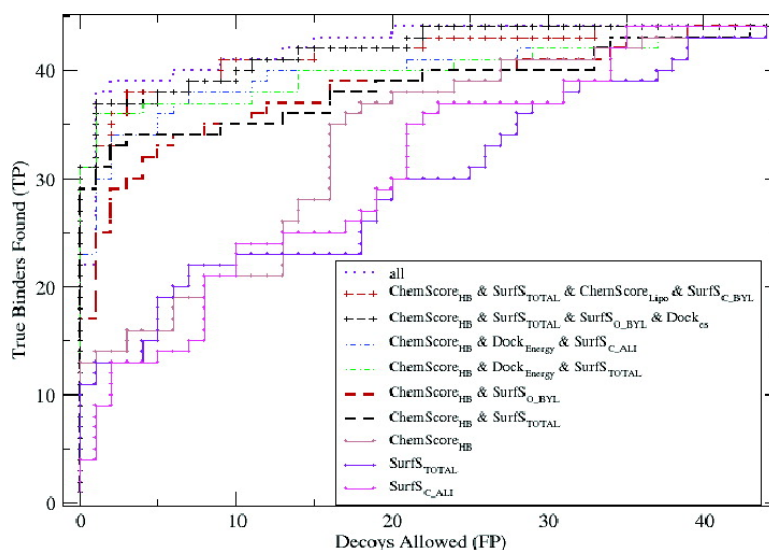


PostDOCK: A Structural, Empirical Approach to Scoring Protein Ligand Complexes

Clayton Springer, Helgi Adalsteinsson, Malin M. Young, Philip W. Kegelmeyer, and Diana C. Roe

J. Med. Chem., **2005**, 48 (22), 6821-6831 • DOI: 10.1021/jm0493360 • Publication Date (Web): 30 September 2005

Downloaded from <http://pubs.acs.org> on March 29, 2009



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 4 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

[View the Full Text HTML](#)

PostDOCK: A Structural, Empirical Approach to Scoring Protein Ligand Complexes

Clayton Springer,* Helgi Adalsteinsson, Malin M. Young, Philip W. Kegelmeyer, and Diana C. Roe

Sandia National Labs, P.O. Box 969, MS 9951, Livermore, California 94551

Received August 11, 2004

In this work we introduce a postprocessing filter (PostDOCK) that distinguishes true binding ligand–protein complexes from docking artifacts (that are created by DOCK 4.0.1). PostDOCK is a pattern recognition system that relies on (1) a database of complexes, (2) biochemical descriptors of those complexes, and (3) machine learning tools. We use the protein databank (PDB) as the structural database of complexes and create diverse training and validation sets from it based on the “families of structurally similar proteins” (FSSP) hierarchy. For the biochemical descriptors, we consider terms from the DOCK score, empirical scoring, and buried solvent accessible surface area. For the machine-learners, we use a random forest classifier and logistic regression. Our results were obtained on a test set of 44 structurally diverse protein targets. Our highest performing descriptor combinations obtained ~19-fold enrichment (39 of 44 binding complexes were correctly identified, while only allowing 2 of 44 decoy complexes), and our best overall accuracy was 92%.

Introduction

Molecular docking calculations are a well-established tool in drug discovery. Docking is used to screen large databases of small molecules against a given target receptor and to identify those that can reasonably be expected to bind to the receptor. The core function of all docking programs is to identify ligand poses (i.e. orientations and conformations) that match the structural and chemical characteristics of the target receptor. Typically, docking calculations are run on large sets of small molecules (10^6 – 10^9 molecules or larger); thus, execution speed is increased at the expense of accuracy. The top ranking ligands are then considered for further development.

The most accurate calculations of the free energy of binding for a protein–ligand complex use explicit waters in full molecular dynamics/Monte Carlo simulation.^{1–3} These calculations are still at least 6 orders of magnitude too slow to be used in large-scale screening of putative ligands.^{4–7} Although docking does not include explicit representations of solvent water molecules, implicit solvation models, such as generalized Born solvation, have been used successfully in docking simulations,⁸ but the added computational cost of those methods restricts them to much smaller datasets than occur in high throughput screening. Faster docking methods make further approximations to these models.^{9,10}

Due to the compromises between accuracy and speed that have to be made for large scale docking simulations, it is common to refer to the binding value assigned to a docked pose as a “score” rather than as free energy and to refer to the simplified physics models used to arrive at these scores as “scoring functions”. The most common fast scoring functions used for screening large datasets

are: force field based scoring schemes^{11–14} in which the nonbonded electrostatic and Lennard–Jones terms are used from a molecular mechanics force field to estimate intermolecular interactions (e.g., DOCK), knowledge-based scoring schemes^{15,16} in which binding potentials are derived from statistical analysis of cocrystallized protein–ligand pairs (e.g., Pmf, DrugScore), and empirically derived scoring schemes^{17,18} which begin by assuming a chemically intuitive form for the descriptors (e.g., hydrogen bonding and contact surface area) and then fit coefficients to create a potential function (e.g., LUDI and ChemScore). These scoring functions are useful singly or in consensus scoring.¹⁹ Bissantz et al.²⁰ have shown that none of several popular scoring functions is best for both hydrophilic (thymidine kinase) and hydrophobic (estrogen receptor) targets. The functional form of empirically derived scoring schemes can be chosen to optimize computational speed and chemical interpretability, but their predictive accuracy usually does not necessarily extend beyond the target receptor families on which they were trained.

In this paper, we describe the development of a new approach to evaluating and ranking ligand–receptor complexes. The goal is to improve the results obtained by the fast, but inexact, scoring methods in docking by adding a postprocessing filter (the “postDOCK” filter) that can divide the output poses into binders and decoys. Each postDOCK filter uses molecular descriptors to describe the characteristics of the ligand–receptor complex, and pattern recognition to separate out the binding from the nonbinding ligands. The molecular descriptors include ones similar to those used in existing empirical scoring schemes, as well as novel descriptors. We use a random forest classifier.²¹ Random forest is an extension of the classification and regression tree (CART) algorithm. CARTs have previously been used in chemical applications,²² are nonparametric, and train quickly, and their output is a readily interpretable tree.

* Current address: Novartis Institutes for Biomedical Research, 100 Technology Square, Cambridge, MA 02139. Phone (617) 871-7588. clayton.springer@novartis.com.

