Binding Energy Landscape Analysis Helps to Discriminate True Hits from High-Scoring Decoys in Virtual Screening

Dengguo Wei,[†,‡] Hao Zheng,[§,‡] Naifang Su,[§,‡] Minghua Deng,[§,‡] and Luhua Lai*[,†,‡]

BNLMS, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of
Chemistry and Molecular Engineering, School of Mathematical Sciences, and Center for Theoretical Biology,
Peking University, Beijing 100871, China

Received November 26, 2009

Although virtual screening through molecular docking has been widely applied in lead discovery, it is still challenging to distinguish true hits from high-scoring decoys because of the difficulty in accurately predicting protein−ligand binding affinities. Following the successful application of energy landscape analysis to both protein folding and biomolecular binding studies, we attempted to use protein−ligand binding energy landscape analysis to recognize true binders from high-scoring decoys. Two parameters describing the binding energy landscape were used for this purpose. The energy gap, defined as the difference between the binding energy of the native binding mode and the average binding energy of other binding modes in the "denatured binding phase", was used to describe the thermodynamic stability of binding, and the number of local binding wells in the landscapes was used to account for the kinetic accessibility. These parameters, together with the docking score, were combined using logistic regression to investigate their capability to discriminate true ligands from high-scoring decoys. Inhibitors and the noninhibitors of two enzyme systems, neuraminidase and cyclooxygenase-2, were used to test their discrimination capability. Using a five-fold cross-validation, the areas under the receiver operator characteristic curves (AUCs) from the best linear combinations of parameters reached 0.878 for neuraminidase and 0.776 for cyclooxygenase-2. To make a more independent test, inhibitors and high-scoring decoys in a directory of useful decoys (DUD), the largest and most comprehensive public data set for benchmarking virtual screen programs by far, were used as independent test sets to test the discrimination capability of these parameters. The AUCs of the best linear combinations of parameters for the independent test sets were 0.750 for neuraminidase and 0.855 for cyclooxygenase-2. Furthermore, combining these two parameters with the docking scoring function improved the enrichment ratio to 200−300% compared to that using the scoring function alone. This study suggests that incorporating information from binding energy landscape analysis can significantly increase the success rate of virtual screening.

INTRODUCTION

With the availability of increasing numbers of protein structures and the advent of high-performance computing systems, molecular docking, a structure-based virtual screening technique, provides a fast, low-cost alternative to experimental high-throughput screening of large libraries of compounds and reduces the time and effort required to identify candidate drug molecules for further development.[1,2] During virtual screening, an enormous number of small molecules in large databases are docked into a given target, and the calculation of the binding free energies are simplified under various assumptions and estimated by a value known as the "score" derived according to the docked binding conformations.[3] Under the evaluation of different scoring functions, the virtual hit set is significantly enriched in bioactive molecules relative to a random selection. However, the number of compounds satisfying a "good" docking score defined by scoring known inhibitors is much larger than assay throughput,[4] and a high rate of false positives occurs if candidates are chosen for assay only according to the order of the scores.

Numerous attempts have been made to improve the scoring function,[5−10] docking method, and screening strategy.[11−21] Using complicated structural descriptors, Springer and co-workers developed postprocessing filters through machine learning tools to distinguish true binding ligand−protein complexes from docking decoys.[22] Garmendia-Doval et al. attempted to apply genetic algorithms to automatically remove false positives from virtual hit sets given by their in-house virtual screening platform, rDock.[23] To prioritize virtual screening hits for experimental assays and assess true/false positives, Lerman and co-workers presented two statistical methods for mining large databases: a general scoring metric based on the virtual screening signal-to-noise level within a compound neighborhood and a neighborhood-based sampling strategy for reducing database size, in lieu of property-based filters.[4] However, the improvements of the reported methods are far from satisfactory.

Development might depend on the accurate prediction of ligand−protein binding affinities, which requires a synergy of exhaustive conformational sampling and accurate evalu-

* Corresponding author phone: 86-10-62757486; fax: 86-10-62751725; e-mail: lhlai@pku.edu.cn.

[†] College of Chemistry and Molecular Engineering.

[‡] Center for Theoretical Biology.

