JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Improvement of Virtual Screening Results by Docking Data Feature Analysis

Marcelino Arciniega*[,†,‡] and Oliver F. Lange[‡,§]

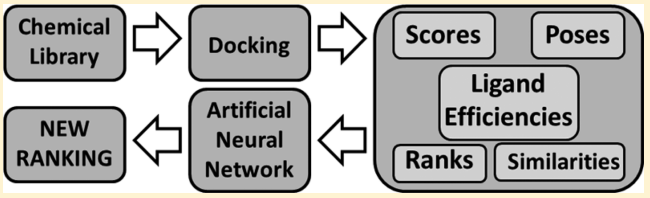
[†]Max Planck Institute Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[‡]Biomolecular NMR and Munich Center for Integrated Protein Science, Department Chemie, Technische Universität München, 85747 Garching, Germany

[§]Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

Supporting Information

ABSTRACT: In this study, we propose a novel approach to evaluate virtual screening (VS) experiments based on the analysis of docking output data. This approach, which we refer to as docking data feature analysis (DDFA), consists of two steps. First, a set of features derived from the docking output data is computed and assigned to each molecule in the virtually screened library. Second, an artificial neural network (ANN) analyzes the molecule's docking features and estimates its activity. Given the simple architecture of the ANN, DDFA can be easily adapted to deal with information from several docking programs simultaneously. We tested our approach on the Directory of Useful Decoys (DUD), a well-established and highly accepted VS benchmark. Outstanding results were obtained by DDFA not only in comparison with the conventional rankings of the docking programs used in this work but also with respect to other methods found in the literature. Our approach performs with similar good results as the best available methods, which, however, also require substantially more computing time, economic resources, and/or expert intervention. Taken together, DDFA represents an automatic and highly attractive methodology for VS.



1. INTRODUCTION

In the early stages of the drug discovery process, large chemical libraries are screened to identify lead molecules that allow the development of new drugs. The large amount of experimental resources that are required renders this search highly expensive and time-consuming.[1] In this context, in sillico virtual screening (VS) has appeared as a fast and economic approach that increases the efficiency of the lead discovery process.[2−4]

There are two main approaches to perform VS: ligand-based and structure-based.[5] In the former, previously known active ligands are used to identify other molecules with similar characteristics; in the latter, protein and ligand structural models at atomic resolution are used to evaluate their binding affinity. Structure-based methods perform generally better than ligand-based methods in identifying new lead compounds.[6] In structured-based methods, the screening is performed by docking each of the library's molecules into the receptor's active site while optimizing the atomic interactions between the binding partners.

Docking methods are foremost developed to identify the ligand's actual binding mode from the large set of sampled conformations tested.[7] In this regard, docking software like Autodock, AutodockVina, and RosettaLigand have achieved high performances.[8−10] Compared to the discrimination of correct and incorrect conformations of the same ligand, it is far more challenging to discriminate active from inactive ligands, as it is required for structured-based screening. This complication arises from the difficulty to evaluate free energy terms of the unbound ligand state, such as solvent and conformational entropies. Whereas these terms cancel out when comparing conformations of the same ligand, they do not cancel out when treating different molecules. Nevertheless, due to the lack of better ranking methods, it is common practice to rank ligands based only on the docking score of their best docked conformations.

Many attempts have been made to improve the ranking of ligands beyond the accuracy obtained by using the plain docking score.[11−16] For instance, García-Sosa et al.[14] reported that a better correlation between docking scores and experimental binding energies can be achieved by dividing the docking score by the ligand's size; resulting in a descriptor often referred to as ligand efficiency (LE). Others have analyzed several high ranking conformations per ligand, rather than considering only the best scored conformations. For example, Seok et al.[15] augmented the binding energy evaluation by adding an entropy term that was estimated from the populations of clusters of high ranked binding modes. To establish a new ranking, Wallach et al.[16] compared scores of query molecules with that of physically similar (molecular weight, number of rotational bonds, number of hydrogen acceptor/donor, etc.) but chemically dissimilar (different

Received: January 14, 2014
Published: May 5, 2014

   dx.doi.org/10.1021/ci500028u | J. Chem. Inf. Model. 2014, 54, 1401−1411

topology and functional groups) decoys. For each molecule in the screening library, a set of decoys is generated. The approach is based on the assumption that an active ligand should obtain a significantly higher score than the decoys' score distribution. Although the method turned out as a success, its main disadvantage is that the amount of molecules to be docked increases by 2 orders of magnitude.

The common theme of the methods mentioned above is to rank the screening library by a modified form of a single scoring function. An alternative class of approaches attempts to overcome the deficiencies of individual scoring functions by employing two or more docking programs in a consensus scheme.[17] One of the most popular ways to combine multiple ranked lists is the "rank-by-rank" approach.[18] For each molecule the average of its ranks is computed and used to establish the final ranking list. A different method would be to average scores from different scoring functions. It has been shown that both of these methods rely crucially on the diversity and high quality of the scoring functions.[19] A more sophisticated analysis was developed by Jacobsson et al.[20] In their work, data mining techniques were used to create "if−then" rules that yielded upper and lower bounds to seven scoring functions.

Here we describe a novel framework to predict the ligand activity based on a diverse set of docking features rather than focusing on a single kind; such as the docking score. This framework, which we named docking data feature analysis (DDFA), converts this set of docking features into a feature score. The signal conversion is performed by an artificial neural network (ANN) that can be trained to work with data from either single or several docking programs. In our particular case we performed the analysis using three programs and five docking features: (i) best docking score, (ii) ligand efficiency, (iii) scores from similar molecules, (iv) the position of the ligand's poses within the general rank, and (v) structural consistency of the ligand's poses. These features were selected to capture different aspects that are typically employed in a human expert analysis to identify active binding molecules from the VS ranking. Bearing this in mind, the docking score feature represents the traditional approach, which assumes correlation between score and activity. The ligand efficiency feature contributes to a size independent comparison among ligands. Monitoring the performance of chemically similar molecules is inspired by the structural activity relationship (SAR) central idea, which is that similar molecules have similar binding energies. The feature that monitors the ranks of the ligand poses assumes that poses from an active molecule are not distributed randomly through the entire rank. The pose variability feature exploits that active ligands often show better converged poses. It is important to mention that the DDFA can be easily extended and/or adapted to include other features.

To test the DDFA approach, we docked the broadly used Directory of Useful Decoys[21] (DUD) using three different docking programs—Autodock4,[22] Autodockvina,[9] and RosettaLigand[23]—and predicted ligand activity. DUD is widely accepted for benchmarking VS protocols. It consists of 40 receptors of pharmaceutical relevance and a screening library of over 100 000 molecules. To predict the ligand activities for a receptor of DUD benchmark, the DDFA ANN was trained using 22 receptors from the remaining 39 data sets. The 22 receptors of the training set were randomly selected after removing receptors with similar biological activity or with reported positive cross-enrichment, with respect to the receptor to be evaluated. We repeated this process with a different receptor left out of the training set each time to obtain ligand activity predictions for each receptor in DUD. As a control, DDFA's performance was compared to that of the individual docking scores of the used programs and a consensus ranking; with the latter generated by the ligand's best rank in any program. The performance evaluation was carried out using well established and broadly accepted metrics, such as enrichment factor (ef) and the area under the curve (auc) from the receiver operator characteristic (ROC) curve.

2. METHODS

2.1. Docking Programs. In this study three docking programs Autodock4.2[22] (AD4), Autodockvina1.2[9] (ADV), and RosettaLigand3.4[23] (RL) were used. Although AD4 and ADV were developed by the same lab, they differ in the sampling methods and weights of individual score terms. RL is part of the Rosetta's software suite for modeling macromolecular structures. We used AD4 and ADV with a rigid receptor model and RL with flexible side-chains for the receptor.

2.1.1. Docking Using AD4 and ADV. The receptors and ligands were prepared following the standard setup protocols using Gasteiger partial charges.[22] The grid sizes were set up to 27 Å × 27 Å × 27 Å in both programs, using as grid center the center of mass of the ligand provided by the DUD to localize the binding pocket. For AD4, the receptor grid was generated using autogrid4 with 0.375 Å of grid spacing. The docking parameter file was generated with the prepare_dpf42.py script in AutoDockTools.[22] The Lamarckian genetic algorithm with default parameters was selected as pose search method.[8] Ten output poses were requested. For ADV, a maximum of ten output poses was kept using a restriction of 3 kcal/mol in the score difference between the best and worse poses. The global search exhaustiveness parameter was set to 16 (default value 8).