Table 1. Selected Examples of Protein–Ligands and Decoys

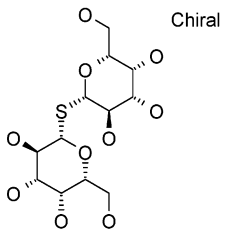
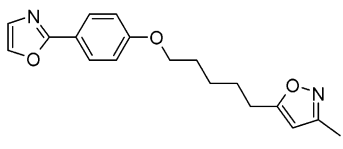
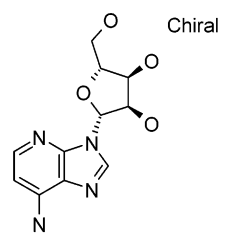
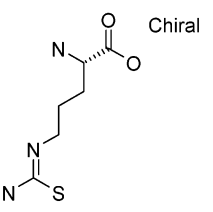
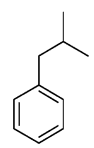
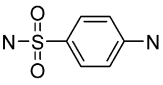
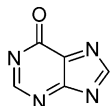
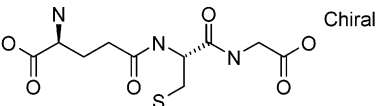
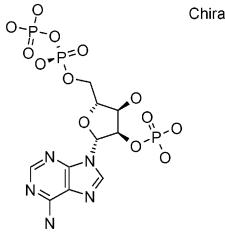
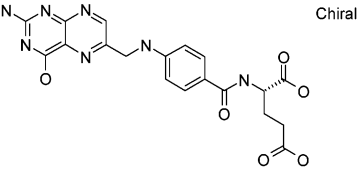
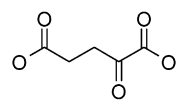
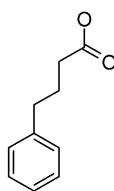
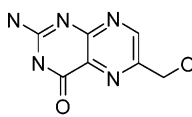
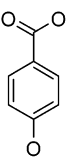
Protein target Tanimoto Similarity	Native ligand;	Decoy Ligand	
1a78_tdg 1rug_w35	PDB: TDGP_1A781 	PDB: W35P_1RUG1 	0.241
1add_1da 3nod_sci	PDB: 1DAP_1ADD1 	PDB: SCIP_L3NOD7 	0.204
184l_i4b 1aj0_san	PDB: I4BP_184L2 	PDB: SANP_L1AJ03 	0.198
1a9r_hpa 6gsy_gtt	PDB: HPAP_L1A9R1 	PDB: GTTP_6GSY1 	0.241
1alu_tar 1ra8_fol	PDB: ATRP_1RA82 	PDB: FOLP_1RA81 	0.482
1aib_2og 1thl_clt	PDB: 2OGP_L1AIB2 	PDB: CLTP_L1THL1 	0.376
1aj0_ph2 1pbe_phb	PDB: PH2P_L1AJ02 	PDB: PHBP_L1PBE2 	0.210

Table I (Continued)

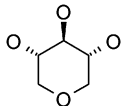
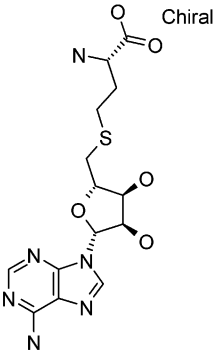
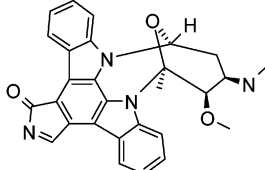
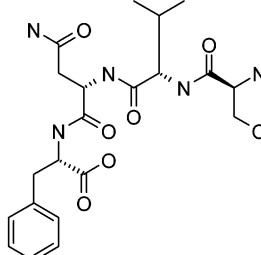
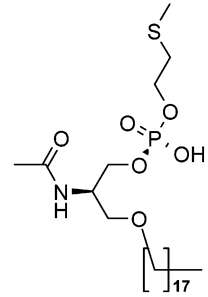
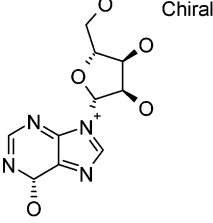
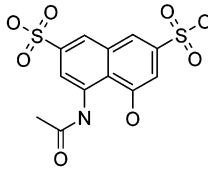
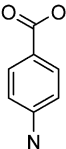
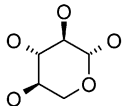
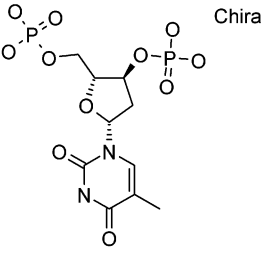
Protein target	Native ligand;	Decoy Ligand	
Xysp_1apv4 sahp_1aqi1	PDB: XYSP_1APV4  Chiral	PDB: SAHP_1AQI1  Chiral	0.162
1aq1_stup lyth_1yth1	PDB: STUP_1AQ11  Chiral	PDB: PEPP_1YTH1  Chiral	0.318
1ayp_inb 2ada_hpr	PDB: INBP_1AYP7  Chiral	PDB: HPRP_2ADA2  Chiral	0.306
1a5w_y3 1iut_pab	PDB: Y3_P_1A5W1 	PDB: PABP_L1IUT2 	0.281
1bcx_xys 2enb_ptp	PDB: XYSP_L1BCX3  Chiral	PDB: PTPP_2ENB1  Chiral	0.235

Table 2. Descriptors Used

name of descriptor	type of descriptors	number of descriptors
DOCK	DOCK direct interaction: vdw and electrostatics	2+1 ^a
SASA	solvent accessible surface area (1.4 Å probe) calculated by SURF	11+1 ^a
SASA-L	solvent accessible surface area (6 Å probe) calculated by SURF	11+1 ^a
ChemScore	hydrogen bonding, metal binding, lipophilic, rotatable bonds	4+1 ^a
Total		32

^a The additional (+1) descriptor is from treating the sum of the descriptors as an additional descriptor. This was included because decision trees do not automatically detect linear combinations.

To address the high dimensional space created by using multiple descriptors, described below, we used ensembles of learners and feature set selection techniques. To provide broad applicability, postDOCK was trained on 152 protein families^{23,24} that represent a sequentially diverse set of protein targets. Using a test set of 44 structurally diverse protein targets, the post-DOCK filter exhibits a 19-fold enrichment for identifying correctly docked vs decoy complexes over random filtering.

Computational Methods

Construction of the Database of Binders and Decoys.

To construct the binding complexes for the data sets, all protein–ligand complexes from the protein data bank (PDB) were scanned. We eliminated ligands that had missing heavy atom coordinates, were common crystallographic solvents, that were monomeric sugars containing less than 10 atoms, that had more than 100 heavy atoms, or that had greater than 30 rotatable bonds. Missing side chains and hydrogen atoms of the proteins were added to all complexes in the final data sets using Sybyl. Ligands were atom typed and hydrogens added by hand. To remove steric clashes, the position of the ligand in the receptor site was optimized using 100 steps of the simplex optimizer in DOCK4.0.1.²⁵ Removing steric clashes is also important because we do not want steric clashes to be associated with true binding poses, even if they are found frequently in crystallographic poses.

To determine the performance of pattern recognition algorithms on the relevant problem requires proper selection of the training set. Since the postDOCK filter is intended as a general docking filter, we constructed the training and test sets using the broadest range of protein targets available. The 152 training complexes and 44 test complexes were selected from this set of PDB complexes using the Fold classification based on Structure-Structure alignment of Proteins (FSSP) fold tree^{23,24} to achieve diversity across all known proteins. The training set was a sequentially diverse set generated by selecting one protein target from each cluster in the lowest tier of the FSSP hierarchy. From the ~2400 clusters in FSSP, most of which had no ligands, there were a total of 152 ligand–protein complexes for the training set (Table S1). The validation set was drawn from the FSSP fold tree's structurally diverse top tier. The top tier of FSSP has ~600 structurally diverse protein folds. If we limit ourselves to no more than one protein–ligand complex from each fold that is not already used in the training set we are left with 44 complexes for the validation set (Table S2). Some selected examples of protein targets, their true binders, and the decoy used are provided in Table 1.