[§] School of Mathematical Sciences.

10.1021/ci900463u
Published on Web 09/27/2010

ation of the delicate balance among aspects involved in the description of ligand−protein association, such as van der Waals forces, electrostatic interactions, solvation effects, and conformational entropy.[24] All of these issues remain major methodological and technical challenges.[24] In fact, in the process of practical virtual screening, it is the high-scoring decoys that make it challenging to choose hits among decoys. Further attempts aimed at discriminating between the inhibitors and the high-scoring decoys might improve the success ratio of virtual screening.

Energy landscape analysis provides an alternative, and it has been successfully utilized in describing and understanding much of the complexity in protein folding and bimolecular binding studies.[25−40] Monte Carlo simulations of protein folding using a lattice model have shown that the energy spectrum of a folding sequence exhibits a pronounced energy gap between the native state and all other conformations, a feature that is not typically found in the spectra of random sequences,[26−30] and this feature describes the thermodynamic stability of the native folding mode. In their model to identify a folding sequence, Ebeling and co-workers found that, in addition to the spectra of conformation energies, possible routes connecting the corresponding conformations in conformation space also affect the feasibility of a sequence folding.[31] Evidence from theory and simulation indicates that amino acid sequences with minimal frustration are likely to fold reliably.[32]

With the increasing recognition that protein folding and molecular binding share similar phenomena and mechanisms, energy landscape analysis is being further employed in ligand−protein binding studies. Successes and failures of molecular docking have been explained by the topology of the underlying binding energy landscape,[33−38] which is often insensitive to the accuracy of the energy models used. A minimally frustrated energy surface with a unique and stable native complex was shown to be an important prerequisite for receptor-specific binding.[35,37,38] On the basis of binding energy landscape analysis, Wang et al. proposed a quantitative optimal criterion and tested it in the system of cyclooxygenase-2 to measure the binding specificity of ligands.[39,40]

In the current study, we have attempted to use protein−ligand binding energy landscape analysis to recognize true binders from high-scoring decoys. Two parameters, the thermodynamic stability and kinetic accessibility, to describe the binding energy landscape of protein−ligand binding were used for classification purposes. These two parameters were successfully applied to the influenza virus neuraminidase and cyclooxygenase-2 systems.

DATA SETS OF INHIBITORS AND NONINHIBITORS

To rigorously evaluate the assumptions and parameters proposed, we used two systems, neuraminidase and cyclooxygenase-2, for which inhibitor binding modes are clearly understood and a large number of inhibitors with chemical structure diversity are available. Considering the distribution of physicochemical properties, 77 inhibitors of neuraminidase with 10 distinctive structural scaffolds and 78 inhibitors of cyclooxygenase-2 with 8 different basic structural scaffolds were collected from the literature and were used as “true inhibitors” (see Tables S1 and S2 in the Supporting Information for details). The IC_{50} values (half-maximal inhibitory concentrations) of these selected inhibitors were at least lower than 10 μM. The three-dimensional structures of the inhibitors were built using Sybyl 6.91 and were then subjected to energy minimization using a conjugate gradient minimization algorithm with the Tripos force field until a gradient convergence of 0.01 kcal/(mol·Å) was achieved.

In addition to the collection of known inhibitors, a large pool of confirmed inactive compounds was also needed. However, large numbers of experimentally confirmed inactive molecules are generally not available. It is a common practice to randomly select molecules from a large molecule database and consider them to be inactive. The molecular weights of decoys chosen for the analysis have a significant impact on the evaluation. Decoys with low molecular weights achieve lower docking scores on average and are therefore likely to obtain lower rankings than larger active counterparts. To avoid the influence of molecular weight, we selected decoy molecules with molecular weights in the same range as the chosen inhibitors. Two thousand molecules with molecular weights in the range of 300−550 Da were randomly selected from the ACD-3D database (Available Chemical Database, Release ACD 3D 2002.2) and considered as noninhibitors of neuraminidase, and 2000 molecules with a molecular weight range of 350−500 Da from ACD-3D were randomly selected as the noninhibitors of cyclooxygenase-2.