2.1.2. Docking Using Rosetta Ligand. For Rosetta Ligand (RL) the receptor side chain conformations were first optimized with the *fixbb* application of Rosetta.[24] The ligands were adapted to the RL format using scripts provided in the Rosetta distribution (molfile_to_params.py).[24] RL searches for docking poses by cycling through a predetermined library of intraligand conformations simultaneously to optimizing the ligand's rigid body degrees of freedom and receptor sidechain dihedral angles. Usually the ligand conformational library is generated with the external program OpenEye's Omega.[25]

In the context of this work, ligand conformations were already available through the AD4 and ADV docking output, and thus, all output poses from AD4 and ADV were used for the ligand conformation library of RL. For every run, the ligand initial placement was provided by the center of mass of a randomly selected member of the conformation library. Docking was performed using the RossetaScripts[26] application with the parameters reported by Davis et al.[27] The number of runs per ligand was set to 50. The top ten structures in interface score were selected for analysis and comparison with the other docking software.

2.2. RAW Rankings. The screening library of each DUD receptor was docked using the docking programs AD4, ADV, and RL a ranking based exclusively on the docking score was generated. A fourth ranking (ALL) was created by assigning to each ligand the best achieved position within any of the individual rankings. Tied cases were resolved by comparing the ligand's standardized docking scores of the individual programs. Docking scores were standardized by subtracting the average

and dividing by the standard deviation of the score-distribution. This standardization procedure is commonly known as Z-Score. We refer to this set of scores as RAW (RAW-AD4, RAW-ADV, RAW-RL, and RAW-ALL) since they represent the most straightforward approach to establish a ranking of a docked library.

2.3. DDFA Rankings. In the DDFA approach, a feature vector is assigned to each ligand and used as input layer of a feed-forward ANN. The term "docking feature" refers to characteristic information computed from the docking data of the screened library. Details of the docking features used in this work are given in section 2.4. The analysis was performed considering docking data from either a single docking program (DDFA-AD4, DDFA-ADV, and DDFA-RL) or from analyzing all three sources simultaneously (DDFA-ALL). ANN's architecture and training procedure is described in section 2.5.

2.4. Docking Features. Docking data is analyzed to derive features that help to discriminate between active and inactive ligands. In this work five features are used in the analysis (DockScore, DockLE, DockSimi, DockPoses, and DockRmsd) and are described in the following sections.

2.4.1. DockScore. This feature is given by the best docking score of the ligand poses. It represents the traditional approach, in which the docking score helps to provide enough information to discriminate an active molecule from an inactive one. Prior to analysis the docking scores were standardized as Z-scores.

2.4.2. DockLE. The ligand efficiency (LE) was computed as the quotient between the best ligand's score and the number of heavy atoms of the ligand.

2.4.3. DockSimi. The DockSimi feature of a ligand is the weighted average of the best docking scores of the five most similar ligands in the docked library. The Tanimoto coefficients (Tc) were used as both similarity measures and weighting factors in the computation of the average. The FP2 molecular fingerprints as implemented in OpenBabel[28] version 2.3.1 were used to compute the Tc. Only ligands with Tc > 0.70 were considered as similar. Whenever no similar ligands existed in the docked library, DockSimi was set to zero.

2.4.4. DockPoses. This feature is a five-dimensional vector composed of the number of ligand poses that are within the top 5%, 10%, 15%, 20%, and 25%, respectively, of all pose-scores in the docked library.

2.4.5. DockRmsd. This feature is a five-dimensional vector given by the RMSD of the second−sixth ranked poses of a ligand when superimposed to the first ranked pose.

2.5. Evaluation of Docking Features Using Artificial Neural Networks. *2.5.1. Architecture of the ANN.* Artificial neural networks (ANNs) are known to perform well on pattern recognition and classification problems.[29] Here we train an ANN to identify active molecules based on the information provided by docking features.

Figure 1 shows a schematic representation of the ANN topology. It consists of 13, 8, and 1 nodes for the input, hidden, and output layers, respectively (Supporting Information Figure S1). The network has full-connectivity among the layers, with linear, sigmoidal, and softmax activation functions for the input, hidden, and output layers, respectively. The ANN was constructed using the PyBrain[30] package. Given the ligand's docking features at the input layer, the returned value at the ANN's output node can be interpreted as a confidence assessment on the ligand's activity chances. Consequently the ligands of the screened library are ranked based on the ANN's output. If DDFA is applied to a single docking program, the docking features give rise to 13 input nodes as follows: one node for features DockScore, DockLE, and DockSimi and five nodes each for the features DockPose and DockRmsd. In the DDFA-ALL, where DDFA is applied to three docking programs simultaneously, the number of input nodes is tripled.

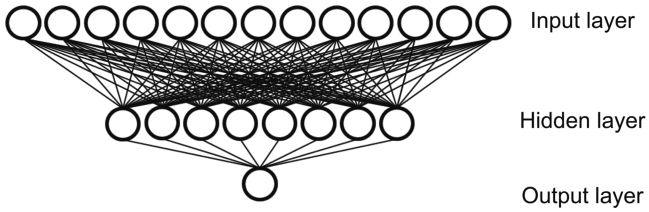


Figure 1. Schematic representation of the ANN used with single docking program. The ANN has a feed-forward architecture consisting of three layers with 13 and 8 nodes in the input and hidden layers, respectively, and a single output node. There is full connectivity among the layers using linear, sigmoid, and softmax as activation functions for the input, hidden, and output nodes, respectively. A detailed description of the input values taken by each of the input nodes can be found in the Methods section. In the case where the information from three docking program is used, the input layer nodes are triplicated.

2.5.2. Training and Application of the ANN. In order to apply the DDFA approach to any of the DUD receptor's, the docking data was divided into three nonoverlapping sets: training, validating, and testing (Figure 2). The training set is used during parameter estimation; the validation set is used to control hyperparameters and to monitor training progress; the test set is used to measure the performance of the methodology as in the reported results. To test the method we use a leave-one-out approach. Thus, one receptor and it's DUD ligands are used as the testing set and remaining receptors and associated DUD ligands are used for training and validation. However, because similarities between the receptor used for testing and the receptors used for training or validation might cause overestimation of the performance for truly new and unseen cases, we further remove any receptors similar to the test receptor from training and validation sets. As similar we consider receptors in the same biological class (Table 1) and also receptors for which positive cross-enrichment has been reported[21] (Table 1, column 2). Because this would cause varying numbers of receptors in training and validation sets, we further reduce their number to always get a total of 22 receptors, which reflects the smallest number of nonsimilar receptors which would ever occur. This final selection is done randomly. Thus, the analysis on the testing set represents a realistic evaluation of the DDFA performance and similar performance would be expected for unknown receptors and screening ligands.

To generate balanced training and validation sets, all active molecules are taken together with the same amount of randomly selected decoys. From this pool with a balanced active ligand to decoy ratio, 70% was used for training and 30% for validation (Supporting Information Figure S2). The training process was conducted under a back-propagation protocol with a value of 0.001 for all three training parameters: weight decay, learning rate, and momentum (Supporting Information Figure S3). The testing set was used to monitor the ANN performance over the training epochs. Training was terminated when a plateau for the test-set performance had been reached. This plateau occurred after 800 epochs for the AD4, ADV, and the

