JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Multi-Step Protocol for Automatic Evaluation of Docking Results Based on Machine Learning Methods—A Case Study of Serotonin Receptors 5-HT$_6$ and 5-HT$_7$

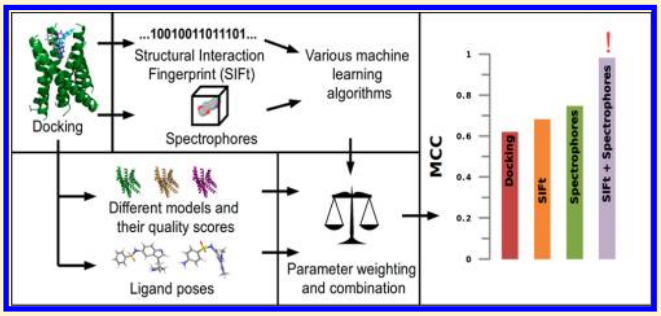
Sabina Smusz,[†,‡] Stefan Mordalski,[†] Jagna Witek,[†] Krzysztof Rataj,[†] Rafał Kafel,[†] and Andrzej J. Bojarski*[,†]

[†]Department of Medicinal Chemistry, Institute of Pharmacology, Polish Academy of Sciences, 12 Smętna Street, 31-343 Kraków, Poland

[‡]Faculty of Chemistry, Jagiellonian University, 3 Ingardena Street, 30-060 Kraków, Poland

Supporting Information

ABSTRACT: Molecular docking, despite its undeniable usefulness in computer-aided drug design protocols and the increasing sophistication of tools used in the prediction of ligand−protein interaction energies, is still connected with a problem of effective results analysis. In this study, a novel protocol for the automatic evaluation of numerous docking results is presented, being a combination of Structural Interaction Fingerprints and Spectrophores descriptors, machine-learning techniques, and multi-step results analysis. Such an approach takes into consideration the performance of a particular learning algorithm (five machine learning methods were applied), the performance of the docking algorithm itself, the variety of conformations returned from the docking experiment, and the receptor structure (homology models were constructed on five different templates). Evaluation using compounds active toward 5-HT$_6$ and 5-HT$_7$ receptors, as well as additional analysis carried out for beta-2 adrenergic receptor ligands, proved that the methodology is a viable tool for supporting virtual screening protocols, enabling proper discrimination between active and inactive compounds.



INTRODUCTION

Molecular docking is a widely used method for predicting the conformation of a chemical compound within the binding cleft of a receptor. The speed and relatively high accuracy[1] of this method has made it a crucial step in many hierarchical virtual screening (VS) protocols.[2] Although scoring functions, both provided with the docking software and developed independently,[3−8] show acceptable predictions of the energetic component of the ligand binding, the biological meaning of output docking conformations and crucial interactions have to be assessed either by visual inspection or other tools.

The need for additional analysis of post-docking results becomes troublesome in regard to processing large amounts of data, such as those generated in a screening experiment, and reliable tools are therefore needed to support the filtering of the active compounds from the numerous predicted ligand−receptor complexes. Recent advances in data fusion approaches combined with machine-learning (ML) techniques show great potential of post-docking analysis but on the other hand still leave room for improvement.[9−11]

We propose a multi-step protocol that exploits ML models for automatic evaluation of docking results and verify its performance using the serotonin receptors 5-HT$_6$[12] and 5-HT$_7$.[13] The target choice was dictated by the scope of research conducted in the Institute of Pharmacology Polish Academy of Sciences (IP PAS) in order to provide the reliability of the results and assurance that the results are not prone to errors obtained in the previous stages, such as homology models construction. The method incorporates Structural Interaction Fingerprints (SIFts), Spectrophores descriptors, and a multi-step results analysis, taking into account the quality of the receptor model, various conformations of the docked compound, and the performance of the particular classification algorithm. It was proved that such an approach is a successful way of post-processing of the docking results, and compared to raw docking score, the proposed multi-step protocol provides improvement of discrimination between active and inactive compounds.

METHODOLOGY

Construction of Homology Models of 5-HT$_6$ and 5-HT$_7$. The structures of the 5-HT$_6$ and 5-HT$_7$ receptors were developed by means of homology modeling. Ten crystal structures of class A G protein-coupled receptors (GPCRs), acquired from the PDB repository, were used as templates: adenosine A$_{2A}$ receptor,[14] adrenergic receptor beta1,[15] adrenergic receptor beta2,[16] CXC chemokine receptor type

Received: September 16, 2014
Published: March 25, 2015

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823−832

4,[17] dopamine 3 receptor,[18] histamine 1 receptor,[19] muscarinic receptors M_2,[20] M_3,[21] and serotonin receptors 5-HT_{1B}[22] and 5-HT_{2B}[23] (a detailed list of all crystal structures used for modeling is provided in Table S1 of the Supporting Information). The sequences of the human 5-HT_6 and 5-HT_7 receptors were obtained from the UniProtKB/Swiss-Prot database.[24] The sequence alignments were performed manually by means of Discovery Studio[25] (only transmembrane helices; loops were not considered). Two hundred models were generated for each aligned template, using Modeller 9v8.[26,27] The obtained structures were validated with a two-step docking protocol, as described by Rataj et al.[28] Compounds for docking were fetched from the ChEMBL database,[29] and their three-dimensional conformations were generated with the use of LigPrep.[30]

For each individual template, the model providing the best discrimination between active (compounds with K_i < 100 nM) and inactive molecules (compounds with K_i > 1000 nM) was selected on the basis of the Area Under Receiver Operating Characteristic (AUROC) values (ROC graphs were constructed by plotting the true positive rate against the false positive rate for the subsequent cutoff levels of Glide Score).[31] Graphs were generated only for models that successfully passed through the first step of evaluation; those for selected models are shown in Figure S1 of the Supporting Information.[28] AUROC values for all the models are gathered in Table 1.

Table 1. AUROC Values Characterizing the Best Models Obtained for a Particular Template

receptor	template	AUROC value
5-HT_6	**beta2**	**0.730**[a]
	beta1	–[b]
	A_{2A}	**0.693**
	D_3	**0.689**
	H_1	0.605
	M_2	0.639
	M_3	**0.661**
	CXCR4	**0.718**
	5-HT_{1B}	0.499
	5-HT_{2B}	–
5-HT_7	**beta2**	**0.757**
	beta1	**0.786**
	A_{2A}	0.709
	D_3	**0.764**
	H_1	**0.828**
	M_2	0.717
	M_3	**0.749**
	CXCR4	0.669
	5-HT_{1B}	0.441
	5-HT_{2B}	–

[a]Bold values indicate five models selected for further study with the highest AUROC. [b]'–' indicates that the model did not pass the first step of evaluation.