Molecular Docking. AutoDock 3.05 was used to sample the binding energy landscapes of small molecules and their corresponding proteins. The protein coordinates used for docking were taken from the X-ray crystal structures of influenza virus neuraminidase in complex with 4-acetamido-3-hydroxy-5-nitro-benzoic acid (PDB entry: 1ivd) and of cyclooxygenase-2 in complex with 4-[5-(4-bromophenyl)-3-(trifluoromethyl) pyrazol-1-yl] benzenesulfonamide (PDB entry: 6cox). The coordinates of proteins were prepared using Sybyl 6.91, and all ligands were removed. The coordinates were converted to AutoDock format files using the graphical user interface AutoDock Tools (ADT). All other atomic parameters were generated automatically by ADT. Grids of 60 Å × 60 Å × 60 Å with 0.375-Å spacing, centered at the center of mass of the corresponding small-molecule inhibitors in the crystals were calculated for 10 ligand atom types using AutoGrid. The binding conformations of the inhibitors and noninhibitors in both enzymes were predicted by AutoDock 3.05. For each small molecule, 200 separate docking calculations were performed. Each docking calculation consisted of 25 million energy evaluations and a maximum of 270000 generations using the Lamarckian genetic algorithm local search method. All dockings described in this article were performed with a population size of 300, and 300 rounds of Solis and Wets local search were applied with a probability of 0.06. A mutation rate of 0.02 and a crossover rate of 0.8 were used to generate new docking trials for subsequent generations, and the best individual from each generation was propagated to the next generation.

Choice of the Native Binding Conformations. We assumed that, in the binding energy landscape, the distributions of native binding conformations relative to all of the other binding conformations should be different for the true binders and the high-scoring decoys. Thus, it is the first step to choose native binding conformations. The best choice would be the binding conformations in the known complex crystal structures; however, it is unlikely that complex structures are available for each compound. In molecular docking studies, the top-scoring molecular conformation among all of the predicted binding conformations is generally accepted as the native binding conformation. However, because of approximations in deriving the docking scores, uncertainty exists in estimating the binding energy values using scoring functions. The native binding conformation does not always have the highest docking score, and sometimes, it has a score that is slightly below the highest.[41] Another method to identify the native binding conformations is to cluster the binding conformations according to the average root-mean-square deviation (rmsd) first and then to assume that the conformation with the largest occupancy is most likely the native binding conformation (where "occupancy" refers to the number of conformations that are grouped in one cluster). Considering the assumption of the largest occupancy and the uncertainty in docking score, we used the following strategy to select the native binding conformation from the 200 binding conformations generated for each compound: first, all possible binding conformations of a ligand were ordered according to their scores; then, starting from the conformation with the highest score, the rmsd of each conformation with respect to all the previous higher-score conformations was computed. (Note that all of the heavy atoms and polar hydrogen atoms were considered in our calculation of the rmsd.) If the rmsd was lower than a given threshold (2.0 Å), then the compared conformation would be considered as belonging to the cluster of the reference binding conformation; otherwise, it was added to a list of distinct ligand binding conformation clusters. In this way, 200 conformations of each molecule were divided into a list of clusters, rmsd values between representative conformations of different clusters were larger than the given threshold, and the number of docking conformations clustered to each distinct representative binding conformation was defined as the occupancy. Then, the representative binding conformations with scores less than 1.0 kcal/mol below the highest score were used as candidates, from which the one with the largest occupancy was selected to be the native binding conformation.

Construction of Inhibitor Data Sets and High-Scoring Decoy Data Sets. Using the approach proposed above, the native binding conformation of each compound was selected, and the docking score of the selected conformation was used as an approximation of the binding free energy. The compounds with very low scores clearly have great probability to be noninhibitors; however, it is difficult to judge whether the compounds with high scores are true ligands. In the present study, we focused on true binders and these high-scoring decoys.