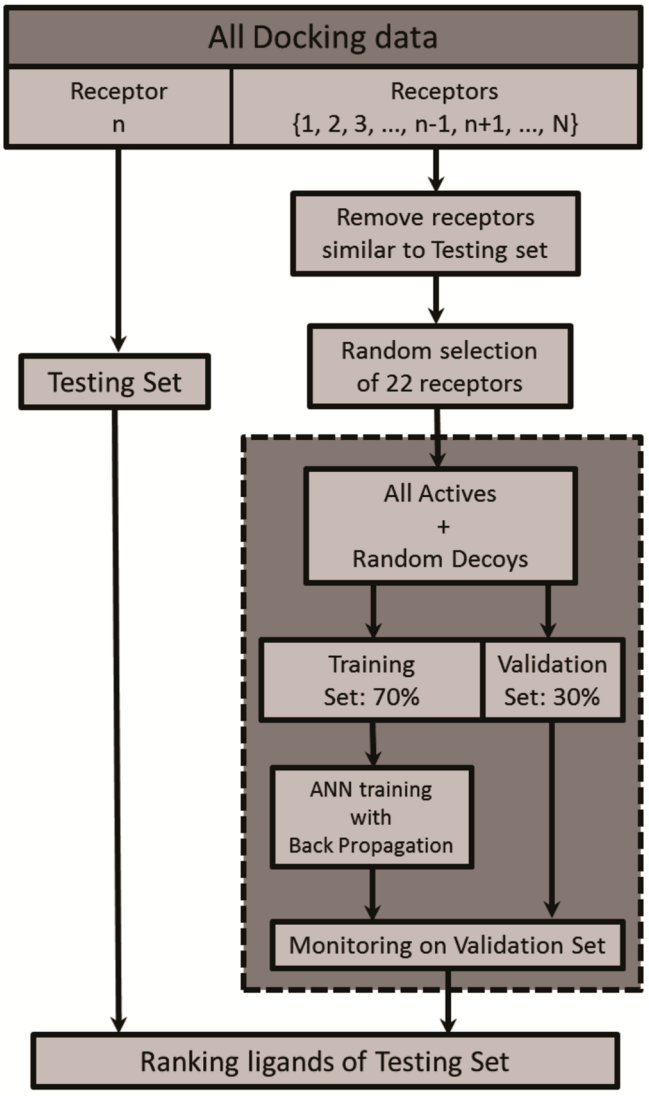


Figure 2. Flowchart depicting the cross-validation and application procedures of the DDFA approach. VS data is separated into three nonoverlapping sets: training, validation, and testing. The receptor to be analyzed constitutes the testing set; any of its data is considered during the training and validation processes (enclosed dashed region). The training and validation sets are formed using 22 receptors. None of these receptors can be similar to the receptor used for testing (Table 1). Docking data belonging to all active molecules, together with same amount of information from random selected decoys, is partitioned into training and validation sets in a 70:30 ratio. The ANN is trained under a back-propagation protocol using the training set, whereas the validation set is used to monitor the ANN performance. As final step, the trained ANN is applied to the test receptor.

ALL-scheme and after 1200 epochs for DDFA-RL (Supporting Information Figure S4). After the training process, ligands from the testing receptor were ranked based on the ANN's output. This ANN training procedure was repeated 40 times, each time with a different set of the 40 receptors of DUD selected as the testing receptor and thus excluded from the training and validation sets.

2.6. VS Performance Evaluation. In order to evaluate the performance of VS experiments, several metrics were computed on the benchmark receptors based on the generated rankings. These metrics are the area under the curve (auc) of the receiver operator characteristic (ROC) curve for sensitivity versus

Table 1. Similarity Relationships between Receptors as Considered to Build Training and Validation Sets[a]

class: nuclear hormone receptors	
receptor	receptors with positive cross-enrichment
AR	TK, ADA, ALR2, PARP, PNP, SAHH
ERagonist	PNP
ERantagonist	*none*
GR	*none*
MR	PARP
PPARg	*none*
PR	*none*
RXRa	COX-1
class: kinases	
receptor	receptors with positive cross-enrichment
CDK2	*none*
EGFr	*none*
FGFr1	*none*
HSP90	*none*
P38MAP	*none*
PDGFrb	*none*
SRC	PDE5
TK	ADA, COMT, ALR2, COX-1, GPB, PARP, PNP, SAHH
VEGFr2	*none*
class: serine proteases	
receptor	receptors with positive cross-enrichment
FXa	DHFR, GART
thrombin	DHFR, ERantagonist
trypsin	PPARg, ADA, DHFR
class: metallo enzymes	
receptor	receptors with positive cross-enrichment
ACE	ALR2
ADA	*none*
COMT	RXRa, ALR2, AmpC, PNP
PDE5	P38MAP
class: folate enzymes	
receptor	receptors with positive cross-enrichment
GART	PPARg
DHFR	PPARg
class: other enzymes	
receptor	receptors with positive cross-enrichment
AChe	FXa
ALR2	GART, ACE, RXRa, PPARg, AmpC, COX-1, COX-2
AmpC	GART, ACE, RXRa, PPARg, ALR2, COX-1, COX-2
COX-1	ALR2, COX-2
COX-2	HSP90, ALR2, PARP
GPB	COMT
HIVPR	*none*
HIVTR	PNP
HMGR	RXRa, ACE, GART, ARL2,AmpC,COX-1
InhA	*none*
NA	PPARg, thrombin, trypsin, ADA
PARP	COX-1, PNP
PNP	TK, ADA, COMT, COX-1, GPB, PARP, SAHH
SAHH	TK, ADA, COMT, COX-1, PARP, GPB, PNP

[a]The 40 DUD receptors sorted in 6 biological classes. For a given receptor used as test set, all receptors within the same classification and in the list of reported cross-enrichment[21] are excluded from training and validation sets.

specificity [eqs 1 and 2] and enrichment factor (ef) [eq 3]. These metrics were chosen due to their popularity and acceptance in the field.[31]

$$\text{sensitivty} = \frac{\text{true positives}}{\text{total actives}} \quad (1)$$

$$\text{specificiy} = \frac{\text{true negatives}}{\text{total decoys}} \quad (2)$$

$$\text{ef}_{X\%} = \frac{\text{actives found at } X\%}{\text{molecules at } X\%} \frac{\text{total molecules}}{\text{total actives}} \quad (3)$$

To estimate the significance of the difference ΔX in a metric X between a pair of methods, with X being either the auc or ef, we compute the p-value on the average difference[31]

$$p = \frac{1}{2}\left\{1 - \text{erf}\left[\langle \Delta X \rangle \sqrt{\frac{N}{2\text{Var}(\Delta X)}}\right]\right\} \quad (4)$$

Where erf is the error function, N is the number of receptors in the DUD benchmark, and $\text{Var}(\Delta X)$ denotes the variance of ΔX.

3. RESULTS AND DISCUSSION

We have developed a novel method for improving virtual screening (VS) results called docking data feature analysis (DDFA). In this approach, all ligands are docked several times with different docking programs. Features derived from the full library of docked poses and scores are assessed by an artificial neural network (ANN) to identify potential active molecules. To test our approach, we used the Directory of Useful Decoys[21] (DUD), which is an established VS benchmark consisting of 40 receptors, each of them having its own screening library with a 1:36 active to decoy ratio. The DUD is a challenging test for receptor-based VS algorithms since the decoys were selected specifically to be similar to the active molecules of each receptor.[21] Each of the 40 DUD libraries were docked using three different programs: Autodock4.2[22] (AD4), Autodockvina[9] (ADV), and RosettaLigand3.4[23] (RL). Docking was conducted with a rigid receptor molecule in AD4 and ADV and with flexible receptor sidechains in RL. Two rankings were generated from each of the three data sets: (i) based on the docking score (RAW) and (ii) based on the novel feature score (DDFA). Additionally, RAW and DDFA rankings were generated by combining all three docking data sets (denoted as RAW-ALL and DDFA-ALL). In the following we compare the VS performance between the two ranking approaches (RAW and DDFA) applied to the four docking data sets (AD4, ADV, RL, ALL). To evaluate docking performance, we computed the area under the curve (auc) of the receiver operator characteristic (ROC) curve given by the various rankings. An AUC of 0.5 reflects a random selection, whereas a value of 1.0 reflects the perfect identification of active compounds. As a second performance measure, we computed the enrichment factor (ef), which compares the active-to-decoy ratio computed at a given cutoff rank.

Compared to all three individual docking programs, DDFA-ALL significantly improves the auc (Figure 3A, C, and E). Notably, DDFA-ALL yields performances above the random level (auc > 0.5) for all the receptors, with 30 of them registering auc values above 0.7 (Table 2). In contrast, RAW-AD4, RAW-ADV, and RAW-RL, yield good performance (auc > 0.7) only in 11, 15, and 15 cases, respectively (Table 3). Even more striking differences are observed in the number of receptors performing poorly (auc < 0.5); with 13, 7, 7, and 0 for RAW-AD4, RAW-ADV, RAW-RL, and DDFA-ALL, respectively. In line with these results, DDFA-ALL obtains an average auc of 0.77, which exceeds the corresponding values of RAW-AD4, RAW-ADV, and RAW-RL by 28%, 20%, and 18%. DDFA-ALL not only clearly outperformed the individual scoring programs in the auc metric but also in the enrichment factor (ef) (Figure 3B, D, and F). Within the first 20% of the ranking, DDFA-ALL's ef is around 50% larger than the efs of the conventionally obtained rankings. Taken together, these findings indicate that the DDFA-ALL is a robust method for evaluating VS experiments, not only because it effectively yields higher average performance in terms of auc and ef but also due to its strong reduction of poor performing receptors.