Out of the set of best models for particular templates, the five models with the highest AUROC values were selected for further study (presented in bold in Table 1).

Preparation of Sets of Compounds. The ChEMBL database was the source of compounds for the research. All compounds with experimentally verified activity toward 5-HT_6R and 5-HT_7R were fetched from ChEMBL; however, only the ones tested in assays on human- or rat-cloned or native receptors and with activity quantified in K_i or IC_{50} were taken into account. In addition, sets of compounds with assumed inactivity were generated from the ZINC database[32] for each of the considered targets based on the Directory of Useful Decoys (DUD) methodology.[33] For each chemical structure used, the following descriptors were calculated using tools provided by ChemAxon:[34] molecular weight (MW), number of hydrogen bond acceptors (HBA), number of hydrogen bond donors (HBD), number of rotatable bonds (rotB), and logP. For each active ligand, structures with the same number of HBA, HBD, and rotB as well as those with logP and MW values differing by no more than 10% were selected out of the ZINC database using an in-house script. Then, the Daylight-type fingerprints were calculated by means of RDKit software,[35] and the set of compounds was restricted to those with Tanimoto coefficient values of less than 0.7 for a particular ligand (according to the assumption that DUDs should resemble active compounds in terms of physicochemical properties but bear different topology at the same time). Thirty-six decoys per active molecule (with the lowest values of the Tanimoto coefficient) were picked, and as the set of DUDs for both targets exceeded 30,000 compounds, 2000 examples were randomly selected out of this original set for the purpose of the study. The three-dimensional conformations for all the compounds (actives, inactives, and DUDs) were generated by LigPrep, and they were all docked into the selected homology models (five for each of the targets) with a maximum of five conformations allowed for a given ligand. Because incriminated receptors belong to the aminergic subfamily of class A GPCRs, two other physicochemical features were examined: logD and pK_a. Their distribution for all sets of compounds is visualized on the histograms in Figure S2 of the Supporting Information.

Representation of Docking Results. For all ligand–receptor complexes obtained from the docking procedure, SIFts[36] and Spectrophores[37] were calculated. The former ones were generated with the use of the Schrödinger software package[38] and adjusted to the appropriate format with the use of an in-house script, whereas the Spectrophores descriptors were calculated using tools from the Open Babel package.[39]

SIFts are binary fingerprints representing interactions between a docked ligand and a protein. They are generated for each compound separately and are divided into chunks that characterize the contacts between a compound and particular amino acids of the target. In this paper, nine bits are used to describe those associations (any, backbone, side chain, polar, aromatic, hydrophobic, H-donor, H-acceptor, charged).

Spectrophores provide information about the three-dimensional conformation of a molecule through a set of atomic properties—atomic partial charges, atomic lipophilicity, atomic shape deviations, and atomic electrophilicity. They are generated by surrounding a given compound with a set of points, which serve as anchors for the interactions with the compounds' atoms calculated in terms of the above-mentioned set of properties (the total interaction between the artificial points and the molecule for a given property is minimized).

An exemplary calculation of SIFts and Spectrophores for the selected compound is presented in Figure S3 of the Supporting Information.

Machine-Learning Experiments and Analysis of the Results. The ML experiments were performed with the use of WEKA package[40] for five different classification algorithms: Naïve Bayes (NB),[41] Sequential Minimal Optimization (SMO),[42] k-nearest neighbor algorithm (IBk),[43] decision tree

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823–832

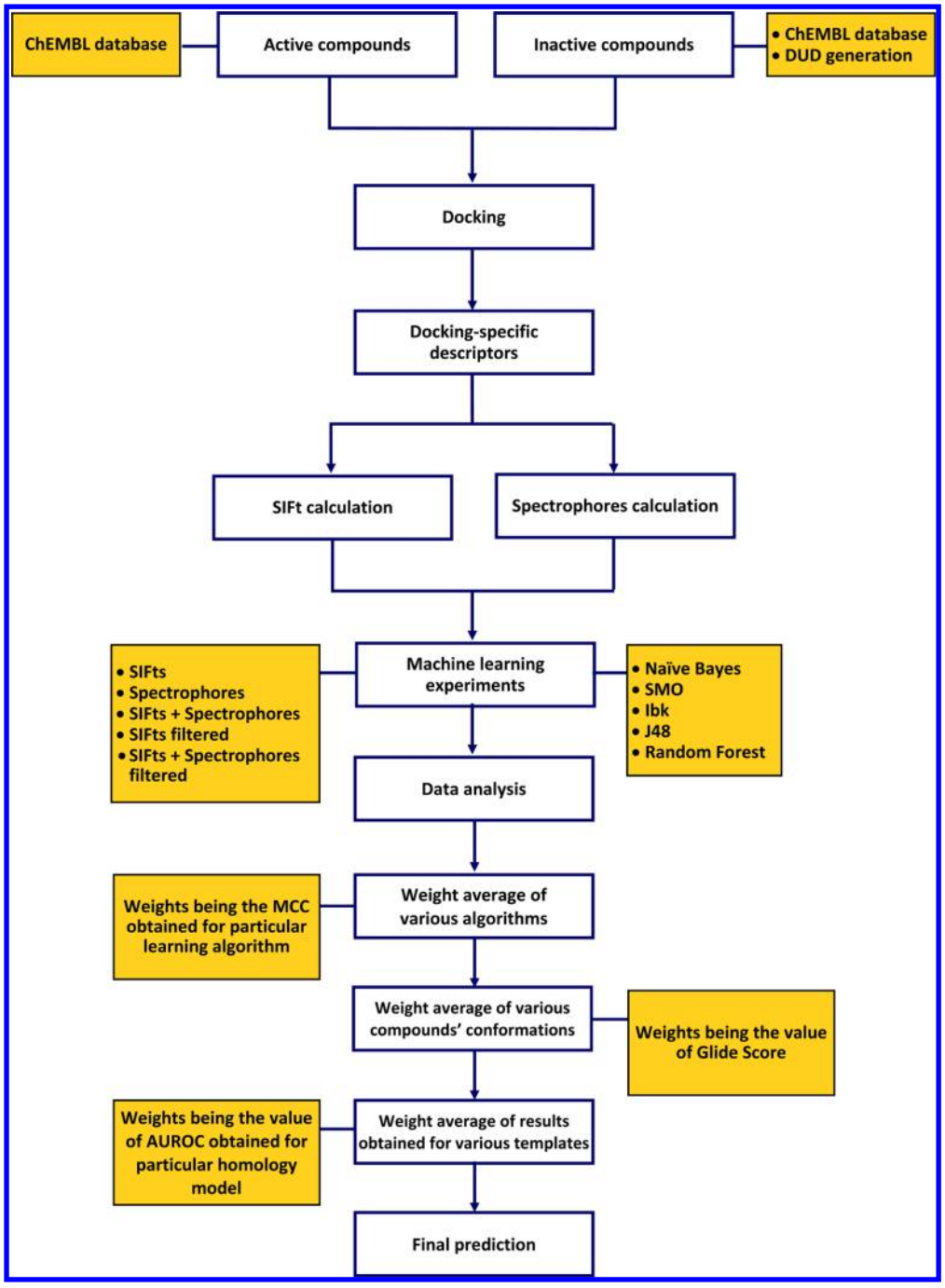


Figure 1. Scheme of the multi-step protocol for post-docking analysis.