We hypothesized that, in the case of protein−ligand binding, noticeable differences in the binding energy landscapes of inhibitors and high-scoring decoys should be observed. For neuraminidase, the calculated docking scores of the 77 inhibitors were in the range from −14.9 to −9.95 kcal/mol; thus, after the noninhibitors with smaller absolute docking scores had been deleted, 552 noninhibitors with docking scores below −10.0 kcal/mol were selected as high-scoring decoys. For cyclooxygenase-2, the docking scores for most inhibitors (about 95% of the inhibitors in the training set) were lower than −10.0 kcal/mol, so we also used −10.0 kcal/mol as a cutoff, and 836 noninhibitors with docking scores below −10.0 kcal/mol served as the high-scoring decoys. All of the inhibitors and the high-scoring noninhibitors formed data set_1.

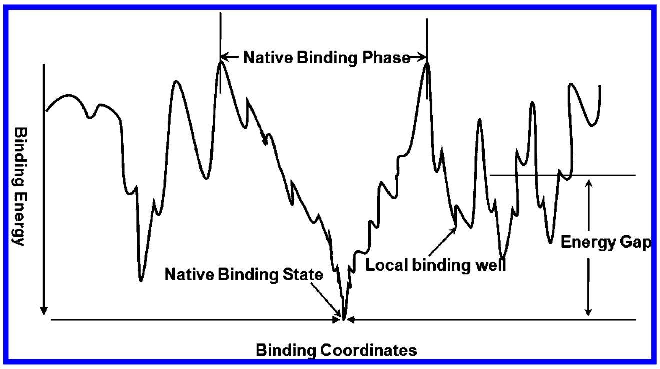

Figure 1. Illustration of the funneled energy landscape of biomolecular binding.

PARAMETERS DESCRIBING THE BINDING ENERGY LANDSCAPE

Definitions of the Native Binding Phase and the Denatured Binding Phase. In the folding and unfolding process of many small single-domain proteins, such as chymotrypsin inhibitor 2, at most two main states, the native and the denatured states, are observed on the longest time scales under physiological conditions. The denatured states correspond to a high-entropy, high-energy, and disordered phase, whereas the native folded states correspond to a lower-entropy and low-energy phase.[42−44] For the case of protein−ligand binding, we assumed that the available binding conformations of a small molecule could also be classified into a "native binding phase" and a "denatured binding phase" (Figure 1). The conformations within a certain rmsd from the native binding conformation were considered to be in the native binding phase (Figure 1), and all other conformations were classified as being in the denatured binding phase. The predefined rmsd values describe the breadth of the native binding phase, and the breadth depends on the property of the protein binding pocket studied. This value can be defined from experience, and different rmsd cutoffs (from 4 to 10 Å, at 0.5-Å intervals) were tested.

Thermodynamic Stability. In protein folding studies, the energy gap between the native state and all of the other conformations is used as a measure of the thermodynamic stability of the native state, which is used to judge whether the sequence has a stable folding state.[30] In the present study, the thermodynamic stability of the native binding conformation was measured by the energy gap between the binding energy of the native binding state and the average binding energy of the binding modes in the denatured binding phase (Figure 1). This energy gap was used as a criterion to distinguish inhibitors from noninhibitors by assuming that the larger the energy gap, the higher the probability that the small molecule is an inhibitor. For compounds with all of the sampled binding conformations located in the native binding phase, their binding thermodynamic stability was considered to be large enough to attract all of the initial binding configurations to the native binding phase. In such cases, we set the energy gap to be −5 kcal/mol. Considering the advantage of receiver operator characteristic (ROC) curves in evaluating the performance of different virtual screening protocols over other methods,[45] we used ROC curves and the area under the curves (AUC) to describe the discrimination capability of the energy gaps derived at different rmsd cutoffs.

Kinetic Accessibility Parameter. The energy gap quantifies only the stability of the native binding conformation, but the barriers along the way down to the native binding model control its kinetic accessibility. Evidence from theory and simulation indicates that amino acid sequences with minimal frustrations are more likely to fold than those with a rough energy landscape.[32] When it comes to binding, inhibitors are expected to exhibit minimally obstructed pathways to the native binding conformations in the binding energy landscapes leading to a stable binding mode, whereas noninhibitors have rougher energy landscapes leading to multiple binding modes. The obstructions can be described by the number of local binding wells (Figure 1) along the energy landscape, and a large number of binding wells and high-energy barriers would cause more difficulties during the association process.