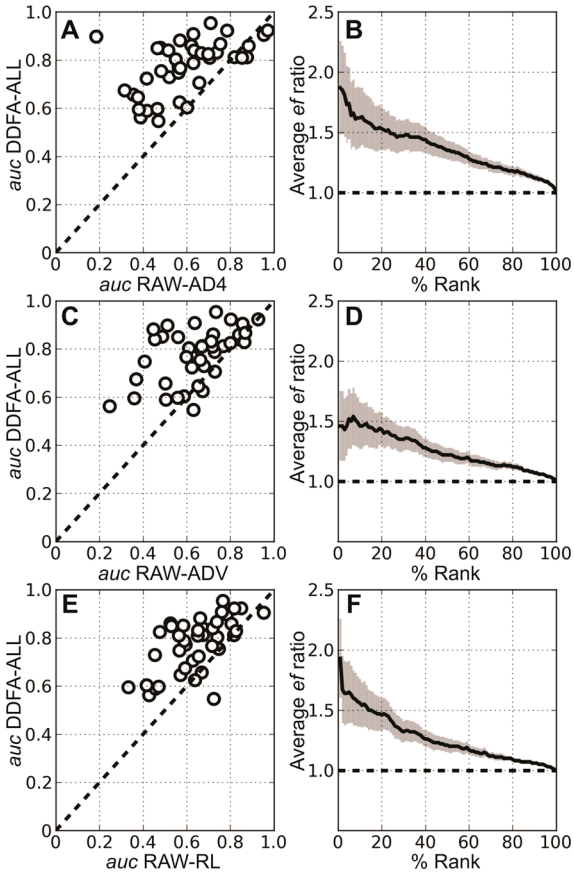


Figure 3. DDFA-ALL vs individual RAW rankings. DDFA-ALL is compared against RAW-AD4 (A, B), RAW-ADV (C, D), and RAW-RL (E, F). In plots comparing the auc (A, C, and E) the circles represent each of the 40 DUD receptors. Plots comparing ef (B, D, and F) show the DDFA-ALL to individual RAW average ratio. In all the plots, the dashed line indicates the limit where both methods perform equally.

Next we asked whether the remarkable gain in performance of DDFA-ALL stems from the feature-based analysis of the docking data or from the combination of complementary docking programs. With this objective, we applied the DDFA approach to the data from single docking programs. Interestingly, these individual versions of DDFA still outperform the RAW approach (Figure 4) yielding 28, 27, and 28 receptors with auc > 0.7 for DDFA-AD4, DDFA-ADV, and DDFA-RL, respectively, which has to be compared to the 11, 15, and 15 cases of good performance for RAW approaches

dx.doi.org/10.1021/ci500028u | J. Chem. Inf. Model. 2014, 54, 1401−1411

Table 2. DDFA Rankings Metrics[a]

	AD4				ADV				RL				ALL			
	ef_{max}	$ef_{2\%}$	$ef_{20\%}$	auc	ef_{max}	$ef_{2\%}$	$ef_{20\%}$	auc	ef_{max}	$ef_{2\%}$	$ef_{20\%}$	auc	ef_{max}	$ef_{2\%}$	$ef_{20\%}$	auc
average	12.9	9.6	2.8	0.74	13.7	10.3	2.8	0.75	13.1	9.7	3.0	0.76	13.5	10.3	3.0	0.77
conf 95%	3.0	2.2	0.3	0.04	3.0	2.2	0.3	0.03	2.4	2.0	0.3	0.03	2.6	2.0	0.3	0.03
ACE	1.8	0.0	1.7	0.57	4.3	2.1	2.0	0.63	2.9	0.0	2.7	**0.70**	2.3	0.0	1.9	0.57
AChE	19.3	11.9	2.6	0.71	21.3	18.1	3.8	**0.88**	20.1	12.8	2.2	0.67	4.7	3.3	2.2	0.73
ADA	1.0	0.0	0.4	0.40	8.3	6.5	1.7	0.60	13.7	9.1	2.1	**0.70**	5.5	2.6	1.7	0.64
ALR2	20.4	16.3	3.4	0.81	31.4	27.5	4.0	**0.87**	23.5	17.7	3.3	0.74	19.6	15.7	3.9	0.85
AmpC	4.8	4.8	1.2	0.56	4.8	2.4	0.7	0.55	4.8	2.4	1.2	**0.57**	4.8	2.4	0.7	0.51
AR	7.3	5.8	2.7	0.78	23.3	17.5	4.0	**0.88**	5.1	3.2	1.4	0.54	9.7	9.7	3.9	0.82
CDK2	14.4	10.8	3.4	0.83	21.9	16.8	3.1	**0.81**	11.7	8.0	2.9	0.72	13.1	10.2	3.2	0.80
COMT	7.6	0.0	3.2	0.80	43.5	29.0	3.7	**0.88**	4.8	4.8	2.3	0.66	23.9	10.6	4.0	0.87
COX-1	8.7	8.7	2.9	0.75	29.1	16.6	3.8	**0.83**	12.5	10.4	3.2	0.71	24.9	14.5	3.6	0.82
COX-2	10.6	9.8	3.0	0.78	24.7	20.2	3.7	0.84	4.8	4.2	3.3	0.80	13.5	13.0	3.9	**0.86**
DHFR	17.2	12.8	4.0	0.89	12.1	9.4	3.0	0.79	19.5	18.6	4.6	**0.95**	19.5	15.9	4.2	0.91
EGFr	2.0	1.7	1.8	0.67	7.3	7.3	2.4	0.75	13.7	11.7	3.5	**0.84**	11.8	8.9	3.1	0.80
ERagonist	15.1	14.4	3.4	0.71	12.5	12.1	3.4	0.84	15.1	9.1	3.7	0.84	18.2	18.2	4.3	**0.92**
ERantagonist	27.2	21.0	3.3	**0.84**	6.9	6.6	3.2	0.77	19.1	15.8	3.2	0.84	8.2	6.6	3.3	0.79
FGFrl	7.5	7.5	2.0	0.62	5.9	5.0	2.5	**0.73**	2.1	1.3	1.6	0.61	4.2	2.9	1.6	0.57
FXa	9.2	5.6	1.6	0.58	5.6	4.5	1.9	0.62	11.2	6.9	2.3	**0.65**	7.1	5.0	1.6	0.61
GART	2.5	0.0	2.4	0.74	2.6	2.6	1.8	0.68	12.8	11.5	3.5	**0.84**	15.3	7.7	3.4	0.76
GPB	8.2	5.0	1.4	0.60	8.0	5.9	3.2	0.81	18.1	14.7	3.9	0.83	22.5	19.0	4.3	**0.89**
GR	14.2	11.6	2.8	**0.76**	10.5	7.2	1.3	0.65	2.6	1.3	1.2	0.55	4.5	3.9	1.5	0.61
HIVPR	20.3	14.0	2.6	0.72	18.6	14.9	2.9	0.74	16.9	13.2	3.4	**0.80**	15.5	10.9	2.6	0.76
HIVRT	8.4	8.4	2.1	0.57	16.9	10.5	2.1	0.66	12.1	5.9	1.8	0.61	9.7	9.4	2.3	**0.66**
HMGR	7.2	5.8	3.6	0.80	8.7	7.2	1.7	0.64	20.2	11.5	3.2	**0.86**	11.5	5.8	3.6	0.82
HSP90	1.1	0.0	0.8	0.47	3.3	1.4	2.7	0.68	5.5	2.7	1.8	0.67	1.4	0.0	1.4	**0.62**
InhA	25.1	18.2	3.6	0.80	34.6	21.2	3.5	0.81	29.5	17.3	3.2	0.76	23.4	17.5	3.8	**0.84**
MR	38.7	21.4	3.9	**0.88**	16.0	10.0	4.3	0.87	21.7	20.0	4.0	0.82	38.7	21.4	3.6	0.84
NA	8.0	8.0	2.4	0.71	8.3	7.2	2.6	0.68	5.2	5.2	3.7	**0.80**	16.0	11.5	2.5	0.74
P38MAP	2.8	2.8	1.7	0.62	8.0	7.2	3.1	**0.77**	3.1	2.4	2.0	0.67	4.5	3.7	2.5	0.71
PARP	40.7	28.7	4.7	0.95	9.4	7.5	4.4	0.89	33.5	26.4	4.9	**0.96**	31.3	25.7	4.6	0.95
PDE5	15.3	14.3	3.7	0.83	22.3	16.0	3.6	0.83	20.0	16.0	4.3	0.89	21.4	18.5	4.1	**0.90**
PDGFrb	11.9	9.8	3.4	**0.81**	5.3	4.2	1.6	0.66	12.5	9.5	3.3	0.80	5.4	5.1	2.4	0.68
PNP	8.8	7.2	3.2	0.75	13.0	9.3	3.4	0.75	10.9	9.3	2.8	**0.81**	10.9	9.3	3.1	0.77
PPARg	7.8	7.0	2.2	0.72	7.1	7.1	2.6	0.70	9.4	8.9	3.5	**0.82**	5.2	3.8	2.2	0.76
PR	13.3	6.3	4.6	**0.90**	2.6	1.9	2.0	0.67	11.3	11.3	3.0	0.77	19.0	19.0	4.0	0.85
RXRa	33.0	25.6	5.0	**0.96**	17.9	17.9	4.8	0.94	16.5	10.3	4.8	0.93	20.5	20.5	4.3	0.91
SAHH	2.0	0.0	1.5	0.53	3.8	1.6	2.3	0.72	25.7	24.7	4.3	**0.90**	16.1	15.5	3.8	0.80
SRC	20.4	18.0	4.0	**0.87**	12.1	11.1	3.4	0.82	7.0	5.1	2.9	0.77	9.6	8.8	3.0	0.83
thrombin	12.6	12.6	3.4	0.82	23.9	14.7	3.3	**0.85**	15.4	10.5	3.3	0.82	14.0	11.2	3.4	0.84
TK	2.9	0.0	2.5	**0.66**	1.8	0.0	0.7	0.62	3.0	0.0	1.6	0.54	4.6	2.3	2.3	0.61
trypsin	17.8	13.4	4.0	**0.88**	10.3	8.2	2.5	0.72	4.1	2.1	2.7	0.74	13.4	11.1	3.3	0.76
VEGFr2	19.2	13.6	2.9	0.77	22.5	14.0	2.9	0.76	21.4	14.0	3.2	**0.79**	16.8	11.8	2.9	0.77