J48[44] and Random Forest (RF),[45,46] with SIFts and Spectrophores constituting input descriptors. The calculations were performed for each representation type separately and also for their concatenation. Because SIFts provide relatively long descriptions of compounds, calculations (with the use of SIFts both individually and in combination with Spectrophores) were carried out twice—once for the original fingerprints and once after data preprocessing (Attribute filtering by Correlation-based Feature Subset Selection[47] was carried out using genetic algorithm[48] as a searching method). The ML experiments were performed in the 3-fold and 5-fold cross-validation (CV) mode.

In order to verify the ability of the model to make a discrimination between active and inactive compounds that were not included in the training set, the clustering procedure was carried out. All sets of compounds were hierarchically clustered in Canvas[49] with MOLPRINT2D[50,51] for compounds representation and Buser similarity. After the examination of the number and composition of clusters obtained according to the Kelley criterion, the number of clusters was fixed to 20. The experiments were performed in the leave-one-cluster out approach (each cluster was once excluded from the training set and formed the test set).

For the sake of proper data consolidation and analysis, the experiments were conducted in several consecutive steps. First, each ligand−receptor complex was considered as a separate instance; therefore, the calculation of evaluating parameters at this stage could be used for the assessment of ML algorithms.

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823−832

Table 2. Docking Statistics for 5-HT$_6$ and 5-HT$_7$ Homology Models

	initial number of compounds			number of compounds after Ligprep				number of docked compounds		
target	actives	true inactives	DUDs	actives	true inactives	DUDs	template	actives	true inactives	DUDs
5-HT$_6$	1388	320	2000	2545	597	3002	beta2	2127	415	2153
							CXC	2441	519	2636
							A_{2A}	2136	424	2193
							D_3	1801	332	1624
							M_3	2488	545	2752
5-HT$_7$	624	293	2000	1239	589	2589	H_1	910	423	1876
							beta1	907	415	1762
							D_3	822	402	1712
							beta2	787	367	1490
							M_3	963	443	1917

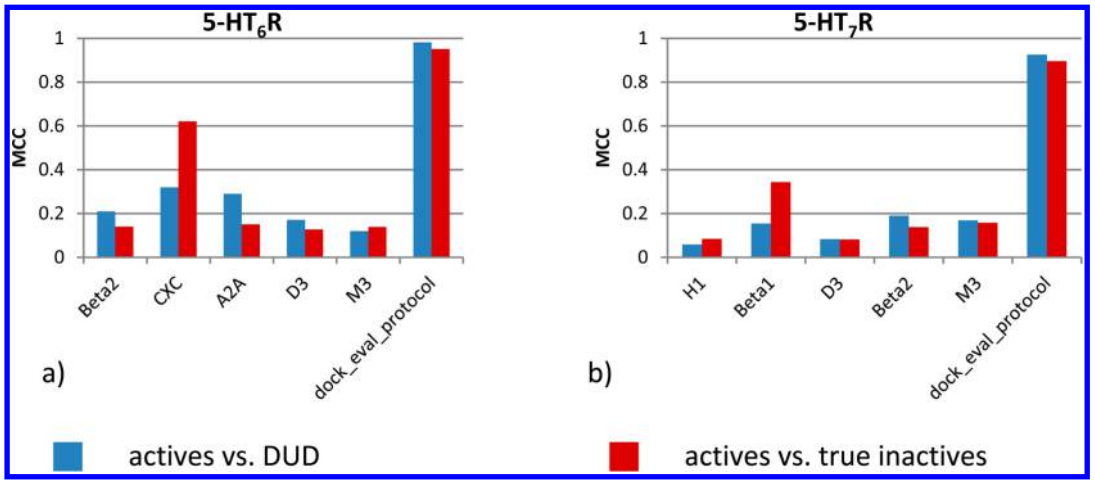


Figure 2. Comparison between MCC values calculated for the raw docking results for the five best models and those obtained after the application of the docking results analysis protocol for 5-HT$_6$ (a) and 5-HT$_7$ (b) ligands.

In the next step, for each instance, the consensus from all learning algorithms was generated by calculating the weighted average, with weights provided by the performance (measured by the Matthews Correlation Coefficient (MCC)[52]) of ML methods from the previous step (the better the performance of the algorithm, the higher the weight). A final prediction for a particular ligand docked into a given receptor model was produced, with weights being value of the scoring function provided by the docking program (Glide Score[53]). The final step was connected with consensus making, that is, a weighted average of the results obtained for the various receptor models built on different templates, with weights being the values of AUROC calculated during the construction of the homology models. The performance of the ML methods was measured by the MCC calculated after each step of the developed protocol with the following formula:

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}$$

The results were compared with the MCC values calculated on the basis of docking results using the above formula, with TP (true positives) − the number of successfully docked active molecules, FP (false positives) − the number of successfully docked inactive compounds, FN (false negatives) − the number of active compounds that failed to dock, and TN (true negatives) being the number of inactive molecules that failed to dock. In order to provide the objectivity of this comparison, the assignment into the docked/undocked group was performed on the basis of the GlideScore function values (cutoff levels were tested in the range from −10 to −3 with a step of 0.1; the highest MCC obtained for various cutoffs was taken as a raw docking result).

The ML-based protocol for post-docking analysis is presented in Figure 1, and details of its way of working for selected active and inactive molecule are presented in Figure S4 of the Supporting Information.

Using weights based on different conformations posed a problem for compounds that did not produce ligand−receptor complexes with any of the protein models. To overcome this issue, 11 values of weights assigned to instances that were unable to dock were tested (ranging from −0.5 to 0.5 in steps of 0.1). Similarly, various thresholds were applied to determine the class assignment after the Glide Score weighting (values from −0.5 to 0.5 with 0.1 step), giving a total of 121 tested combinations.

Additional Validation for Beta-2 Adrenergic Receptor. Additional validation of the protocol was carried out also for the crystal structures of the beta-2 adrenergic receptor (beta-2 AR) (with crystal resolution as the quality measure), with crystal structures fetched from the PDB repository (five crystals with the highest resolution, covering agonist-bound and inverse agonist-bound conformations) and sets of compounds prepared in an analogous way as in the case of serotonin receptor ligands (Table S2, Supporting Information).