In the present study, the sampled binding conformations in the binding energy landscape of a small molecule were clustered into several distinctive local binding wells. These wells were regarded as obstructions in the process of binding and interfered with the path to the native binding state. More obstructions presented more challenges for the binding process. The number of local binding wells was regarded as a measurement of the kinetic accessibility of binding in the present study. Inhibitors were expected to have fewer local binding wells in the landscapes than noninhibitors. The standard of clustering the sampled binding conformations into local binding wells depends on the binding property of the particular protein under investigation, and different rmsd cutoffs (rmsd = 1.0, 1.5, 2.0, 2.5, and 3.0 Å) were tested. The clustering method was similar to that described in the process of obtaining the native binding conformations, and the initial reference conformations were set to be the selected native binding conformations. ROC curves and AUC values were used to describe the discrimination capability of the number of local binding wells at different rmsd cutoffs. To explore the influence of the cutoff values, the discrimination capability of the number of local binding wells clustered at other cutoffs was also tested (Table S3 in the Supporting Information).

Combination of the Parameters Proposed above and Docking Scores. Thermodynamic stability and kinetic accessibility are closely related aspects that determine whether a reaction can really occur. To test their cooperativity for inhibitor discrimination, linear combinations of the energy gap and the number of local binding wells were analyzed using logistic regression [a linear method in the glm (generalized linear model) package of R]. This method is a variation of ordinary regression and is used when the dependent (response) variable is dichotomous (i.e., takes only two values, which usually represent the occurrence or nonoccurrence of some outcome event, usually coded as 0 or 1) and the independent (input) variables are continuous, categorical, or both. In the present study, the dependent variable was defined as 1 (inhibitors) or 0 (high-scoring decoys), and the independent variables were the energy gap and the number of local binding wells. Further combinations of these two parameters with the docking scores derived from the native binding conformations were also tested.

Cross-Validation and Independent Test Sets. To validate the prediction capabilities of different parameter combinations derived from the logistic regressions, five-fold cross-validation was employed to analyze the constructed data (data set_1). Inhibitors and high-scoring decoys were randomly divided into five groups. The data for four groups of inhibitors and four groups of high-scoring noninhibitors were combined to form the training set to derive a model to predict the data in the fifth groups. In turn, all of the data were predicted by the models derived from 80% of the data. To avoid the randomness generated from one round of five-fold cross-validation, five-fold cross-validations were performed 10 times. Average true positive and false positive rate values derived from every cutoff value were utilized to derive ROC curves. AUC values for the ROC curves were utilized to describe the discrimination capabilities of different parameter combinations.

To further test the predictive capabilities of the derived models, a directory of useful decoys (DUD),[46,47] the largest and most comprehensive public data set for benchmarking virtual screening programs by far, was used as an independent test set. The DUD contains inhibitors and corresponding decoys for 40 different targets. During construction of the decoys, only compounds with Tanimoto coefficients (based on CACTVS type 2 fingerprints) less than 0.9 with respect to any annotated inhibitors were selected, and the molecular weight, number of hydrogen-bond acceptors, number of hydrogen-bond donors, number of rotable bonds, log p value, and number of important functional groups (amine, amide, amidine, and carboxylic acid) were used in the similarity analysis to select the most similar decoys for each inhibitor. The DUD presented a challenge to discriminate the inhibitors from the corresponding decoys. There are 49 inhibitors of neuraminidase, 349 inhibitors of cyclooxygenase-2, and their corresponding decoys in DUD. AutoDock 3.05 was used to sample the binding energy landscapes of these small molecules and the corresponding enzymes. For simplicity of calculation, ligands with more than six different atom types were not included. The possible native binding conformations of these molecules and the corresponding parameters were determined by the procedure described above. Thirty-one inhibitors of neuraminidase and the corresponding 1734 decoys and 235 inhibitors of cyclooxygenase-2 and the corresponding 7751 decoys constituted the independent test sets (data set_2) for these two enzymes. (The inhibitors in data set_1 were not included in the independent test set.) Models derived from the full data in data set_1 were used to analyze the data in the independent data set (data set_2) and to discriminate inhibitors and noninhibitors.