[a]Enrichment factor (ef) at 2%, 20%, and maximal reached, in addition to the ROC auc. The bold values indicate the highest auc value achieved in the given receptor.

reported above. Also the number of receptors with auc < 0.5 remains low; the only two observed cases are angiotensin converting enzyme (ACE) and heat shock protein 90 (HSP90), for which auc of 0.40 and 0.47, respectively, are obtained with DDFA-AD4 (Table 2). The average auc values, for DDFA-AD4, ADDF-ADV, and ADDF-RL, are 0.74, 0.75, and 0.76, respectively. These results still correspond to improvements of 23%, 17%, and 17% with respect to their RAW counterparts. Also the ef improves with the DDFA individual versions (Figure 4B, D, and F). For DDFA-AD4 and DDFA-RL, the ef is around 50% larger than that of the corresponding RAW version over the first 10% of the ranking, whereas for DDFA-ADV this degree of improvement is just observed at the starting point of the ranking. This analysis confirms the robustness of the DDFA approach, since a significant enhancement in performance is already obtained even when information from a single docking program only is used.

The above-mentioned observation suggests that some part of the performance gain in DDFA-ALL stems from the combination of different docking programs. To assess this influence, the RAW rankings of the docking programs were combined in to a single list, RAW-ALL. Indeed, the RAW-ALL ranking also outperforms individual RAW rankings (Figure 5A, C, and E), although to a lesser extent than the DDFA-ALL (Figure 5G). In the RAW-ALL approach, 21 proteins reported auc values above 0.70; which exceeds the 11, 15, and 15 of these cases for RAW-AD4, RAW-ADV, and RAW-RL, respectively; but, it is still inferior to the 30 cases for DDFA-

dx.doi.org/10.1021/ci500028u | J. Chem. Inf. Model. 2014, 54, 1401−1411

Table 3. RAW rankings metrics[a]

	AD4				ADV				RL				ALL			
	ef_{max}	$ef_{2\%}$	$ef_{20\%}$	auc	ef_{max}	$ef_{2\%}$	$ef_{20\%}$	auc	ef_{max}	$ef_{2\%}$	$ef_{20\%}$	auc	ef_{max}	$ef_{2\%}$	$ef_{20\%}$	auc
average	7.8	5.6	2.0	0.60	9.7	7.1	2.1	0.64	7.7	6.2	2.1	0.65	9.3	7.4	2.6	0.70
conf 95%	2.2	1.9	0.4	0.06	2.7	2.3	0.3	0.05	1.7	1.6	0.3	0.04	2.1	2.0	0.4	0.05
ACE	2.1	0.0	1.0	**0.38**	1.3	0.0	0.8	0.36	1.0	0.0	0.5	0.33	1.2	0.0	1.2	0.35
AChE	1.9	1.9	1.3	0.52	5.4	4.8	2.8	**0.68**	1.0	0.0	0.7	0.45	2.9	2.4	2.3	0.61
ADA	1.0	0.0	0.0	0.36	1.2	0.0	0.9	0.50	2.3	0.0	1.3	**0.67**	1.5	0.0	0.6	0.57
ALR2	2.9	0.0	2.6	0.62	9.8	3.9	2.7	0.72	5.9	5.9	1.5	0.53	6.2	2.0	2.7	**0.73**
AmpC	1.1	0.0	0.2	0.39	1.0	0.0	0.2	0.25	1.0	0.0	0.5	**0.43**	1.0	0.0	0.2	0.30
AR	9.0	7.0	2.7	0.70	19.4	18.8	3.7	0.80	5.1	3.2	1.5	0.48	15.4	14.7	3.9	**0.84**
CDK2	8.6	4.3	2.1	0.56	13.1	10.9	2.1	0.62	4.4	2.9	1.6	0.60	8.6	6.5	2.5	**0.63**
COMT	1.4	0.0	1.4	0.48	32.6	14.5	1.4	0.49	4.8	4.8	0.9	**0.58**	10.9	9.7	1.8	0.55
COX-1	2.2	2.2	0.8	0.52	16.6	10.4	3.8	**0.84**	6.2	6.2	1.6	0.67	12.5	12.5	3.6	0.83
COX-2	7.8	7.3	2.8	0.75	25.6	23.0	4.0	0.87	4.0	3.3	2.6	0.74	18.8	17.3	4.3	**0.90**
DHFR	17.7	14.7	4.8	**0.95**	11.1	8.9	3.5	0.86	17.5	17.4	4.7	0.95	14.5	12.5	4.7	0.94
EGFr	1.7	1.5	1.2	0.55	2.5	1.1	1.7	0.61	7.0	6.0	2.6	**0.74**	4.9	3.6	2.2	0.72
ERagonist	21.2	18.2	3.1	0.79	18.2	18.2	3.3	0.80	16.7	12.1	3.6	0.82	18.2	15.9	4.6	**0.93**
ERantagonist	21.8	19.7	3.5	**0.82**	13.6	9.2	2.2	0.67	13.6	11.8	2.3	0.67	13.2	13.2	3.3	0.77
FGFrl	1.0	0.8	0.7	0.42	1.2	0.4	0.9	**0.50**	1.0	0.0	0.8	0.46	1.0	0.0	0.7	0.46
FXa	2.1	1.8	1.2	0.57	2.4	2.4	1.8	**0.67**	9.1	6.6	1.9	0.64	5.5	5.5	1.8	0.66
GART	4.4	1.3	4.0	**0.88**	2.9	0.0	2.6	0.77	5.6	5.1	3.5	0.82	3.9	1.3	3.9	0.85
GPB	1.0	0.0	0.1	0.19	2.9	2.9	1.2	0.51	10.4	9.8	3.5	0.78	6.3	2.9	3.2	**0.82**
GR	6.5	5.8	2.1	**0.60**	7.9	5.2	1.4	0.59	1.3	0.7	0.6	0.42	9.1	6.5	1.6	0.57
HIVPR	5.1	4.1	2.5	0.66	5.8	5.8	2.7	0.73	6.8	5.8	2.3	0.63	6.5	4.8	2.6	**0.73**
HIVRT	2.5	2.4	0.7	0.38	12.1	7.0	1.9	**0.65**	7.3	5.9	1.9	0.57	9.7	7.0	2.0	0.63
HMGR	3.9	2.9	1.7	0.63	1.2	0.0	0.7	0.45	2.9	1.4	0.9	**0.72**	1.9	1.4	1.4	0.62
HSP90	1.2	0.0	0.4	0.47	1.7	0.0	1.2	0.63	2.4	0.0	2.4	**0.72**	1.6	0.0	1.0	0.64
InhA	22.7	13.5	1.9	0.47	19.1	12.4	1.8	**0.56**	10.3	5.1	1.6	0.53	15.4	11.1	2.2	0.53
MR	15.5	10.7	4.6	**0.88**	23.3	23.3	4.0	0.84	16.7	16.7	3.3	0.81	21.7	16.7	4.0	0.84
NA	2.3	1.2	0.9	0.56	1.1	0.0	0.4	0.41	4.1	3.1	1.6	**0.57**	2.1	2.1	0.8	0.49
P38MAP	1.6	1.2	0.7	0.42	3.1	2.0	2.3	0.62	3.6	2.0	1.9	**0.65**	2.4	2.3	2.0	0.63
PARP	25.1	21.1	2.7	0.71	9.4	4.5	2.8	0.73	15.2	13.2	3.1	0.77	15.2	14.7	4.3	**0.91**
PDE5	11.7	6.9	2.3	0.63	11.7	8.0	2.0	0.64	10.6	9.2	2.6	**0.76**	15.3	9.7	2.9	0.75
PDGFrb	7.7	4.7	0.9	0.32	5.3	3.0	0.6	0.37	4.2	3.6	1.7	**0.59**	6.5	3.5	1.3	0.50
PNP	6.1	3.1	2.4	0.63	4.8	4.1	2.6	0.73	5.6	3.1	1.9	0.59	4.2	2.1	3.1	**0.79**
PPARg	1.2	0.0	1.1	0.48	4.7	3.5	1.8	0.66	7.1	5.9	2.7	**0.75**	3.5	1.8	1.9	0.63
PR	13.3	8.4	1.7	0.57	1.9	0.0	1.1	0.45	19.7	15.0	1.9	0.66	15.8	9.4	3.2	**0.81**
RXRa	16.5	15.4	5.0	0.97	33.0	28.2	4.3	0.93	17.9	17.9	4.0	0.85	28.2	28.2	5.0	**0.98**
SAHH	2.6	0.0	2.1	0.67	22.5	20.1	3.5	**0.86**	17.0	17.0	3.5	0.83	13.3	12.4	3.3	0.81
SRC	14.0	10.7	2.8	0.70	5.0	4.1	2.4	0.72	7.6	4.4	2.0	0.65	12.1	10.1	3.9	**0.85**
thrombin	8.4	7.0	2.8	0.74	11.2	8.4	3.0	0.71	7.0	7.0	2.4	0.65	11.2	9.1	3.1	**0.79**
TK	1.5	0.0	0.9	0.47	1.6	0.0	0.7	0.56	4.6	2.3	1.1	0.47	2.3	2.3	1.1	**0.57**
trypsin	17.8	14.5	3.7	**0.85**	8.2	4.1	2.5	0.67	3.1	3.1	1.7	0.56	8.2	6.2	3.5	0.83
VEGFr2	15.6	8.9	1.9	0.57	14.2	8.7	2.1	0.60	16.6	9.9	2.7	**0.72**	20.2	13.4	2.8	0.70