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823−832

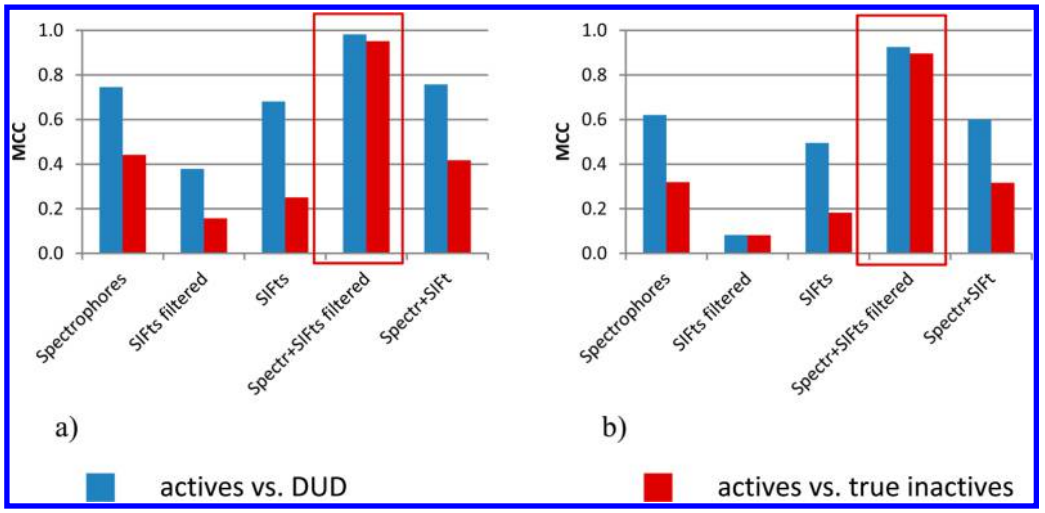


Figure 3. MCC values obtained after applying the docking results analysis protocol for various ligand–protein complex representations for 5-HT$_6$ (a) and 5-HT$_7$ (b) ligands. The representation providing the best results is identified by a red rectangle.

Table 3. Numerical Values of MCC Parameters Obtained for Various Representations of Ligand–Protein Complexes[a]

target	type of experiment/MCC value				
5-HT$_6$			**actives vs DUDs**		
	Spectrophores	SIFts filtered	SIFts	Spectr+SIFts filtered	Spectr+SIFts
	0.746	0.379	0.681	**0.982**	0.757
			actives vs true inactives		
	Spectrophores	SIFts filtered	SIFts	Spectr+SIFts filtered	Spectr+SIFts
	0.442	0.157	0.251	**0.951**	0.417
5-HT$_7$			**actives vs DUDs**		
	Spectrophores	SIFts filtered	SIFts	Spectr+SIFts filtered	Spectr+SIFts
	0.620	0.082	0.495	**0.925**	0.600
			actives vs true inactives		
	Spectrophores	SIFts filtered	SIFts	Spectr+SIFts filtered	Spectr+SIFts
	0.319	0.081	0.182	**0.896**	0.316

[a]Highest MCC values in a particular row are represented in bold.

RESULTS AND DISCUSSION

The input sets of compounds (Figure 1) were docked into a collection of homology models of the 5-HT$_6$ and 5-HT$_7$ receptors, after preprocessing with LigPrep.[30] The docking results for each model selected for further study (sorted by descending AUROC value) are collected in Table 2. Docking statistics for additional evaluation performed for beta-2 AR ligands are presented in Table S2 of the Supporting Information.

The CV experiments were performed for the number of folds equal to 3 and 5; the results of the 3-fold experiments can be found in Figures S5–S8 and Table S3 of the Supporting Information, together with the numerical values of the obtained MCCs.

Cross-Validation. *Overall Performance.* The results for the ML-based post-docking analysis demonstrate its undisputable advantage compared with the raw docking results. The MCC values calculated on the basis of the docking outcome (separately for each template) and those calculated after the application of the developed protocol (in 5-fold CV) are presented in Figure 2. The latter shows the best values obtained for all tested combinations of thresholds and all tested compound representations.

The results of screening based solely on the docking score clearly show the lack of discriminative power of such an approach. Although the docking programs always provide a docking score of some sort, its values do not unequivocally determine the compound activity (as in the example shown in Table 1, where the AUROC values fluctuated around a value of approximately 0.7). Application of a multi-step protocol for the docking results analysis provided an almost perfect discrimination between active and inactive molecules (both for assumed inactives generated according to the DUD methodology and true inactives picked from the ChEMBL database). The MCC values in all cases are close to or above 0.9, which corresponds to very high classification efficiency. As expected, the discrimination between actives and DUDs was a little bit more effective than the discrimination between actives and true inactives.

Similar tendencies regarding the overall performance of the docking evaluation protocol can be observed for the experiments performed for beta-2 AR ligands; the protocol application led to significant improvement of the predictions activity (Figure S9, Supporting Information).

Representation-Dependent Analysis. The docking protocol was evaluated for five distinct configurations of ligand–protein complex representations: Spectrophores, SIFts, SIFts with attribute filtering, Spectrophores and SIFts combined, and Spectrophores and SIFts after attribute filtering.

Those representations were found to provide various ML methods performance in terms of their application in ML-based experiments. The MCC statistics are presented in Figure 3, and

