-based predictions from ligand-based predictions. In Figures 3 and 4, we plotted the MedusaDock scores versus QSAR_no_AD predictions. The 2D/3D correlation reached R^2 = 0.67 for UK and only 0.42 for CHK1. These values are indeed important to analyze because they measure the level of concordance between the two different modeling approaches for the CSAR ligands and can be computed without the knowledge of the experimental values of the compounds. The challenge is thus to find new ways to use these correlation plots for establishing rules to calculate a new type of 2D/3D consensus. Also, it seems logical that one way to assess the potential benefit of the 2D/3D consensus requires the calculation of their correlation coefficient for modeling set compounds (and thus there is a need for docking the modeling set compounds as well).

Fourth, compared to the other research teams who participated in the CSAR benchmarking, the reasonable prediction performances obtained by our QSAR models ranked our group in the top 10%. Moreover, our QSAR models occupied top two and top three positions for ranking both CHK1 actives and inactives, respectively. Our models were ranked fifth for ERK 2 prediction reliability. We have processed

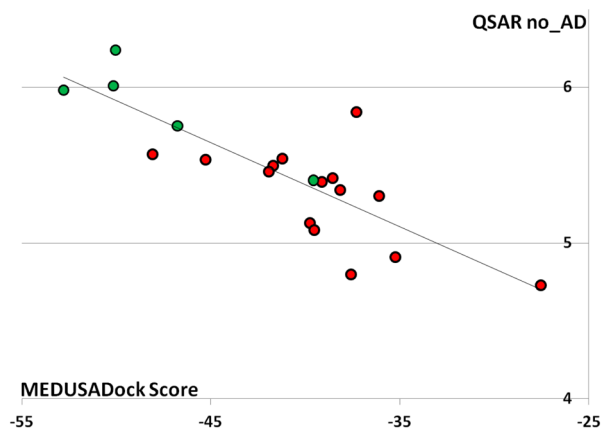**Figure 2.** Structural neighbors of the CSAR_ERK2_1 compound retrieved in the ChEMBL database.

Figure 3. MedusaDock scores plotted versus QSAR_no_AD pKi predictions for the 20 CSAR ligands toward UK ($R^2 = 0.67$). UK binders (pKi $\geq$ 7) are colored in green, whereas nonbinders are colored in red.
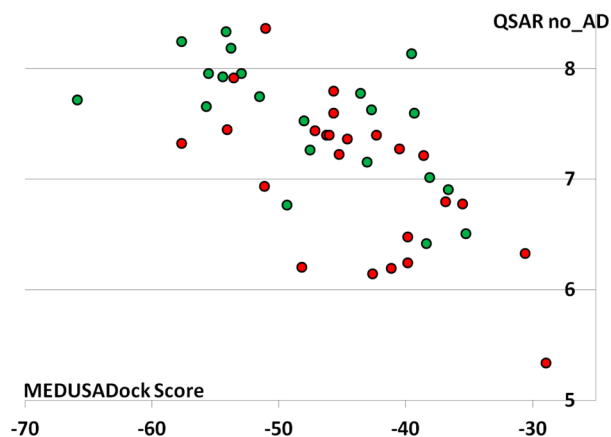

Figure 4. MedusaDock scores plotted versus QSAR_no_AD pKi predictions for the 45 CSAR ligands toward CHK1 ($R^2 = 0.42$). CHK1 inhibitors (pIC$_{50}$ $\geq$ 7) are colored in green, whereas inactive compounds are colored in red.

only three targets, and the overall performance among all the targets cannot be estimated for our models. However, based on the results for separate targets, we can expect that our group was ranked among top three research teams.

Overall, structure-based approaches should not be viewed as intuitively better or more predictive than ligand-based QSAR models, and this CSAR benchmarking exercise serves to illustrate this point. It is well known that correlation between docking scoring functions and experimental binding affinities is typically low[25] or moderate.[26] Furthermore, as shown by our collaborators at UNC for the same CSAR sets,[8] structure-based approaches (and MedusaDock especially) are accurate in generating native-like poses. However, as this study shows, the docking scores for those native-like poses do not correlate with experimental binding affinities well (Table 5) and thus do not allow a correct ranking. This observation highlights a known fact that different scoring functions are needed for predicting ligand poses versus predicting binding affinities. Lastly, we should stress that unlike universal scoring functions used in docking studies, QSAR models are specifically trained and selected toward a given target using a set of respective ligands with experimental activities. Thus, it may be under-appreciated but not necessarily surprising that ligand-based QSAR models can, in fact, have better accuracy than most of the structure-based docking approaches in prognosticating target-specific ligand binding affinities.