[a]Enrichment factor (ef) at 2%, 20%, and maximal reached, in addition to the ROC auc. The bold values indicate the highest auc value achieved in the given receptor.

ALL. On the side of poor-performing receptors (auc < 0.5), their number is reduced to four, which certainly improves with respect to the individual RAW rankings, but not in comparison with DDFA-ALL, with its zero cases with auc < 0.5. An equivalent picture is observed with the ef metrics. RAW-ALL outperforms the individual rankings (Figure 5B, D, and F), but not DDAF-ALL where RAW-ALL is at least 20% smaller over the initial 10% of the ranking (Figure 5H). These results provide evidence on the beneficial effect that is obtained from the combination of three docking data sources.

To evaluate the significance of the observed differences between methods in the performance metrics auc and $ef_{2\%}$, we computed their *p*-value[31] (Methods). The comparison of the four different versions of RAW and DDFA yields remarkably low *p*-values ($<1 \times 10^{-3}$; Table4A). The further improvement in auc achieved by DDFA-ALL with respect to DDFA-AD4, DDFA-ADV, and DDFA-RL is confirmed by the low *p*-values, 0.02, 0.04, and 0.07, respectively (Table 4B). In contrast, DDFA-ALL does not yield significantly better $ef_{2\%}$ than the individual versions of DDFA (Table 4B). Taken together, DDFA is significantly better than RAW in both metrics, whereas DDFA-ALL outperforms the individual versions of DDFA only in the auc metric.

After confirming the significance of the results yielded by DDFA, we wanted to assess their stability with respect to the number of receptors used during the training and validation process. The systematic reductions of receptors used for training causes a gradual decay in performance for all four

dx.doi.org/10.1021/ci500028u | J. Chem. Inf. Model. 2014, 54, 1401−1411

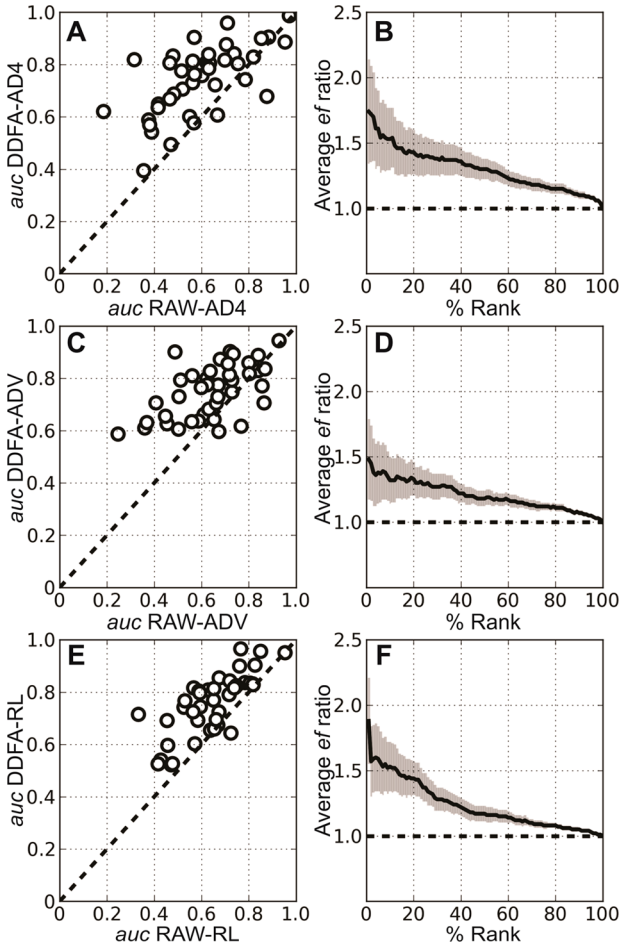


Figure 4. Individual DDFA vs individual RAW rankings. Individual versions of both, DDFA and RAW, are compared for AD4 (A, B), ADV (C, D), and RL (E, F). In plots comparing the auc (A, C, and E) the circles represent each of the 40 DUD receptors. Plots comparing ef (B, D, and F) show the individual DDFA to individual RAW average ratio. In all the plots, the dashed line indicates the limit where both methods perform equally.

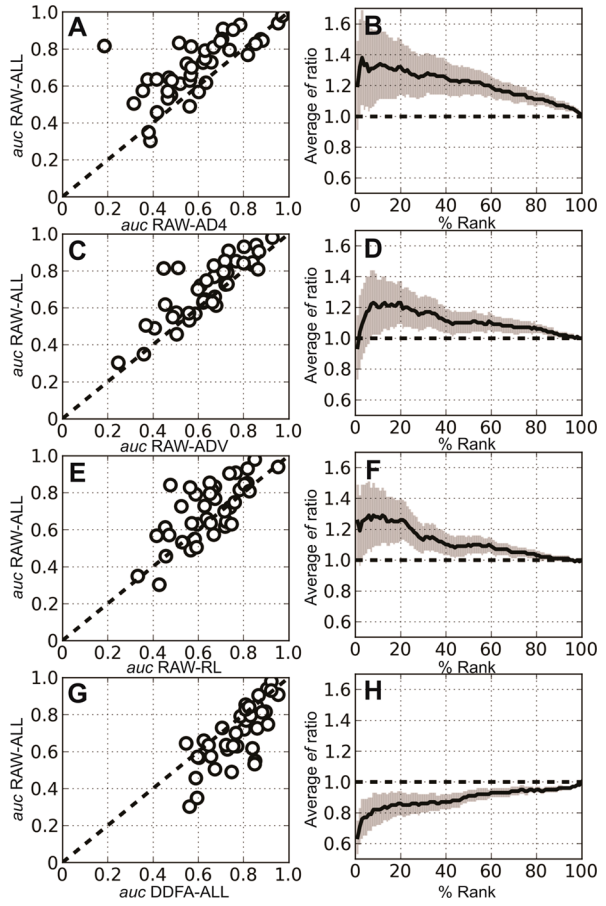


Figure 5. RAW-ALL vs individual RAW and DDFA-ALL rankings. RAW-ALL is compared against RAW-AD4 (A, B), RAW-ADV (C, D), RAW-RL (E, F), and DDFA-ALL (G, H), respectively. In plots comparing the auc (A, C, E, and G) the circles represent each of the 40 DUD receptors. Plots comparing ef show the RAW-ALL to individual RAW average ratio (B, D, and F), and the RAW-ALL to DDFA-ALL average ratio. In all the plots, the dashed line indicates the limit where both methods perform equally.