We also created training and test sets of “decoy” ligands from the other ligands in the PDB. Separate decoys were generated for each protein in the training and test sets. The DOCKing process began with a DOCK sphere-set generated using the coordinates of the native ligand to identify the binding site. All ligands were DOCKed into each protein target (500 orientations with a distance tolerance of 0.25 Å, a distance minimum of 2.0 Å, and a node matching minimum of 3 and maximum of 10 nodes). Flexible docking was used with simultaneous search of torsions and default flexible parameters. The protein and ligand were assigned AMBER all atom

charges and Gasteiger-Marsili partial charges respectively by Sybyl, followed by simplex optimization with DOCK's default force field parameters. Those ligands that were successfully DOCKed and had a daylight fingerprint Tanimoto coefficient³¹ of <0.5 with the native binder were candidate decoys. From this pool, the decoy ligand was selected at random. At this Tanimoto threshold we hoped to reduce the false negative rate in the decoy dataset no more frequently than a binder is found through random screening. Although not the focus of their paper, Martin et al.³² show that the probability of compounds being active steadily decreases away from an active compound. For the MAO inhibitors, 0.5 is the lowest Tanimoto distance shown, and it appears to have a slightly higher probability of being active than the average in the set. Table 1 shows selected pairs of binders and decoys. The diversity of the ligands with respect to hydrogen bonding acceptors and donors and CLogP³³ for the training and test sets are shown in Figures 1 and 2.

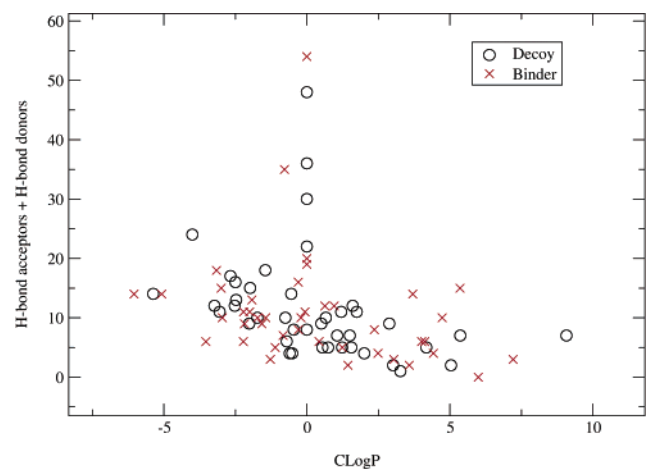
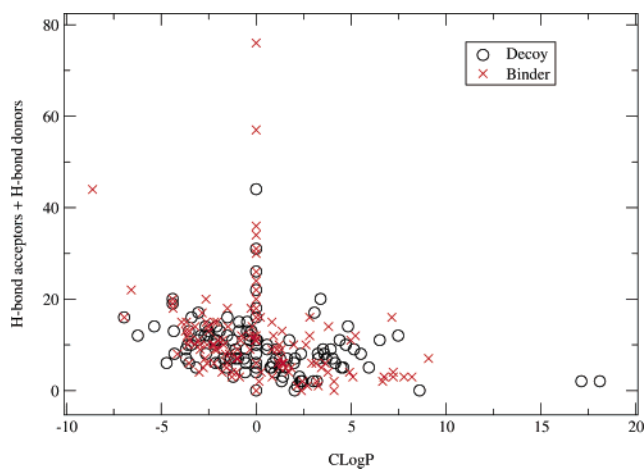
Structural Descriptors. Structural descriptors were used to characterize the protein–ligand complexes. First the structures were minimized using the DOCK total score to remove any steric clashes from both the binders and decoys. This ensured that only valid DOCK poses were evaluated. The set of feature descriptors used to characterize the ligand–protein complexes are summarized in Table 2. The DOCK total score provided the nonbonded van der Waals and electrostatic interaction force field terms, but the other descriptors were calculated in a postprocessing step. These included descriptors from our in-house implementation of the Eldridge ChemScore function.¹⁸ The Eldridge function consists of terms for hydrogen bonding, metal bonding, lipophilicity, and the number of bonds whose rotation is blocked by the presence of the receptor. In addition, we calculated a series of buried solvent accessible surface area (SASA) descriptors, using SURF³⁴ with a 1.4 Å probe, as a computationally inexpensive way of gauging solvation effects. We then calculated the surface attributable to each of a set of atom types, in an atomic solvation parameter³⁵ inspired approach, and used each as a separate descriptor. The atom types were based on SYBYL MOL2 atom types (see Table S3). As there is evidence that more than one length scale is needed to describe the self-association of hydrogen bonding water in aqueous solvation,⁶ we also created a novel descriptor called SASA-L. Our goal with this descriptor was to capture the clathrate ring length scale effects of the hydrogen bonding-associated network, so we selected a 6 Å probe (as compared to 1.4 Å for a water-sized SASA probe) to calculate the SASA-L. Again we calculated how much surface was attributable to each surface atom type (see Table S3) to generate a set of descriptors. Finally, because trees consider each descriptor individually, rather than in a linear combination, decision trees can benefit from including totals of our DOCK, SASA, SASA-L, and ChemScore descriptors as well as the 28 component descriptors (for a total of 32 descriptors considered). The correlation matrix of the selected descriptor set is provided in the Supporting Information.

The Pattern Recognition Algorithm. We used R's package of random forest (called randomForest).²¹ Important individual features were selected by variable importance,²¹ and subsets were selected from high performing individual features. To select which feature sets to use, we began with the the variables selected by random forest's variable selection procedure. From the top 20 of these all pairs, triplets and quartets were assessed, and those with the highest performance are noted. For a single learner (either tree or linear

Table 3. Test Set Performance of Random Forest (each row shows the performance for a subset when used to build an ensemble of decision trees)

DOCK ^a	SASA ^a	SASA-L ^a	ChemScore ^a	Percent Binders correctly identified at best performance ^b	Percent decoys incorrectly identified at best performance ^b	Enrichment identified at 2 FP
		all ^c		86.4	2.3	18.5
es	Total and O_BYL ^e		hb ^f	86.4	2.3	19
	Total and C_BYL		hb and lipo	86.4	4.5	19
Total	Total		hb and rot	84.1	4.5	18.5
es	Total and C_ALI		hb	84.1	4.5	18.5
Total	Total	C_ALI	hb	84.1	2.3	18.5
Total	Total		hb	79.5	2.3	18
Total	C_ALI		hb	81.8	4.5	18
	Total and C_ARO		hb	79.5	2.3	17.5
Total	Total		hb	81.8	4.5	18
	Total		hb	75.0	2.3	16.5
	O_BYL		hb	68.2	4.5	15
	C_ALI		hb	72.7	6.81	14.5
			Total	84.1	40.9	7
	Total			50.0	13.6	7
	C_ALI			47.7	18.2	7
es				27.3	6.8	5
Total				29.5	6.8	3
vdW				45.4	27.3	3.5

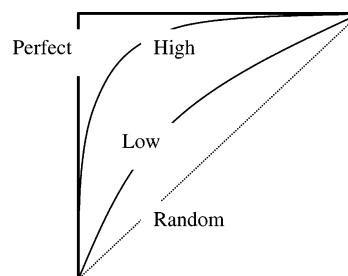
^a The leftmost four columns describe which descriptors are in each subset. ^b 50% threshold is the max performance of the ROC curve. The error bar for these columns is about 3.5%. ^c “all” means that each component was included. ^d “Total” means the summed descriptor values. ^e “C_BYL”, “O_BYL”, and “C_ALI” are components of the SASA. ^f “hb”, “lipo”, and “rot” are components of ChemScore.