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823–832

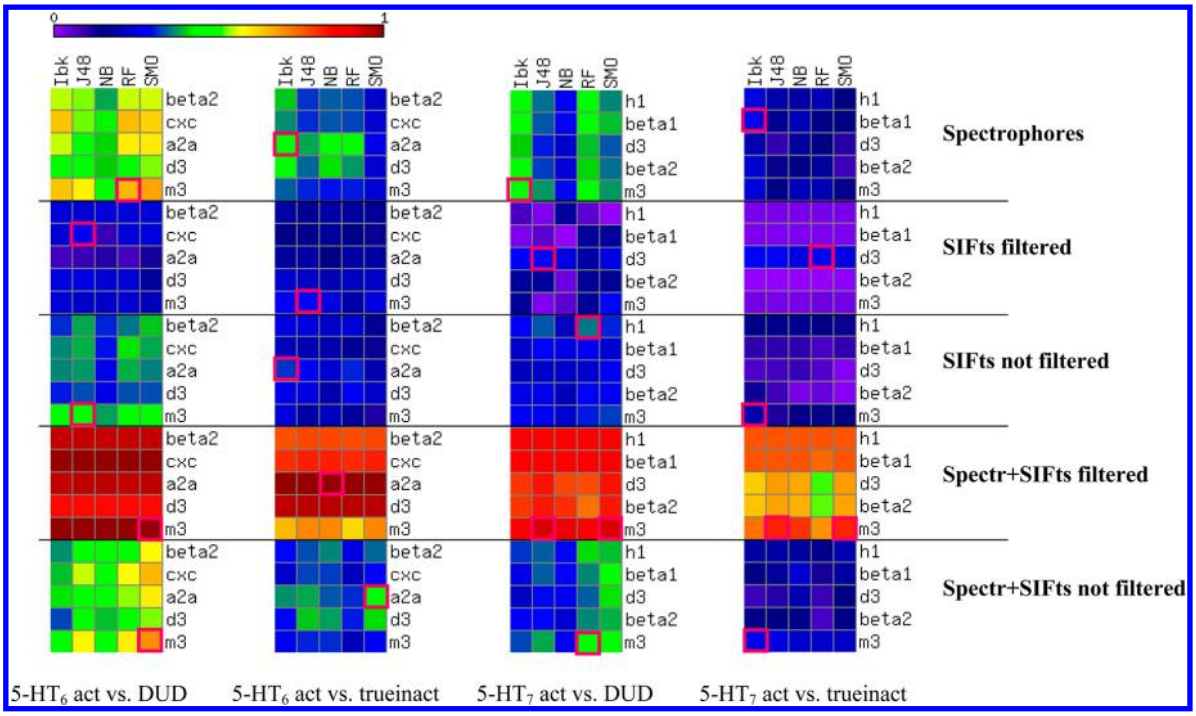


Figure 4. MCC values obtained for individual methods in 5-fold CV experiments. Red squares indicate the best result obtained for a given representation.

the numerical values of the parameter evaluations are gathered in Table 3.

The results presented in Figure 3 clearly show that choosing an appropriate representation of docking results is crucial for obtaining high classification efficiency. The best results were obtained when the ligand−protein complexes were represented by the combination of Spectrophores and SIFts. For both targets, the MCC was close to or over 0.9 both for actives vs DUD and actives vs true inactives experiments.

For all representations tested, the MCC values were higher by 0.2−0.3 on average when active compounds were distinguished from DUDs in comparison to actives vs true inactives experiments. Interestingly, the attribute filtering procedure had different influences on the results for the different representations used. In the case of joint SIFt +Spectrophores representation, attribute filtering improved the results significantly (with more than a 0.5 increase in MCC in some cases when discriminating actives vs true inactives), whereas when SIFts were applied individually, attribute filtering led to lower MCC values, especially for the actives vs DUD classification (up to 0.3).

Performance of ML Algorithms. The performance of different ML methods was also examined (the results are presented in Figure 4 in the form of a heat map of MCC values; red squares indicate the highest MCC values obtained in a particular set of experiments).

The results revealed that the best representation of docking results in ML is a combination of SIFts and Spectrophores with a reduced number of attributes (the red bar of squares referring to filtered SIFts+Spectrophores clearly indicates this fact).

In addition, the analysis of the results showed that there is no best ML method that can be universally applied in all experiments of this type, as each method was indicated at least once as being the one providing the best performance. The MCC values obtained for Spectrophores+SIFts representation for both types of classification (actives vs DUD and actives vs true inactives) are very high, approaching to over 0.9.

The results were organized in order of decreasing AUROC values calculated during the homology modeling procedure. The results for the receptors analyzed in this study indicated that the most effective templates in terms of AUROC did not provide the highest performance for the ML methods. For example, templates indicated as the best ones (beta-2 AR for 5-HT_6R and H_1R for 5-HT_7R) were not found to provide the highest classification efficiency of ML methods even once. In contrast, M_3R, which was indicated as the least effective template for discrimination of actives and inactives on the basis of docking results, was found to provide the highest efficiency in nine of the ML experiments.

The results of the protocol shift the ranking of "the best" templates, yet they do not imply the unviability of the template selection method based on AUROC because the objectives of each stage are slightly different. AUROC-based selection of the templates aims for picking the homology models of the best quality, being the good compromise between discrimination power and the number of docked compounds. In addition such an approach is easy to automate and so can easily rule out the majority of simply bad models from the vast number of input receptor conformations. The ML approach, on the other hand, as more sophisticated and time and resources consuming, is used to squeeze out the most of the docking results obtained for given set of models. The MCC scores obtained for different templates do not reflect the AUROC-based ranking because the structural features of the ligand−receptor complexes used in ML are different from ones important for docking itself.

Efficiency of Individual ML Methods in Target-Specific Experiments. The comparison of the experiments providing the highest classification efficiency for a given ligand−receptor representation, along with the performance of the whole docking results evaluation protocol is presented in Figure 5.

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823−832

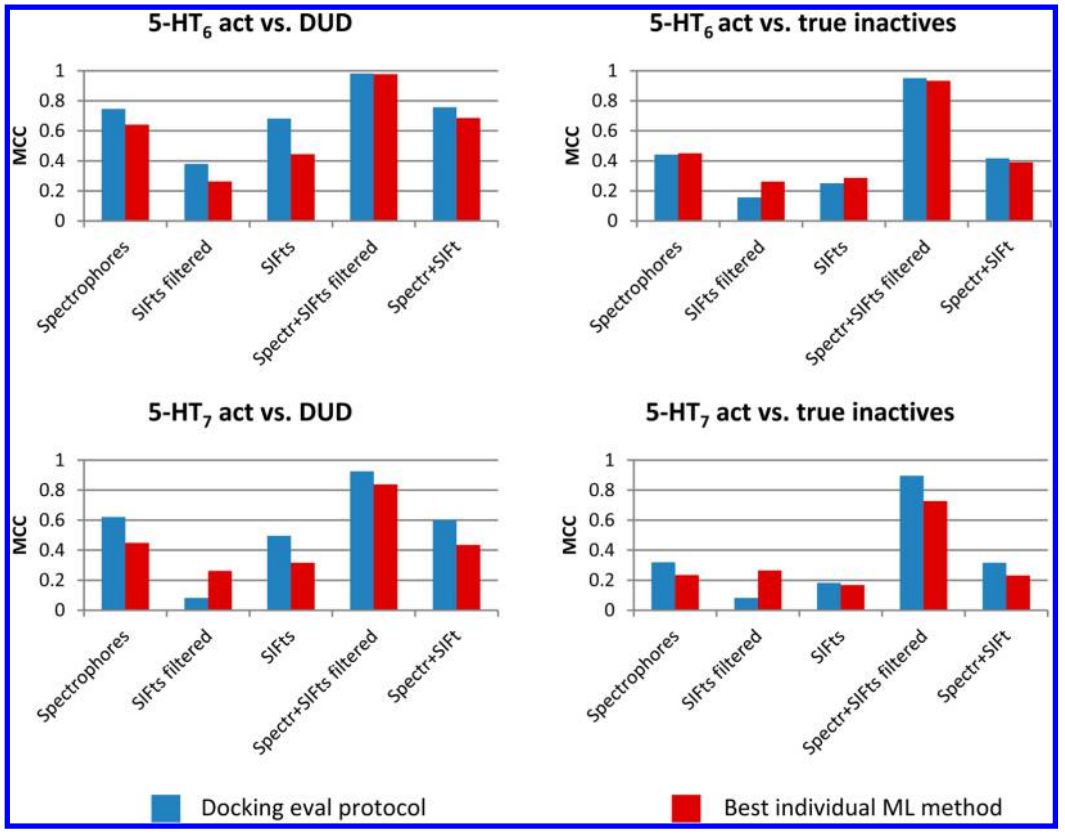


Figure 5. Comparison of the developed protocol with the performance of individual ML experiments providing the highest classification efficiency.