DDFA cases (Figure 6), as expected when using less training data. Nevertheless, DDFA performance is always at least as good as the RAW performance (Figure 6; Table 3), such that one can say with confidence that DDFA is robust in the sense, that it never hurts average performance to use it. Moreover, we should note that in this test we did not reoptimize the hyperparameters that control training for each number of receptors such that one might be able to improve performance by hyperparameter tuning (Supporting Information Figure S5). The result in Figure 6 also shows that the method has potential to achieve an even better performance than demonstrated here, if more receptors than 22 were available for training.

In addition to benchmarking the DDFA approach on DUD, our study also provides valuable insight into the VS performance of the individual docking programs (Table 3). On average, RAW-ADV yielded better results than RAW-AD4 and RAW-RL. The superior performance of RAW-ADV over RAW-AD4 in not surprising, since it matches with previously reported observations.[9,32,33] This is the first time, however, that results for RL obtained on the DUD benchmark were published. RL obtained averages values of 0.65 and 6.19 in auc and $ef_{2\%}$, respectively, thereby yielding similar results to ADV and better than AD4 (Table 3). This result is in line with the outstanding performances that RAW-RL obtained in pose recovery benchmarks.[10,27,34] Table 3 shows that for three receptors none of the docking programs reached auc values above the random level: (i) angiotensin converting enzyme (ACE), (ii) Amp-C beta lactamase (AmpC), and (iii) fibroblast growth factor receptor 1 (FGFr1). These are the three receptors for which DDFA-ALL yielded also the poorest results with auc of 0.57, 0.51, and 0.57 for ACE, AmpC, and FGFr1, respectively. This observation suggest that the improvement produced by DDFA-ALL is somewhat limited by the quality of the individual docking results. Another interesting example is the platelet derived growth factor receptor (PDGFrb), a receptor for which another seven different scoring functions report auc values under 0.5.[35] In our hands, the aucs yielded by RAW-AD4 and RAW-ADV for PDGFrb are also below 0.5, whereas RAW-RL obtains an auc of 0.59. In contrast, the performances of the individual version of DDFA are undoubtedly better; 0.81, 0.66, and 0.80 for DFFA-AD4, DDFA-ADV, and DDFA-RL, respectively.

As shown above DDFA represents a highly attractive alternative to traditional ranking approaches for analyzing VS experiments. This finding is also supported by comparing DDFA performances with those found in the literature (Table

Table 4. *p*-Values of the Difference in Metrics (auc, $ef_{2\%}$) between Each Pair of Methods[a]

	AD4		ADV		RL		ALL	
(A) Significance of the Difference in Performance between RAW and DDFA								
	RAW	DDFA	RAW	DDFA	RAW	DDFA	RAW	DDFA
	(0.60, 5.6)	*(0.74, 9.5)*	*(0.64, 7.1)*	*(0.75, 10.3)*	*(0.65, 6.2)*	*(0.76, 9.7)*	*(0.70, 7.4)*	*(0.77, 10.3)*
p-value in auc	$<1 \times 10^{-5}$		$<1 \times 10^{-5}$		$<1 \times 10^{-5}$		$<1 \times 10^{-5}$	
p-value in $ef_{2\%}$	$<1 \times 10^{-5}$		3×10^{-4}		$<1 \times 10^{-5}$		3×10^{-4}	

	DDFA		DDFA		DDFA	
(B) Significance of the Difference in Performance between DDFA-ALL and the Individual Versions of DDFA						
	AD4	ALL	ADV	ALL	RL	ALL
	(0.74, 9.5)	*(0.77, 10.3)*	*(0.75, 10.3)*	*(0.77, 10.3)*	*(0.76, 9.7)*	*(0.77, 10.3)*
p-value in auc	0.02		0.04		0.07	
p-value in $ef_{2\%}$	0.27		0.59		0.42	

[a]The lower the *p*-value, the more significant the differences in performance.

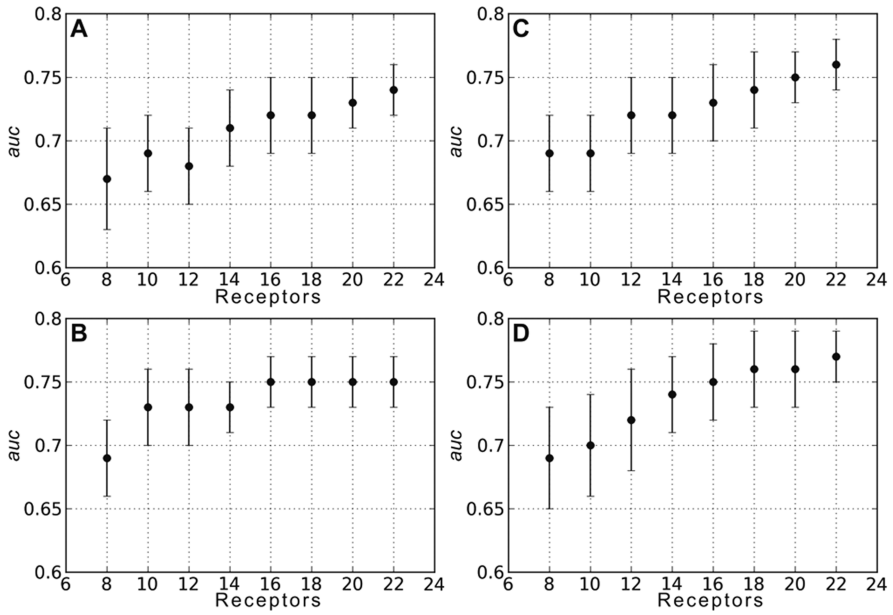


Figure 6. Average auc on the DUD benchmark yielded by (A) DDFA-AD4, (B) DDFA-ADV, (C) DDFA-RL, and (D) DDFA-ALL, with a different number of training receptors. The plotted values correspond to the average over five independent runs using a different subset of receptors. Error bars correspond to the associated standard deviations.

5). Considering structured-based methodologies tested on the DUD benchmark, the different versions of the DDFA approach (ALL, AD4, ADV, and RL) obtained performances that situate them among the best methods available. Certainly, the commercial docking software, ICM and Glide SP, achieve the top performances in the auc and $ef_{2\%}$ metrics, respectively. Nonetheless, their corresponding performances fall within the 95% confidence limits of DDFA-ALL; auc 0.77 ± 0.03 and ef 10.3 ± 2.0 (Table 2). One of the best methods we also found is the methodology developed by Durrant et al.[37] in which NNScore[38] is used. This methodology resembles ours in the sense that it combines academic docking software with an artificial neutral network. However, while NNScore is trained on the characteristic interactions of protein−ligand complexes, thus proposing an interaction rescoring scheme, our DDFA is trained on the characteristic features of the docking data associated with active molecules, thereby representing a reranking scheme. Additionally our DDFA approach also yields high ef values at 2%, which, together with the averaged ef curves presented previously (for example Figure 3), provide confidence on the performance stability that our approach has on this metric. These findings, together with the inherent flexibility of DDFA (easily extended to combine several docking programs and docking features), render our novel approach as highly attractive for analyzing VS experiments.

3. CONCLUSION

The DDFA approach introduced in this work was able to improve considerably the selection of active compounds from the output of popular docking programs. This was achieved by extending the analysis of the docking data beyond the traditional docking score. Although the usefulness of rescoring, consensus rankings, and machine learning methods has already been noted,[39−41] what distinguishes our study is that we could convincingly show a possible way to combine all these elements together synergistically. It must be emphasized, however, that the success on combining several docking features and/or scoring programs resides in their diversity.[42,43] Each element should account for different characteristics that contribute to the active-decoy discrimination. Although establishing the

Table 5. Reported Performances on DUD[a]

methodologies	auc	$ef_{2\%}$
ICM [ref 36][b]	0.79	
AutodockVina-NN1 [ref 37]	0.78	
Glide SP [ref 35][b]	0.77	12.2
DDFA-ALL	**0.77**	**10.3**
DDFA-RL	**0.76**	**9.7**
AutodockVina-NN2 [ref 37]	0.76	
DDFA-ADV	**0.75**	**10.3**
normalization score [ref 16][c,d,e]	0.75	
DDFA-AD4	**0.74**	**9.5**
Glide HTVS [ref 37]	0.73	
Surflex [ref 35]	0.72	12.0
Glide HTVS [ref 35]	0.72	10.7
ICM [ref 36]	0.71	
RAW-ALL	**0.70**	**7.4**
Autodock Vina [ref 37]	0.70	
eHiTS [ref 16][d,e]	0.70	
Glide SP [ref 16][d,e]	0.70	
Surflex [ref 35][b]	0.66	7.9
RosettaLigand	**0.65**	**6.2**
AutodockVina	**0.64**	**7.1**
NNScore 1.0 [ref 38][d]	0.64	
ICM [ref 35]	0.63	8.0
FlexX [ref 35]	0.61	7.2
Autodock4.2	**0.60**	**5.6**
PhDock [ref 35]	0.59	7.7
NNScore 2.0 [ref 32][d]	0.59	
AutodockVina [ref 32][d]	0.58	
Dock [ref 35]	0.55	8.2
$Autodock_{fast}$ [ref 32][d]	0.51	
$Autodock_{rigorous}$ [ref 32][d]	0.50	

[a]Methodologies reported in this work are highlighted in bold letters. [b]Tuned by expert knowledge. [c]Computational expensive. [d]Subset of the DUD receptors. [e]Subset of decoys used.

optimal docking feature selection for a given set of scoring programs is a challenging task, it certainly opens a pathway to possible further improvements.