**Figure 1.** The test set's hydrophobic (CLogP) and hydrophilic (H-bond donors plus acceptors, calculated by Daylight) properties.**Figure 2.** The training set's hydrophobic (CLogP) and hydrophilic (H-bond donors plus acceptors, calculated by Daylight) properties.

model), reducing the number of descriptors improves performance because the relatively small number of complexes in the PDB-based training set is spread over fewer dimensions.^{36–38} Finally, the most important and appealing aspect of using fewer descriptors is to produce a simpler, more physically interpretable model.

Once optimal descriptor subsets were determined, we developed linear models to complement the ensemble of decision trees for these sets. Our linear models are logistic regression (LR)^{39,40} with the glm package of R.⁴¹

Creating a ROC Curve. To calculate the performance of our models over the full range of specificity and sensitivity, we have calculated receiver operator characteristic (ROC) curves.⁴² To generate an ROC curve, we adjust the threshold at which the ensemble of learners determines that a test example is a binder (or decoy). At one extreme, we begin at the bottom, left corner where the threshold is set so high that all the examples are declared decoys (see Figure 1). As the threshold is decreased incrementally, more examples are declared binders. On the ROC curve a true binder being declared a binder moves the curve up, and a decoy being declared a binder moves the curve to the right. In a perfect ROC curve all of the binders are found first and the curve goes to the top, left corner. As the threshold is increased, all the

**Figure 3.** A schematic set of ROC curves showing perfect performance, high performance, low performance, and random performance.

examples are declared binders and the curve goes from the bottom, left to the top, right corner.

Results and Discussion

Searching for Optimal Descriptor Subsets. To capture more of the essential biochemistry of complex stability, we considered a large number of descriptors (see Table 2). With 32 descriptors, there are $2^{32} - 1$ (or

Table 4. Test Set Performance from Selected Logistic Regression Subsets

DOCK ^a	SASA ^a	ChemScore ^a	percent of binders correctly identified at best performance ^b	percent of decoys incorrectly identified at best performance ^b	enrichment factor at 5% FP
Total	Total	Hb	84.1	6.8	16.5
		Hb	70.5	2.3	15.5
Total		Hb	68.2	2.3	15
Total	Total		63.6	13.6	9
		Hb	70.5	6.8	13.5
Total			63.6	13.6	9
	Total		61.3	11.3	10.5

^a The four left-most columns describe which descriptors are in each subset. ^b The error bar for this column is about 7.5%.

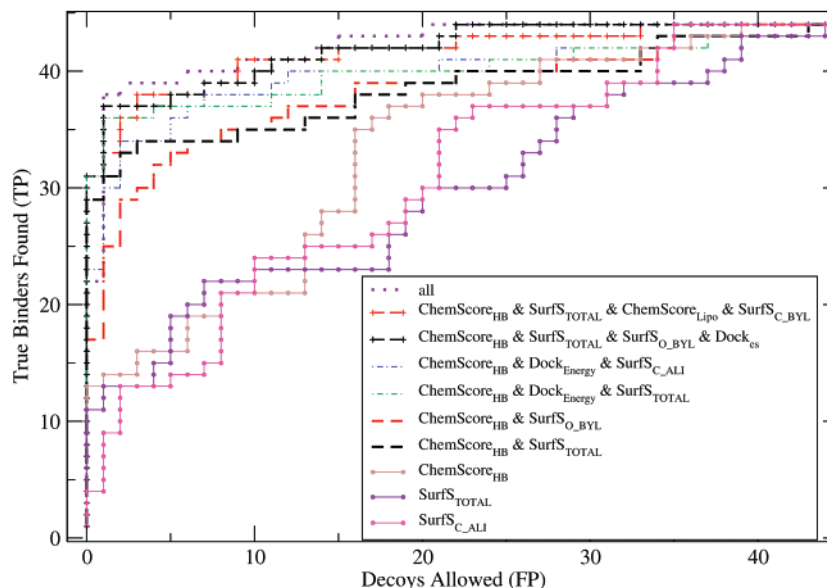


Figure 4. Full ROC curves for selected descriptor subsets using a random forest learner (see also Table 3). Overall performance is determined by proximity to the top, left corner. The single descriptor subsets make up the lowest performers. Selected points are presented numerically in Table 3. Figure 5 gives a detailed view of the high performing region.

4.29×10^9) possible descriptor subsets. To reduce the choice of subsets to a computationally tractable problem, we applied feature set selection methods. This data set is fairly straightforward so forward feature set selection and variable selection all identify very similar descriptor subsets as promising. Selected high-performing subsets are shown in Table 3. This table shows the number of binders and decoys correctly and incorrectly identified at the best performing point on the ROC curve (i.e., closest to the top, left corner). We expect statistical fluctuations on the order of $\sqrt{(N)p(1-p)} \sim \pm 9$ for the enrichment factor in the top performing sets. All top subsets include ChemScore hydrogen bonding (H-bond) and a buried surface area (SASA) term. Many of these subsets also include terms from the DOCK force-field score and the large sphere SASA-L.

Performance of Descriptor Subsets. To evaluate the full performance of the descriptor subsets on our test set, we examined each over its entire range of sensitivity and specificity using ROC curves (see Figures 3 and 4). We see that the lowest performing subsets plotted in Figure 4 are composed of single descriptors; conversely, all of the top performing feature sets rely on feature synergies. Table 3 shows the midpoint on the ROC curve for each descriptor subset. To estimate enrichment performance, we also note the number of true positives discovered when 2 (of 44) false positives are allowed for each descriptor subset. We note that although the existing scoring schemes DOCK and

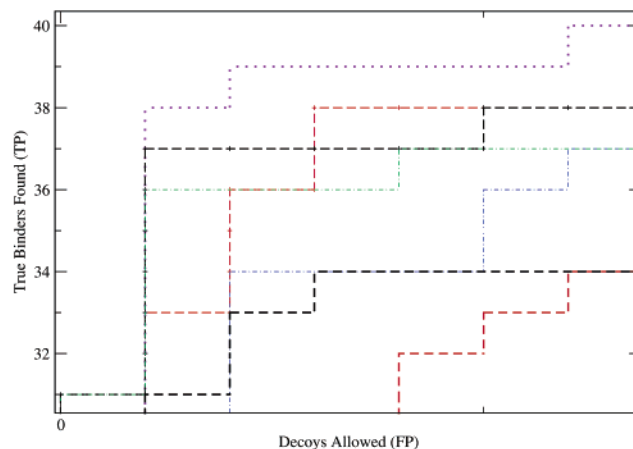


Figure 5. Detail of the high-performing region of Figure 4 (top, left). All the best subsets contain SASA_{TOTAL} and ChemScore_{PH-BOND}, and all the lowest performing subsets contain only a single descriptor. Those without a subscript are using all the descriptor from that source (Table 2). (This figure uses the same line colors and symbols that are used in Figure 4.)