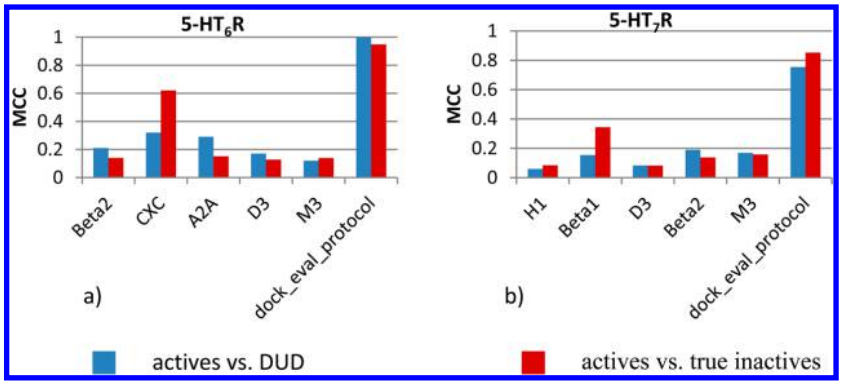


Figure 6. Comparison between MCC values calculated for the raw docking results for the five best models and those obtained after the application of the docking results analysis protocol for 5-HT$_6$ (a) and 5-HT$_7$ (b) ligands in the leave-one-cluster-out approach.

In the majority of cases, the application of the developed protocol led to an improvement in the obtained results. However, for the best representation of the docking results (Spectrophores+SIFts filtered), the level of this improvement is not very high and is caused by the relatively high efficiency of individual ML methods (close to 0.9 in MCC). Therefore, it was difficult to provide substantial further improvement in the results. Spectrophores+SIFts filtered were the representation for which the MCC values were higher in each case (actives vs DUD and actives vs true inactives experiments for both examined targets) when individual ML methods were considered in comparison to the whole docking results evaluation protocol.

It should also be noted that the effectiveness of the protocol was compared with the best ML method under particular conditions. However, before conducting an experiment, the best conditions for given experiments are unknown. Therefore, by applying the described protocol, one can take into account as much information as possible—the different conformations of docked compounds, the effectiveness of various ML methods, and the value of the docking score function. Such a comprehensive analysis is much better than selecting and determining the best experimental conditions for each particular case.

Influence of Weighting Thresholds. Two stages in the developed evaluation protocol require optimization of some parameters (assigning weights for compounds that failed to dock into the binding site and setting a threshold for activity determination after GlideScore[53] weighting).

The threshold for class assignment had no influence on the final results, and for the given value of weight for the compounds that failed to dock, the evaluation parameters'

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823−832

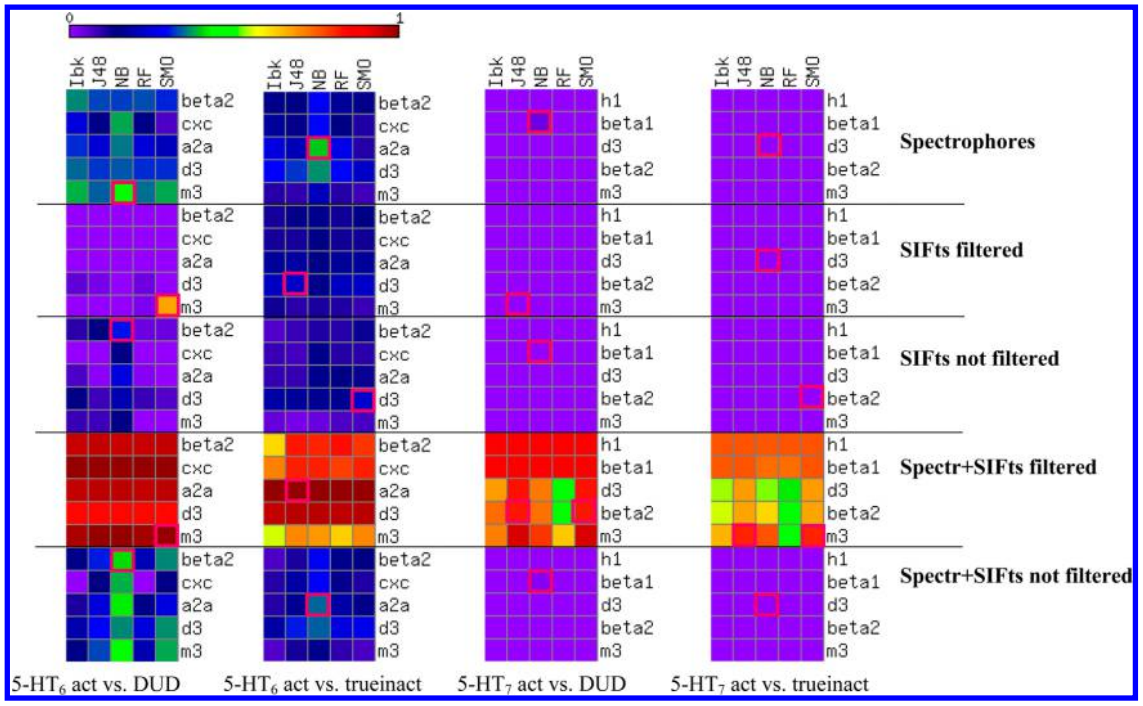


Figure 7. MCC values obtained for individual methods in 5-fold CV experiments in the leave-one-cluster-out approach. Red squares indicate the best result obtained for a given representation. Red squares indicate the best result obtained for a given representation.

values remained unchanged, despite changes in the threshold values. The list of optimal weight values in terms of the highest final performance of the whole protocol together with their standard deviations are gathered in Table S4 of the Supporting Information. In general, zero or negative values were preferred. The weight for compounds that failed to dock had quite a significant influence on the classification efficiency, as the standard deviation of MCC over its various values was equal to several percentage points.

Additional analysis of the influence of the threshold was carried out, during which the results were examined from the point of view of various templates, for a fixed weight of undocked compounds. The results of this part of the study are presented in Figure S10 of the Supporting Information.