In terms of the well-established virtual screening metrics, auc and ef, DDFA performance is statistically similar to that reported by commercial software under expert intervention[35,36] or by methods that increase the computational cost by 2 orders of magnitude.[16] Additionally, DDFA shows an excellent stability in its results and, in strong contrast to simple ranking schemes, performs better than random selection for every single receptor in the DUD benchmark. Overall, DDFA represents a new, simple, and automatic reranking treatment that not only is easy to implement and extend to other docking software or docking data features but also provides high VS performance with minimal extra computing time.

■ ASSOCIATED CONTENT

Supporting Information

Benchmark of the parameters used to setup the ANN of DDFA. This material is available free of charge via the Internet at http://pubs.acs.org. The scripts used to evaluate the ligands with DDFA are available upon request.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: castro@biochem.mpg.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work has been supported by the CONACYT-DAAD scholarship 209523 (M.A.) and the DFG grant LA 1817/3-1 (O.F.L.)

■ ABBREVIATIONS

AD4, Autodock4; ADV, AutodockVina; RL, Rosetta Ligand; ANN, artificial neural network; DDFA, docking data feature analysis

■ REFERENCES

(1) Khanna, I. Drug Discovery in Pharmaceutical Industry: Productivity Challenges and Trends. *Drug Discovery Today* **2012**, *17*, 1088−1102.

(2) Cavasotto, C. N.; Orry, A. J. W. Ligand Docking and Structure-based Virtual Screening in Drug Discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1006−1014.

(3) Tanrikulu, Y.; Krüger, B.; Proschak, E. The Holistic Integration of Virtual Screening in Drug Discovery. *Drug Discovery Today* **2013**, *18*, 358−364.

(4) Kar, S.; Roy, K. How Far Can Virtual Screening Take us in Drug Discovery? *Expert Opin. Drug Discovery* **2013**, *8*, 245−261.

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

(6) Drwal, M. N.; Griffith, R. Combination of Ligand- and Structure-Based Methods in Virtual Screening. *Drug Discovery Today: Technologies* **2013**, *10*, e395−e401.

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

(8) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(9) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(10) Davis, I. W.; Raha, K.; Head, M. S.; Baker, D. Blind Docking of Pharmaceutically Relevant Compounds Using RosettaLigand. *Protein Sci.* **2009**, *18*, 1998−2002.

(11) Zhong, S.; Zhang, Y.; Xiu, Z. Rescoring Ligand Docking Poses. *Curr. Opin. Drug Discovery Dev.* **2010**, *13*, 326−34.

(12) Rajamani, R.; Good, A. C. Ranking Poses in Structure-Based Lead Discovery and Optimization: Current Trends in Scoring Function Development. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 308−15.

(13) Brewerton, S. C. The Use of Protein-Ligand Interaction Fingerprints in Docking. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 356−64.

(14) García-Sosa, A. T.; Hetényi, C.; Maran, U. Drug Efficiency Indices for Improvement of Molecular Docking Scoring Functions. *J. Comput. Chem.* **2010**, *31*, 174−184.

(15) Lee, J.; Seok, C. A Statistical Rescoring Scheme for Protein−Ligand Docking: Consideration of Entropic Effect. *Proteins* **2008**, *70*, 1074−1083.

(16) Wallach, I.; Jaitly, N.; Nguyen, K.; Schapira, M.; Lilien, R. Normalizing Molecular Docking Rankings Using Virtually Generated Decoys. *J. Chem. Inf. Model.* **2011**, *51*, 1817−1830.

(17) Feher, M. Consensus Scoring for Protein-Ligand Interactions. *Drug Discovery Today* **2006**, *11*, 421−428.
(18) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus Scoring for Ligand/Protein Interactions. *J. Mol. Graph. Model.* **2002**, *20*, 281−295.
(19) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134−1146.
(20) Jacobsson, M.; Lidén, P.; Stjernschantz, E.; Boström, H.; Norinder, U. Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data. *J. Med. Chem.* **2003**, *46*, 5781−5789.
(21) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(22) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Autodock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexiblity. *J. Comput. Chem.* **2009**, *16*, 2785−91.
(23) Meiler, J.; Baker, D. ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins* **2006**, *65*, 538−548.
(24) Lemmon, G.; Meiler, J. Rosetta Ligand Docking with Flexible XML Protocols. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Springer: New York, 2012; Vol. *819*, pp 143−155.
(25) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(26) Fleishman, S. J.; Leaver-Fay, A.; Corn, J. E.; Strauch, E. M.; Khare, S. D.; Koga, N.; Ashworth, J.; Murphy, P.; Richter, F.; Lemmon, G.; Meiler, J.; Baker, D. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLoS One* **2011**, *6*, e20161.
(27) Davis, I. W.; Baker, D. ROSETTALIGAND Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* **2009**, *385*, 381−392.
(28) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33.
(29) Dreiseitl, S.; Ohno-Machado, L. Logistic Regression and Artificial Neural Network Classification Models: a Methodology Review. *J. Biomed. Inform.* **2002**, *35*, 352−359.
(30) Schaul, T.; Bayer, J.; Wierstra, D.; Sun, Y.; Felder, M.; Sehnke, F.; Rückstieß, T.; Schmidhuber, J. PyBrain. *J. Mach. Learn. Res.* **2010**, *11*, 743−746.
(31) Nicholls, A. What Do We Know and When Do We Know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.
(32) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor−Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51*, 2897−2903.
(33) Chang, M. W.; Ayeni, C.; Breuer, S.; Torbett, B. E. Virtual Screening for HIV Protease Inhibitors: A Comparison of AutoDock 4 and Vina. *PLoS One* **2010**, *5*, e11955.
(34) Kaufmann, K. W.; Meiler, J. Using RosettaLigand for Small Molecule Docking into Comparative Models. *PLoS One* **2012**, *7*, e50769.
(35) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(36) Neves, M. A.; Totrov, M.; Abagyan, R. Docking and Scoring with ICM: The Benchmarking Results and Strategies for Improvement. *J. Comput. Aided Mol. Des.* **2012**, *26*, 675−686.
(37) Durrant, J. D.; Friedman, A. J.; Rogers, K. E.; McCammon, J. A. Comparing Neural-Network Scoring Functions and the State of the Art: Applications to Common Library Screening. *J. Chem. Inf. Model.* **2013**, *53*, 1726−1735.
(38) Durrant, J. D.; McCammon, J. A. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1865−1871.
(39) Houston, D. R.; Walkinshaw, M. D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *J. Chem. Inf. Model.* **2013**, *53*, 384−390.
(40) Planesas, J. M.; Claramunt, R. M.; Teixidó, J.; Borrell, J. I.; Pérez-Nueno, V. I. Improving VEGFR-2 Docking-Based Screening by Pharmacophore Postfiltering and Similarity Search Postprocessing. *J. Chem. Inf. Model.* **2011**, *51*, 777−787.
(41) Teramoto, R.; Fukunishi, H. Supervised Consensus Scoring for Docking and Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 526−534.
(42) Hsu, D. F.; Chung, Y.-S.; Kristal, B. S., Combinatorial Fusion Analysis: Methods and Practice of Combining Multiple Scoring Systems. In *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*; IGI Global, 2008; pp 1157−1181.
(43) Hsu, D. F.; Kristal, B.; Schweikert, C., Rank-Score Characteristics (RSC) Function and Cognitive Diversity. In *Brain Informatics*; Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., Huang, J., Eds.; Springer: Berlin and Heidelberg, 2010; Vol. *6334*, pp 42−54.