RESULTS

Distribution of Docking Scores. In the example of neuraminidase, the docking scores of inhibitors in the data set _I were well separated from those of the high-scoring decoys in data set_1 (p value = 1.31×10^{8} by the t test) (Figure 2a). To evaluate the discrimination capability of the docking score on the training set, ROC curves were utilized to examine the sensitivity and specificity over the entire range

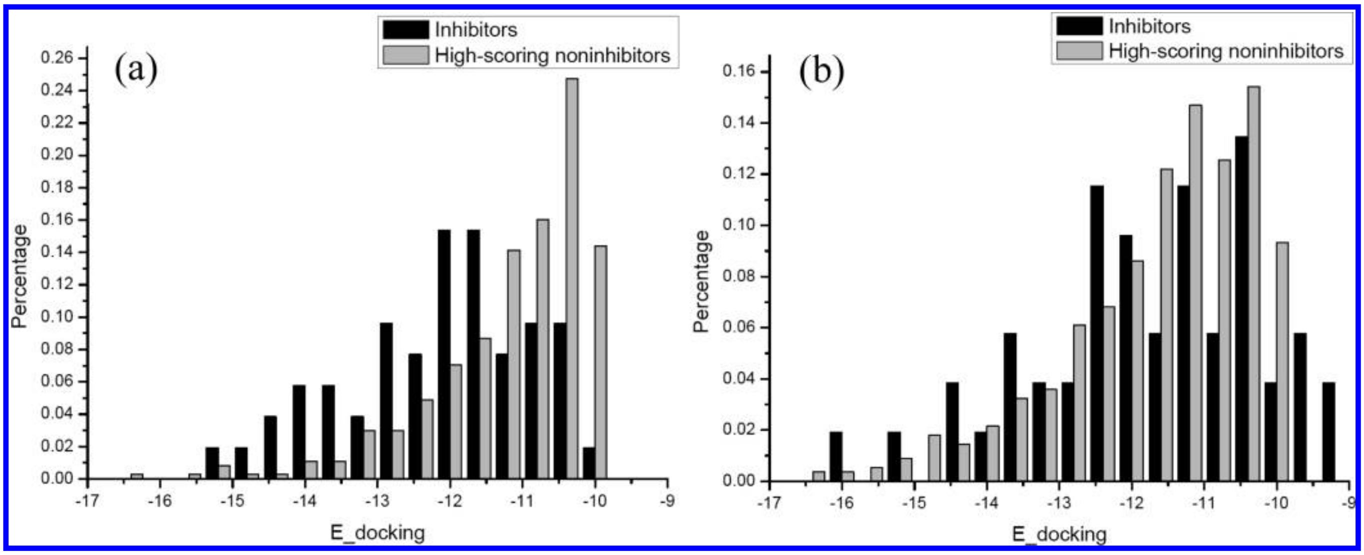

Figure 2. Docking score distribution for inhibitors and high-scoring noninhibitors in data set_1: (a) Neuraminidase system, (b) cyclooxygenase-2 system.

Table 1. Areas under the ROC Curves Calculated for Energy Gaps Obtained at Different rmsd Thresholds for the Inhibitors and the Decoys of Neuraminidase in Data Set_1

rmsd cutoff (Å)	AUC[a]	rmsd cutoff (Å)	AUC[a]
4.0	0.524	7.5	0.763
4.5	0.522	8.0	0.713
5.0	0.577	8.5	0.651
5.5	0.643	9.0	0.604
6.0	0.762	9.5	0.553
6.5	0.800	10.0	0.522
7.0	0.798		

[a] AUC = area under the curve.

Table 2. Areas under the ROC Curves Calculated for Energy Gaps Obtained at Different rmsd Thresholds for the Inhibitors and the Decoys of Cyclooxygenase-2 in Data Set_1

rmsd cutoff (Å)	AUC[a]	rmsd cutoff (Å)	AUC[a]
4.0	0.556	7.5	0.722
4.5	0.556	8.0	0.752
5.0	0.570	8.5	0.743
5.5	0.560	9.0	0.737
6.0	0.575	9.5	0.735
6.5	0.622	10.0	0.726
7.0	0.684		

[a] AUC = area under the curve.