ChemScore, which are each summarized in a single descriptor, eliminate many decoys (5× and 7× enrichment are seen, respectively), by adding additional descriptors we can obtain up to 19× enrichment (Table 3). In screening, the goal is enrichment, so while getting the binding complexes correct is important, it is critical to correctly identify (and then discard) as many decoys

Table 5. Coefficients from Selected Logistic Regression Models^a

models	intercept	ChemScore _{H-bond}	SASA _{total}	DOCK _{total}
SASA _{total} + ChemScore _{H-bond}	-2.64 ± 0.37	-0.314 ± 0.043	-0.00406 ± 0.00097	
DOCK _{total} + ChemScore _{H-bond}	-0.02 ± 0.37	-0.380 ± 0.048		+0.066 ± 0.016
DOCK _{total} + SASA _{total} + ChemScore _{H-bond}	-1.29 ± 0.46	-0.367 ± 0.049	-0.0053 ± 0.0011	+0.076 ± 0.017
DOCK _{total}	+0.10 ± 0.26			+0.0036 ± 0.0086
SASA _{total}	-1.66 ± 0.29		-0.00541 ± 0.00090	
ChemScore _{H-bond}	-1.44 ± 0.20	-0.316 ± 0.041		

^a There is one subset per row and one column for each descriptor. Empty cells indicate that the descriptor is not included in that subset. The “std dev” are as reported from R’s glm package.

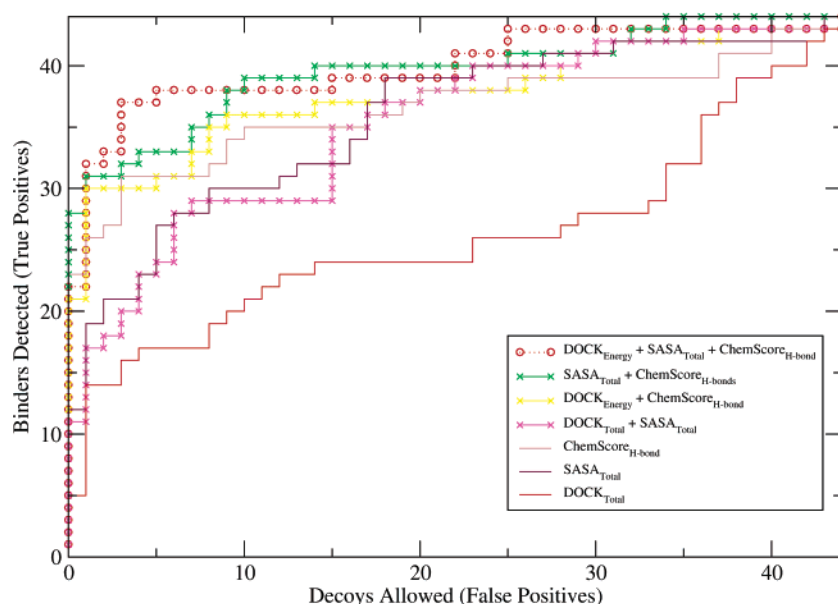


Figure 6. Full ROC curves for selected descriptor subsets using logistic regression. Descriptor sets with a subscript are for a single descriptor. Overall performance is measured by proximity to the top, left corner. The best subset contains SASA_{TOTAL} and ChemScore_{H-BOND}. Lowest performing subsets are composed of a single descriptor. (Table 2). Selected points from this figure are presented numerically in Table 4.

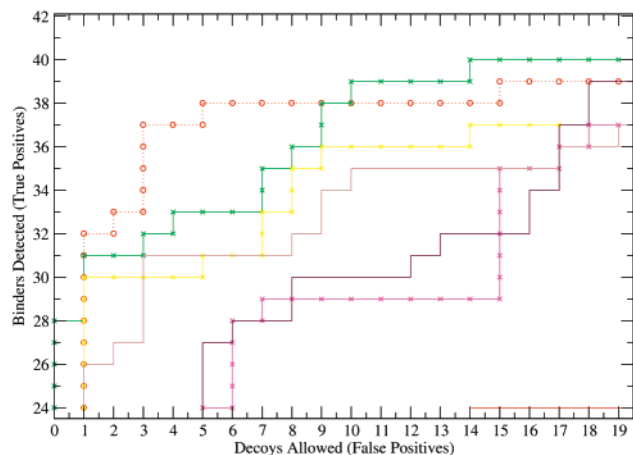


Figure 7. Detail of high performers from Figure 6 and using the same colors and symbols. Overall performance is by proximity to the top, left corner. The best subset contains SASA_{total} and CSCORE_{H-bond}. Lowest performing subsets are composed of a single descriptor.

as possible. In this case, enrichment factor is the best metric to evaluate. Note that ROC curves are asymptotically equivalent to enrichment curves (true positives vs all positives) when the number of false positives becomes very large (e.g., when screening a large, diverse library for hits). For other applications (e.g., obtaining docking poses), a balanced consideration of binding and nonbinding interactions may be desired, and overall

performance is the appropriate metric. One of our key results is to identify the best performing descriptor set(s) for the postDOCK filter for each of these two applications. Apparently, synergy between hydrophobic (either SASA_{total} or ChemScore_{lipophilic}) and hydrophilic (as ChemScore_{H-bonds}) descriptors is essential for both enrichment and overall performance: these descriptors appear in all our best subsets (Figures 3 and 4, and Table 3).

Enrichment Results. The ROC curve’s initial slope (the enrichment factor) is the relevant performance metric in virtual screening where the goal is to minimize the number of false positives for each true positive discovered. Random screening has an enrichment factor of 1-fold; useful procedures have higher enrichment factors. Existing single descriptor scoring schemes DOCK and ChemScore each have 5-fold or 7-fold enrichments. The highest enrichment is found in subsets that contain the features SASA_{Total} and ChemScore_{H-bond}, which is ~20-fold enrichment with 5% false positives. With logistic regression the subset of (DOCK_{total} + buried SASA_{total}, + ChemScore_{H-bond}) has an enrichment factor of ~16-fold.