Leave-One-Cluster-Out Approach. Similarly to the results obtained the CV, the developed protocol was found to be very effective in active/inactive discrimination, even if compounds with similar structures were not present in the training set. The comparison of the developed protocol with the raw docking results are presented in Figure 6, whereas the results obtained for the selected ML methods are shown in Figure 7. The results obtained for this approach prove the outstanding performance of SIFts+Spectrophores after attribute filtering compared to other representations of complexes.

CONCLUSIONS

It was proved that the developed protocol enabled proper discrimination between active and inactive molecules, improving the results provided by the docking procedure also in terms of recognition of new structures. In addition, a combination of SIFts, Spectrophores, and the attributes filtering procedure for the obtained set was found to be the most effective strategy for our case study. Comprehensive predictive models were obtained by taking into account various aspects connected with docking (different conformations of ligands and the impact of the template used for homology model construction) and ML experiments (performance of the particular algorithm). However, as the results show, there is no universal ML approach that gives optimal results, and when using the protocol described in this paper, it is advised that the protocol be optimized first to select the best-performing combination of parameters. Application of the presented protocol also improved the results obtained by individual learning algorithms because classification effectiveness was higher after its application even when compared to the best method in a particular type of experiment. It is also worth indicating that the homology models that performed best in docking were not necessarily best for ML-based experiments, and even those that were evaluated by lower AUROC values were able to provide classification results at the highest level. The developed protocol is not only able to properly evaluate the activity of compounds when similar structures where present in the training set but is also capable of properly discriminating new actives and inactives that was proved in the leave-one-cluster out procedure.

ASSOCIATED CONTENT

Supporting Information

Table S1: Crystal structures used for homology modeling purposes. **Figure S1**: ROC curves obtained for the selected homology models. **Figure S2**: Distribution of log D and pK_a values for 5-HT$_6$- and 5-HT$_7$-oriented sets of compounds. **Figure S3**: Example of SIFts and Spectrophores generation. **Figure S4**: Exemplary analysis of activity of two compounds by docking evaluation protocol. **Table S2**: Docking statistics for beta-2 adrenergic receptor crystal structures. **Figure S5**: Comparison between MCC values calculated for the raw docking results for the five best models and those obtained after applying the docking results analysis protocol for 5-HT$_6$ and 5-HT$_7$ ligands for 3-fold CV. **Figure S6**: MCC values obtained after applying the docking results analysis protocol for various representations of ligand−protein complexes for 5-HT$_6$ and 5-

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823−832

HT_7 ligands for 3-fold CV. **Figure S7**: MCC values obtained for individual methods in 3-fold CV experiments. **Figure S8**: Comparison of the developed protocol with the performance of the individual ML experiment providing the highest classification efficiency in 3-fold CV. **Table S3**: Numerical values of MCC parameters obtained for various representations of ligand−protein complexes for 3-fold CV. **Figure S9**: Comparison between the MCC values obtained for individual beta-2 adrenergic receptor crystal structures (the best ML method) and with the use of docking evaluation protocol in actives/true inactives beta-2 AR ligands classification. **Table S4**: Optimal values and standard deviation of the weights assigned to undocked compounds providing the best performance of the whole evaluation protocol. **Figure S10**: Analysis of the influence of threshold changes on ML protocol performance for Spectrophores+SIFTs representation for a given template. **Data set S1**: Exemplary WEKA input files. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bojarski@if-pan.krakow.pl.

Author Contributions

The manuscript was written with contributions from all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.S., S.M., J.W., K.R., and A.J.B. participate in the European Cooperation in Science and Technology (COST) Action CM1207: GPCR-Ligand Interactions, Structures, and Transmembrane Signalling: an European Research Network (GLISTEN). The study was supported by grant PRELUDIUM 2013/09/N/NZ2/01917 financed by the Polish National Science Centre (https://www.ncn.gov.pl/?language=en).

ABBREVIATIONS

AUROC, Area under Receiver Operating Characteristic; CV, cross-validation; DUD, Directory of Useful Decoys; GPCRs, G protein-coupled receptors; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; IP PAS, Institute of Pharmacology Polish Academy of Sciences; MCC, Matthew's Correlation Coefficient; ML, machine learning; MW, molecular weight; NB, Naïve Bayes; RF, Random Forest; rotB, rotatable bond; SMO, Sequential Minimal Optimization; SIFt, Structural Interaction Fingerprint; VS, virtual screening

REFERENCES

(1) Plewczynski, D.; Łazniewski, M.; Augustyniak, R.; Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput. Chem.* **2011**, *32*, 742−755.

(2) Kumar, A.; Zhang, K. Y. J. Hierarchical Virtual Screening Approaches in Small Molecule Drug Discovery. *Methods* **2014**, *71*, 26−37.

(3) Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Simple, Intuitive Calculations of Free Energy of Binding for Protein−Ligand complexes. 1. Models without Explicit Constrained Water. *J. Med. Chem.* **2002**, *45*, 2469−2483.

(4) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting Binding Modes, Binding Affinities and 'Hot Spots' for Protein−Ligand Complexes Using a Knowledge-Based Scoring Function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.

(5) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein−Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(6) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.

(7) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(8) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

(9) Yang, J.-M. Y J.-M.; Hsu, D. F. Consensus Scoring Criteria in Structure-based Virtual Screening. *Proceedings from the Emerging Information Technology Conference*, Taipei, Taiwan, 2005, p 3.

(10) Sastry, G. M.; Inakollu, V. S. S.; Sherman, W. Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from 2D Fingerprints, Shape, and Docking. *J. Chem. Inf. Model.* **2013**, *53*, 1531−1542.

(11) Kinnings, S. L.; Liu, N.; Tonge, P. J.; Jackson, R. M.; Xie, L.; Bourne, P. E. A Machine Learning-Based Method to Improve Docking Scoring Functions and Its Application to Drug Repurposing. *J. Chem. Inf. Model.* **2011**, *51*, 408−419.

(12) Schechter, L. E.; Lin, Q.; Smith, D. L.; Zhang, G.; Shan, Q.; Platt, B.; Brandt, M. R.; Dawson, L. A.; Cole, D.; Bernotas, R.; Robichaud, A.; Rosenzweig-Lipson, S.; Beyer, C. E. Neuropharmacological Profile of Novel and Selective 5-HT_6 Receptor Agonists: WAY-181187 and WAY-208466. *Neuropsychopharmacology* **2008**, *33*, 1323−1335.

(13) Hedlund, P. B.; Sutcliffe, J. G. Functional, Molecular and Pharmacological Advances in 5-HT7 Receptor Research. *Trends Pharmacol. Sci.* **2004**, *25*, 481−486.