(Figure 5a below). The area under the ROC curve (AUC) calculated from the docking score was 0.735, and when the docking score threshold was set at −11.5 kcal/mol (that is, compounds with docking scores lower than −11.5 kcal/mol were considered to be inhibitors, and compounds with docking scores higher than −11.5 kcal/mol were considered to be noninhibitors), 71% of true positives and 71% of true negatives were identified. This indicates a relatively good discrimination capability of the AutoDock 3.05 score function for inhibitors and high-scoring noninhibitors of neuraminidase. In contrast, in the case of cyclooxygenase-2, no apparent differences were found in the distribution of the docking scores of the high-scoring decoys and the inhibitors in data set_1 (p value = 0.57 by the t test) (Figure 2b). The small area under the ROC curve (Figure 5b below) calculated for the docking score reflected the poor capability of the scoring function in distinguishing the high-scoring decoys and inhibitors of cyclooxygenase-2.

Energy Gap Distribution. The energy gap between the docking score of the native binding conformation and the average docking score of the conformations in the denatured binding phase was utilized to measure the thermodynamic stability of the native binding modes. For the two proteins, the energy gap distribution of inhibitors was distinctly different from that of the high-scoring decoys when the rmsd values were in a certain range, and the difference could be seen from the areas under the ROC curves (Tables 1 and 2). This difference was the largest when the rmsd values were set at 6.5 Å for neuraminidase and 8.0 Å for cyclooxygenase-2 (Figure 3a,b). The conformations with rmsd values from the native binding conformation larger than 6.5 and 8 Å were classified as denatured phases for neuraminidase and cyclooxygenase-2, respectively, and the average scores of conformations in the denatured phases were used to calculate the energy gaps.

Distribution of the Number of Local Binding Wells. The kinetic accessibility of binding was described by the number of local binding wells in the binding energy landscape. The areas under the ROC curves measured the discrimination capabilities of the parameters derived at different rmsd thresholds (Table S3, Supporting Information). For the two systems, the best rmsd varied. For neuraminidase, a maximum AUC value (0.888) was observed when the rmsd was set at 0.5 Å. In the case of cyclooxygenase-2, the maximum AUC value was 0.714 when the rmsd cutoff was set at 3.0 Å. As the rmsd thresholds also depend on other factors, such as the sampling in docking calculations, a unified value was used for comparable purposes. In the literature, the rmsd value of 2.0 Å is commonly used as a criterion to judge whether two conformations belong to one binding state. We used the numbers of local binding wells derived at rmsd = 2.0 Å for the two proteins in further analysis. The distributions of the number of local binding wells derived at rmsd = 2.0 Å are shown in Figure 4a,b. When the rmsd values were set at 2.0 Å, the AUC value for the number of local binding wells was 0.727 for neuraminidase and 0.691 for cyclooxygenase-2.