Overall Performance. For applications where eliminating decoys and discovering binders are of equal importance (for example, lead optimization and pose generation), the best overall performance is the closest that the ROC curve gets to the top, left corner; that is

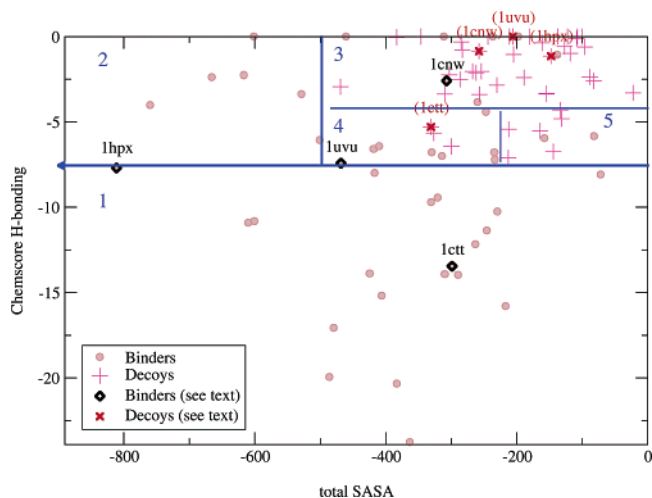


Figure 8. The points show the values of ChemScore_{H-bond} and total SASA_{total} descriptors for the test set for both binders and decoys. The blue lines show the partitioning of the space as given by a single tree on the training set into rectangular regions which are labeled 1 through 5. The regions 1, 2, and 4 contain a majority of binders, and regions 3 and 5 contain a majority of decoys. Eight points (discussed in the text) are labeled by the pdb id of the target, four targets with a binder and decoy each. The black type labels and black diamond points are the binding examples and the red code labels are the decoys. The poses and structures for these points are shown in Figures 9–16.

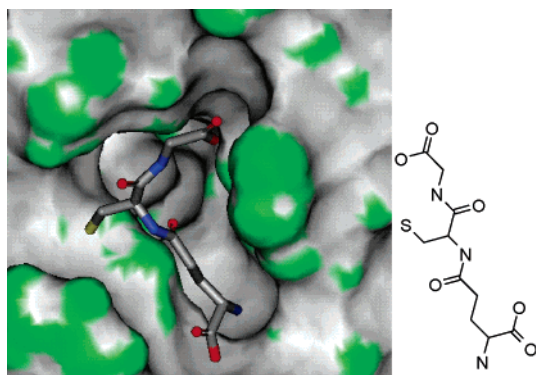


Figure 9. The decoy gdn (from 1hnc) docked into HIV protease (1hpX) falls in region 3 of Figure 8. It does not bury enough surface area or have enough H-bonding to be mistaken for a binder.

when (percent true positives found) – (percent false positives allowed) is at its highest. The overall performance of ChemScore and DOCK alone are finding 37/44 true binders while allowing 18/44 decoys and finding 13/44 true binders while allowing 3/44 decoys, respectively. By contrast, the highest performance found in any subset is finding 38/44 true binders while only allowing 2/44 decoys. To estimate the effects stochastic elements of the random forest we grew 300 random forests. In these 300 the mean enrichment was 18.96 and 25th, 50th (median), and 75th percentiles were 18.5, 19, and 19.5, respectively. The best performance is (percentage of binders found) – (percentage of decoys allowed). The mean best performance is 83% and the 25th, 50th, 75th percentiles were 81.8, 81.8, 84.1%, respectively.

Decoys are Easier to Identify Than Binders.

Traditional scoring functions focus on correctly predicting how well a ligand will bind. Enrichment emphasizes

eliminating decoys. Almost all the descriptor combinations we consider find eliminating decoys easier than finding binders. For example, the DOCK total force field score gets most of its enrichment from eliminating decoys (from Table 3: DOCK_{total} identifies 13 of 44 binders while eliminating 41 of 44 decoys), rather than positively identifying those that do bind. Indeed, for many descriptor subsets, the percentage of decoys correctly identified is roughly 25% higher than the percentage of binders correctly identified at the midpoint on the ROC curve (Figures 3 and 4). Another manifestation of this is that most of the overall performance maxima occurred with fewer false positives than false negatives. Logistic regression also finds decoys easier to detect than binders (Figures 5 and 6).

Logistic Regression. We can connect this work with other empirical scoring functions by making a linear model of our descriptors. Visual inspection of Figure 8 suggests that a linear decision surface would be effective for this training set. We use logistic regression (LR) as our linear method.⁴⁰ The performance of the descriptors in this linear model is a little lower than that of the random forest model (see Table 4). Similar features and feature subsets are important for the two techniques. In particular, SASA_{total}, ChemScore_{H-bond}, and DOCK_{total} are part of the high performing descriptor subset, and once again the single descriptors still greatly underperform the feature combinations; thus, synergies between descriptors are still important. Note the case ChemScore_{H-bond} which is the best individual feature and is better on its own than the DOCK_{total} and SASA_{total} subset. Together they still form the best triplet and subset we found.

The coefficients with their error bars for the LR models are shown in Table 5. For all descriptors, more negative values indicate stronger interactions: those descriptors with positive coefficients are associated with improved binding. (DOCK_{total} + SASA_{total} + ChemScore_{H-bond}) is the best performing subset. LR performs better on descriptor subsets than all the features because unlike the ensemble of trees it is susceptible to noisy and/or low information features included in the entire set.

A Single Decision Tree Example. One advantage of using subsets of descriptors is their ease of physical interpretation. To illustrate this point we show a simple example using only two descriptors. Figure 8 shows the recursive partitioning of the training set using the top performing descriptor pair: SASA_{total} and ChemScore_{H-bond}. Again for simplicity, Figure 8 shows results for a single decision tree instead of an ensemble of trees. Each partition is labeled in the sequential order in which it is derived. Partition 1 is the first partition dividing binders and nonbinders, and this partition (ChemScore_{H-bond} < ~-7 Kcal) has a majority of binders. Partition 2 divides the remaining compounds, and again complexes scoring in this partition (ChemScore_{H-bond} > ~-7 kcal but SASA_{total} < ~-500 Å²) are considered binders. Partition 3 divides the remainder, and all compounds scoring in this partition (ChemScore_{H-bond} > ~-4.8 and SASA_{total} > ~-500 Å²) are considered nonbinders. Similarly, compounds scoring in partitions 4 and 5 are binders and nonbinders, respectively. It is physically reasonable that ligands that have the highest

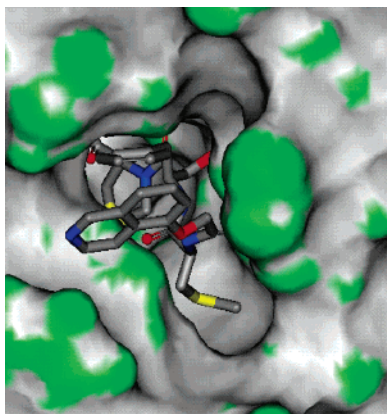


Figure 10. The binding mode of KNI in HIV protease (1hpx) as found in the pdb. It has strong enough interactions to be correctly identified as a binder and falls in region 1 of Figure 8.