4. CONCLUSIONS

In this study, structure-based (molecular docking) and ligand-based (QSAR models) approaches were used both independently and in the form of a 2D/3D consensus model to rank untested ligands based on their predicted potency. In this exercise of blind predictions, QSAR models developed with publicly available experimental data extracted from the ChEMBL database were shown to outperform predictions obtained by several molecular docking approaches. These results confirmed that when QSAR models are rigorously derived using curated chemical data sets and statistically relevant procedures for model selection and validation then their prediction power can be at least as accurate as computationally expensive structure-based docking. Our results

also emphasized the validity of QSAR models as a critical component of a virtual screening platform. Moreover, we showed the potential benefits of using both QSAR and docking predictions altogether to assess and eventually override the presence of activity cliffs in the sets of ligands. However, in this particular CSAR benchmark, we did not notice a dramatic boost in predictions' accuracy using the current implementation of our QSAR/docking consensus model. We believe the CSAR community exercise represents a great initiative to honestly benchmark (i) structure-based scoring functions and docking software with each other as well as with (ii) ligand-based cheminformatics methods, whose prediction accuracy will continue to rise along with the increasing number of experimental data available in online repositories.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: alex_tropsha@unc.edu.

Notes
The authors declare no competing financial interest.

■ REFERENCES

(1) Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B.; Stuckey, J. A.; Carlson, H. A. *J. Chem. Inf. Model.* **2013**, DOI: 10.1021/ci400025f.

(2) Tropsha, A. *Mol. Inf.* **2010**, *29*, 476−488.

(3) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755−63.

(4) Chen, Y.; Shoichet, B. K. *Nat. Chem. Biol.* **2009**, *5*, 358−64.

(5) Jain, S. V; Ghate, M.; Bhadoriya, K. S.; Bari, S. B.; Chaudhari, A.; Borse, J. S. *Org. Med. Chem. Lett.* **2012**, *2*, 22.

(6) Hsieh, J.-H.; Yin, S.; Liu, S.; Sedykh, A.; Dokholyan, N. V; Tropsha, A. *J. Chem. Inf. Model.* **2011**, *51*, 2027−35.

(7) Scotti, L.; Mendonca, F. B. M.; Moreira, D. R. M.; Sobral da Silva, M.; Pitta, I. R.; Scotti, M. T. *Curr. Top. Med. Chem.* **2012**, *12*, 2785−809.

(8) Ding, F.; Dokholyan, N. V. *J. Chem. Inf. Model.* **2012**, DOI: 10.1021/ci300478y.

(9) ChEMBL Database https://www.ebi.ac.uk/chembl/ (accessed Mach 13, 2013).

(10) Fourches, D.; Muratov, E.; Tropsha, A. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.

(11) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Todeschini, R., Consonni, V., Eds.; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000; Vol. *11*, p 667.

(12) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 403−421.

(13) Muratov, E. N.; Artemenko, A. G.; Varlamova, E. V; Polischuk, P. G.; Lozitsky, V. P.; Fedchuk, A. S.; Lozitska, R. L.; Gridina, T. L.; Koroleva, L. S.; Sil'nikov, V. N.; Galabov, A. S.; Makarov, V. a; Riabova, O. B.; Wutzler, P.; Schmidtke, M.; Kuz'min, V. E. *Future Med. Chem.* **2010**, *2*, 1205−1226.

(14) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N.; Volineckaya, I. L.; Makarov, V. A; Riabova, O. B.; Wutzler, P.; Schmidtke, M. *J. Med. Chem.* **2007**, *50*, 4205−4213.

(15) Artemenko, A.; Muratov, E.; Kuz'min, V.; Kovdienko, N.; Hromov, A.; Makarov, V.; Riabova, O.; Wutzler, P.; Schmidtke, M. *J. Antimicrob. Chemother.* **2007**, *60*, 68−77.

(16) Tropsha, A.; Golbraikh, A. *Curr. Pharm. Des.* **2007**, *13*, 3494−3504.

(17) Artemenko, A. G.; Muratov, E. N.; Kuz'min, V. E.; Muratov, N. N.; Varlamova, E. V; Kuz'mina, A. V; Gorb, L. G.; Golius, A.; Hill, F. C.; Leszczynski, J.; Tropsha, A. *SAR QSAR Environ. Res.* **2011**, *22*, 575−601.

(18) Breiman, L. *Mach. Learn.* **2001**, *45*, 5−32.

(19) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E. *J. Chem. Inf. Model.* **2009**, *49*, 2481−8.

(20) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth Publishing: Belmont, 1984; p 358.

(21) Vapnik, V. N. *The Nature of Statistical Learning Theory.*; Springer-Verlag: New York, 1995.

(22) Ding, F.; Yin, S.; Dokholyan, N. V *J. Chem. Inf. Model.* **2010**, *50*, 1623−1632.

(23) Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V *J. Chem. Inf. Model.* **2008**, *48*, 1656−1662.

(24) Maggiora, G. M. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(25) 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. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(26) Biesiada, J.; Porollo, A.; Velayutham, P.; Kouril, M.; Meller, J. *Hum. Genomics* **2011**, *5*, 497−505.