(14) Xu, F.; Wu, H.; Katritch, V.; Han, G. W.; Jacobson, K. A.; Gao, Z.-G.; Cherezov, V.; Stevens, R. C. Structure of an Agonist-Bound Human A2A Adenosine Receptor. *Science* **2011**, *332*, 322−327.

(15) Warne, T.; Moukhametzianov, R.; Baker, J. G.; Nehmé, R.; Edwards, P. C.; Leslie, A. G. W.; Schertler, G. F. X.; Tate, C. G. The Structural Basis for Agonist and Partial Agonist Action on a β(1)-Adrenergic Receptor. *Nature* **2011**, *469*, 241−244.

(16) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-Resolution Crystal Structure of an Engineered Human Beta2-Adrenergic G Protein-Coupled Receptor. *Science* **2007**, *318*, 1258−1265.

(17) Wu, B.; Chien, E. Y. T.; Mol, C. D.; Fenalti, G.; Liu, W.; Katritch, V.; Abagyan, R.; Brooun, A.; Wells, P.; Bi, F. C.; Hamel, D. J.; Kuhn, P.; Handel, T. M.; Cherezov, V.; Stevens, R. C. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science* **2010**, *330*, 1066−1071.

(18) Chien, E. Y. T.; Liu, W.; Zhao, Q.; Katritch, V.; Han, G. W.; Hanson, M. A.; Shi, L.; Newman, A. H.; Javitch, J. A.; Cherezov, V.; Stevens, R. C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* **2010**, *330*, 1091−1095.

(19) Shimamura, T.; Shiroishi, M.; Weyand, S.; Tsujimoto, H.; Winter, G.; Katritch, V.; Abagyan, R.; Cherezov, V.; Liu, W.; Han, G. W.; Kobayashi, T.; Stevens, R. C.; Iwata, S. Structure of the Human Histamine H1 Receptor Complex with Doxepin. *Nature* **2011**, *475*, 65−70.

(20) Haga, K.; Kruse, a C.; Asada, H.; Yurugi-Kobayashi, T.; Shiroishi, M.; Zhang, C.; Weis, W. I.; Okada, T.; Kobilka, B. K.; Haga, T.; Kobayashi, T. Structure of the Human M2 Muscarinic Acetylcholine Receptor Bound to an Antagonist. *Nature* **2012**, *482*, 547−551.

(21) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823−832

Dynamics of the M3 Muscarinic Acetylcholine Receptor. *Nature* **2012**, *482*, 552−556.
(22) Wang, C.; Jiang, Y.; Ma, J.; Wu, H.; Wacker, D.; Katritch, V.; Han, G. W.; Liu, W.; Huang, X. P.; Vardy, E.; McCorvy, J. D.; Gao, X.; Zhou, X. E.; Melcher, K.; Zhang, C.; Bai, F.; Yang, H.; Yang, L.; Jiang, H.; Roth, B. L.; Cherezov, V.; Stevens, R. C.; Xu, H. E. Structural Basis for Molecular Recognition at Serotonin Receptors. *Science* **2013**, *340*, 610−614.
(23) Wacker, D.; Wang, C.; Katritch, V.; Han, G. W.; Huang, X.-P.; Vardy, E.; McCorvy, J. D.; Jiang, Y.; Chu, M.; Siu, F. Y.; Liu, W.; Xu, H. E.; Cherezov, V.; Roth, B. L.; Stevens, R. C. Structural Features for Functional Selectivity at Serotonin Receptors. *Science* **2013**, *340*, 615−619.
(24) Magrane, M.; Consortium, U. UniProt Knowledgebase: A Hub of Integrated Protein Data. *Database (Oxford)* **2011**, *2011*, bar009 DOI: 10.1093/database/bar009.
(25) *Discovery Studio Modeling Environment*, release 4.0; Accelrys Software, Inc: San Diego, 2013.
(26) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.
(27) Nowak, M.; Kołaczkowski, M.; Pawłowski, M.; Bojarski, A. J. Homology Modeling of the Serotonin 5-HT1A Receptor Using Automated Docking of Bioactive Compounds with Defined Geometry. *J. Med. Chem.* **2006**, *49*, 205−214.
(28) Rataj, K.; Witek, J.; Mordalski, S.; Kosciolek, T.; Bojarski, A. J. Impact of Template Choice on Homology Model Efficiency in Virtual Screening. *J. Chem. Inf. Model.* **2014**, *54*, 1661−1668.
(29) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−1107.
(30) *Schrödinger Release 2014-3: LigPrep*, version 3.1; Schrödinger, LLC: New York, 2014.
(31) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861−874.
(32) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(33) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(34) InstantJChem Version 5.8.2, 2011, ChemAxon. www.chemaxon.com (accessed September 15, 2014).
(35) RDKit: Open-Source Cheminformatics. http://www.rdkit.org (accessed September 15, 2014).
(36) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337−344.
(37) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J. P. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *J. Phys. Chem. A* **2002**, *106*, 7895−7901.
(38) *Schrödinger Release 2014-3*; Schrödinger, LLC: New York, 2014.
(39) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33−47.
(40) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD ExplorationsNewsletter* **2009**, *11*, 10−18.
(41) Rish, I. An Empirical Study of the Naive Bayes Classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001; Vol. 3, pp 41−46.
(42) Platt, J. C. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*; Technical Report MSR-TR-98-14; Microsoft Research: Redmond, WA, 1998; pp 1−21.
(43) Cunningham, P.; Delany, S. J. *K-Nearest Neighbour Classifiers*; Technical Report UCD-CSI-2007-4; School of Computer Science and Informatics, University College: Dublin, Ireland, 2007; pp 1−17.
(44) Korting, T. S. C4.5 Algorithm and Multivariate Decision Trees. Image Processing Division, National Institute for Space Research−INPEl. http://www.academia.edu/1983952/C4._5_algorithm_and_Multivariate_Decision_Trees (accessed September 15, 2014).
(45) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(46) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: a Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(47) Hall, M. A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Wakaito, Hamilton, New Zealand, April 1999.
(48) Whitley, D. A Genetic Algorithm Tutorial. *Stat. Comput.* **1994**, *4*, 65−85.
(49) *Schrödinger Release 2014-3: Canvas*, version 2.1, Schrödinger, LLC: New York, 2014.
(50) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol Chem.* **2004**, *2*, 3204−3218.
(51) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools against the DUD Data Set Reveals Limitations of Current 3D Methods. *J. Chem. Inf. Model.* **2010**, *12*, 2079−2093.
(52) Vihinen, M. How To Evaluate Performance of Prediction Methods? Measures and Their Interpretation in Variation Effect Analysis. *BMC Genomics* **2012**, *13* (Suppl 4), S2.
(53) 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.

DOI: 10.1021/ci500564b
J. Chem. Inf. Model. 2015, 55, 823−832