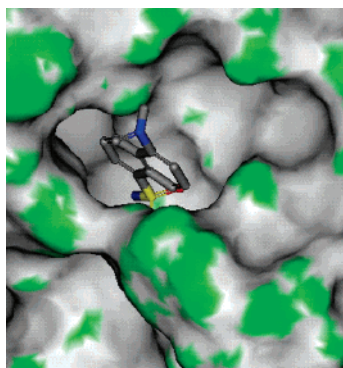
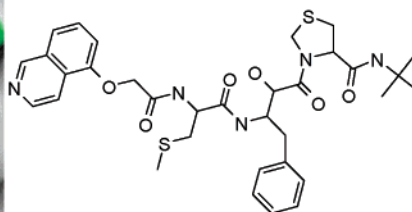


Figure 11. Thrombin (1uvu) with decoy (1okl_mns) ligand DOCKed in. It is correctly identified as a nonbinder and falls in region 3 of Figure 8.

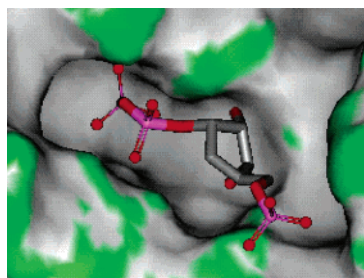
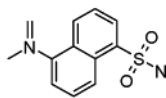


Figure 13. The decoy pcp (1a96) docked into cytidine deaminase (1ctt). This decoy has enough buried surface area and H-bonding to fall in region 4, a majority binding region, and therefore is a false positive in our single tree, two-descriptor learner.

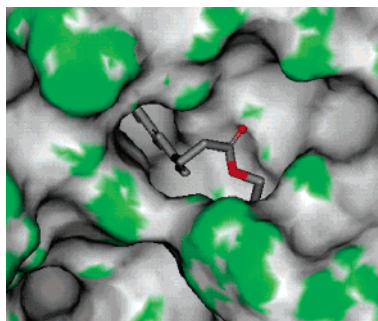
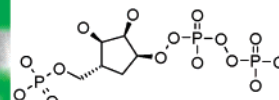


Figure 12. This is thrombin with binder from the pdb. It is correctly identified as a binder and falls in region 4 of Figure 8.

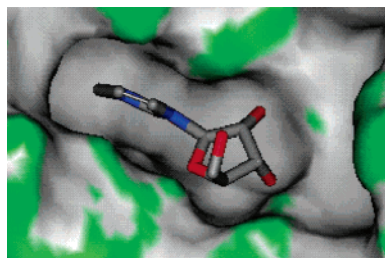
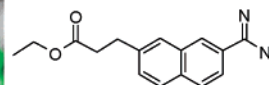


Figure 14. This is native ligand (dHz) docked into cytidine deaminase (1ctt). It has enough interactions to be correctly considered a binder and can be found in region 1 of Figure 8.

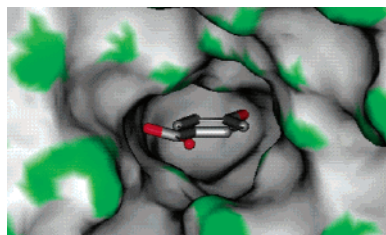
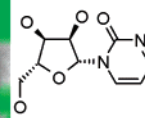
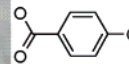


Figure 15. The decoy ligand (pHb) in carbonic anhydrase (1cnw). Because of its lack of interactions it falls in region 3 of Figure 8 and is correctly excluded as a decoy by a single tree learner.



number of hydrogen bonds and/or most buried surface area are partitioned into binding regions.

To show how this simple descriptor set performs on particular examples, we plotted and labeled the compounds from the test set onto this partitioning scheme (see Figure 8). Inaccurate predictions occur when a decoy complex buries a large amount of surface area or makes many hydrogen bonds. Conversely, inaccurate predictions also occur when some true binders do not bury surface area or have enough hydrogen bonds in this simple partitioning to be labeled binders. Although this learner uses only two descriptors, it does a fairly good job of correctly partitioning the test set. With more descriptors, we can better distinguish binders from nonbinders, such as in our top-performing descriptor set ($SASA_{\text{L-carbon}} + SASA_{\text{total}} + \text{ChemScore}_{\text{H-bond}}$).

We consider four specific cases from the test set in the example from Figure 8. Two of them are treated correctly, and the other two either misclassify a decoy or a binder. In our first example, HIV-1 protease (PDB entry 1hpx), both the binder and decoy are properly scored. The crystallographic ligand has a large amount of buried surface area, placing it clearly in a binding

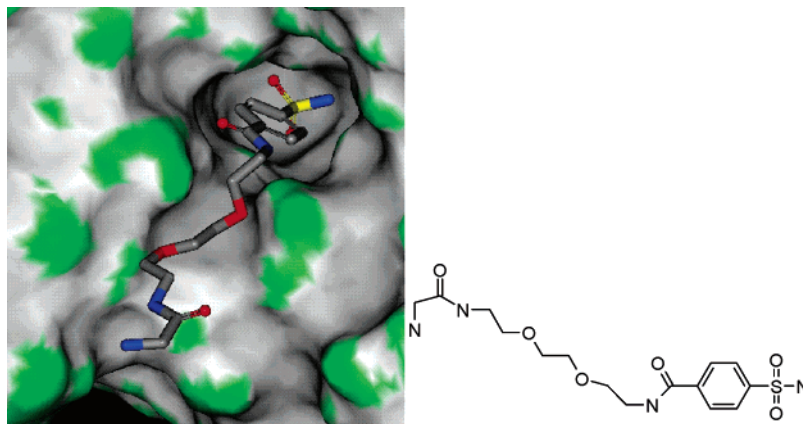


Figure 16. The native ligand (eg1) in carbonic anhydrase (1cnw). This complex does not have enough interactions to be considered a binder. Therefore for this two-descriptor single tree this would be false negative. Metal binding interactions must be considered to properly evaluate this complex.

region. The HIV-1 protease decoy is clearly in the nonbinding partition 3, due to both the lower number of hydrogen bonds and the smaller amount of buried surface area (see Figures 9 and 10). Thrombin (PDB entry 1uvu) is also classified correctly. Again the thrombin PDB complex has much greater H-bonding and buried surface area relative to the decoy (see Figures 11 and 12). The PDB structure of carbonic anhydrase (PDB entry 1cnw) is incorrectly perceived to be a decoy. This is because the H-bonding and SASA terms miss the metal binding interaction that is critical for this target (see Figures 13 and 14). When metal binding is included, this complex is correctly recognized. The cytidine deaminase binder is correctly recognized as such, but the decoy has enough H-bonding that it is predicted to be a binder (see Figures 15 and 16). Inclusion of the DOCK total allows the correct identification of this decoy complex. Using many descriptors and complex pattern recognition methods achieves higher performance; having fairly good performance from a simple model achieves interpretability.

Conclusion

In the virtual screening of large libraries, the vast majority of compounds are of little interest, and limited computation time can be spent on each compound. The task is to eliminate as many nonbinders as possible while retaining the binders. The approach we developed combines fast docking (frozen protein, no explicit solvent) with a pattern recognition algorithm to create a set of postDOCK filters that can be used to triage the implausible binding poses, leaving a greatly reduced set of compounds for further consideration. For almost all descriptor sets, decoys were much easier to detect than binders, and therefore high enrichments were possible without having high accuracy on binders. Existing scoring schemes (DOCK and ChemScore) eliminate many decoys (with 5-fold and 7-fold enrichments). Our best model (constructed using all of our descriptors with random forest learner) has a 19-fold enrichment (recovering 39 of 44 binders while allowing only 2 of 44 of the decoys).

We used machine learning techniques to systematically improve performance by testing new descriptors in combination with existing descriptors. Our results clearly show that the DOCK force field could benefit

from buried surface area and H-bonding terms. Further improvements in performance will come from better descriptors. One source of improvement could come from using partial charge calculation methods that are more accurate than the Gasteiger–Marsili procedure that we used and replacing the distance dependent dielectric with a better implicit solvation method. In this paper, we have tried large sphere SASA (which to our knowledge is a novel descriptor in characterizing protein–ligand complexes) in addition to the more traditional descriptors of protein–ligand complexes. Although it does not contribute to enrichment, the large sphere buried surface area descriptor may improve performance in the mid part of the ROC curve. This effect is not statistically significant and needs to be repeated before we can say that we are seeing aqueous self-association effects.

Because our examples were trained and tested on a diverse set of proteins, we expect the postDOCK filters we listed to be generally applicable to a variety of targets when there is no other a priori information. For cases where there is additional information, such as binding data to a homologous protein, the methods described in this paper could be used to train a dataset specific to the target of interest. Additional descriptors may be added to further tune the postDOCK filter to that target, such as adding metal binding descriptors when the target is a zinc binding protein complex (as in our carbonic anhydrase example). Similarly, this methodology can be applied to tune target specific postDOCK filters using binding assay data.

Acknowledgment. C.S. and H.A. acknowledge postdoctoral funding support from the Sandia National Labs LDRD program.

Supporting Information Available: The pdb id codes of the proteins and ligands that form the training and test sets; the atom typing scheme used for the SASA calculations; pairwise correlation between descriptors in the training set. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Orozco, M.; Tirado-Rives, J.; Jorgensen, W. L. Mechanism for the rotamase activity of FK506 binding protein from molecular dynamics simulations. *Biochemistry* **1993**, *32*, 12864–12874.
- (2) Sham, Y. Y.; Chu, Z. T.; Tao, H.; Warshel, A. Examining methods for calculations of binding free energies: LRA; LIE; PDL-LRA; and PDL/S-LRA calculations of ligands binding to an HIV protease. *Proteins-Struct. Funct. Genet.* **2000**, *39*, 393–407.

- (3) Zhang, L. Y.; Gallicchio, E.; Friesner, R. A.; Levy, R. M.; Guermeur, Y. et al. Solvent models for protein–ligand binding: Comparison of implicit solvent Poisson and surface generalized born models with explicit solvent simulations. *J. Comput. Chem.* **2001**, *22*, 591–607.
- (4) Robinson, G. W.; Cho, C. H.; Urquidi, J. Isobestic points in liquid water: Further strong evidence for the two-state mixture model. *J. Chem. Phys.* **1999**, *111*, 698–702.
- (5) BenNaim, A. A measure of the average cooperativity of a binding system. *J. Chem. Phys.* **1998**, *109*, 7443–7449.
- (6) Lum, K.; Chandler, D.; Weeks, J. D. Hydrophobicity at small and large length scales. *J. Phys. Chem. B* **1999**, *103*, 4570–4577.
- (7) Wallqvist, A.; Covell, D. G. Cooperativity of Water–Solute Interactions at a Hydrophilic Surface. *J. Phys. Chem.* **1995**, *99*, 5705–5712.
- (8) Zou, X. Q.; Sun, Y. X.; Kuntz, I. D. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *J. Am. Chem. Soc.* **1999**, *121*, 8033–8043.
- (9) Majeux, N.; Scarsi, M.; Caffisch, A. Efficient electrostatic solvation model for protein–fragment docking. *Proteins-Struct. Funct. Genet.* **2001**, *42*, 256–268.
- (10) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins-Struct. Funct. Genet.* **1999**, *34*, 4–16.
- (11) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505–524.
- (12) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **1999**, *34*, 4–16.
- (13) 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.
- (14) Pang, Y. P.; Perola, E.; Xu, K.; Prendergast, F. G. EUDOC: A computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.* **2001**, *22*, 1750–1771.
- (15) Muegge, I.; Martin, Y. A general and fast scoring function for protein–ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791–804.
- (16) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein–ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
- (17) Bohm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. **1994**, *8*, 243–256.
- (18) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions 0.1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
- (19) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.
- (20) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.
- (21) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
- (22) Rusinko, A.; MW, F.; Lambert, C.; Brown, P.; Young, S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017–1026.
- (23) Holm, L.; Sander, C. The FSSP Database of Structurally Aligned Protein Fold Families. *Nucleic Acids Res.* **1994**, *22*, 3600–3609.
- (24) Holm, L.; Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **1997**, *25*, 231–234.
- (25) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411–428.
- (26) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505–524.
- (27) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C. et al. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.
- (28) Gasteiger, J.; Marsili, M. *Organ. Magn. Reson.* **1981**, *15*, 353–360.
- (29) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219–3228.
- (30) Marsili, M.; Gasteiger, J. *Croat. Chem. Acta* **1980**, *53*.
- (31) Willett, P. *Similarity and Clustering in Chemical Information Space*; Research Studies Press: Letchworth.
- (32) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecular Have Similar Biological Activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.
- (33) Leo, A. *Programs CLogP and CMR*; BioByte Corp.: Claremont, CA.
- (34) Varshney, A.; Brooks, F. J. Fast Analytical Computation of Richard's Smooth Molecular Surface; UNC at Chapel Hill: Chapel Hill, NC, 1993.
- (35) Eisenberg, D.; McLachlan, A. Solvation Energy in Protein Folding and Binding. *Nature* **1986**, *319*, 199–203.
- (36) Trunk, G. V. A problem of Dimensionality: A Simple Example. *IEEE Trans. Pattern Anal. Machine Intell.* **1979**, *1*, 306–307.
- (37) Brunzell, H.; Eriksson, J. Feature reduction of classification of multidimensional data. *Pattern Recog.* **2000**, *33*, 1741–1748.
- (38) Cover, T. M.; Van Campenhout, J. M. On the possible ordering in the measurement selection problem. *IEEE Trans. Syst., Man, Cybernetics* **1977**, *7*, 657–661.
- (39) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S-plus*; Springer: New York, 1994.
- (40) McCullagh, P.; J. A., N. *Generalized Linear Models*; Chapman and Hall: London, 1989.
- (41) CRAN. The Comprehensive R Archive Network.
- (42) Duda, R.; Hart, Stark, P. *Pattern Classification*; John Wiley and Sons: New York, 2001.

JM0493360