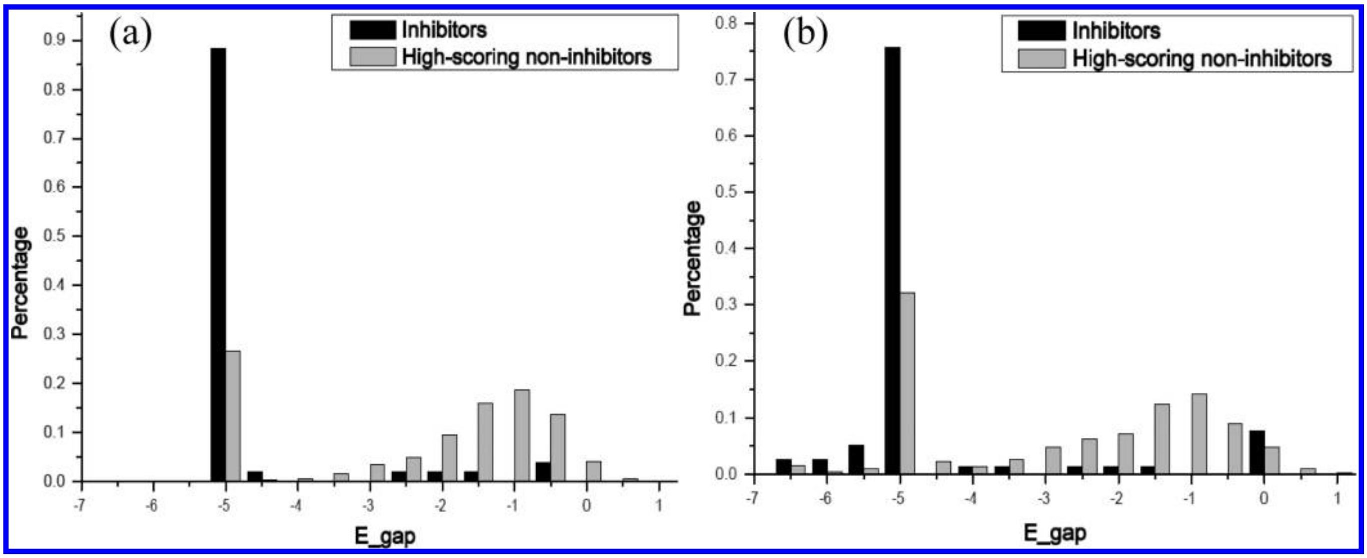

Figure 3. Energy gap distribution for inhibitors and high-scoring noninhibitors in data set_1: (a) Neuraminidase system (E_gap derived at rmsd = 6.5 Å), (b) cyclooxygenase-2 system (E_gap derived at rmsd = 8.0 Å).

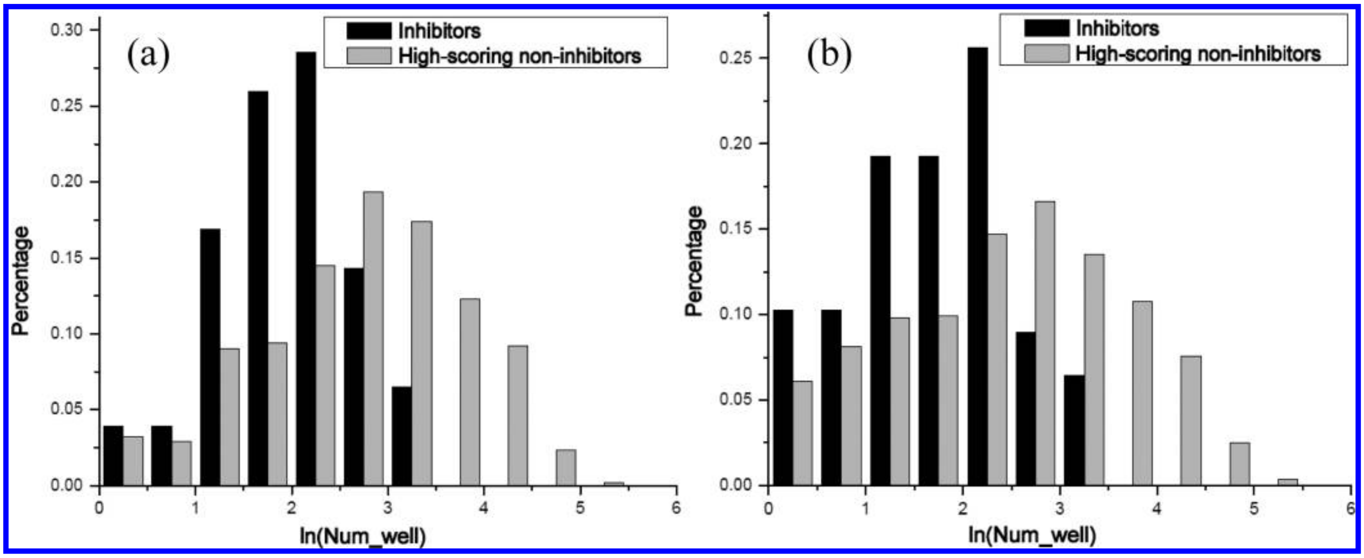

Figure 4. Distribution of the number of local binding wells for inhibitors and high-scoring noninhibitors in data set_1. The value on the *x* axis is the natural logarithm of the number of local binding wells. (a) Neuraminidase system (Num_well derived at rmsd = 2.0 Å), (b) cyclooxygenase-2 (Num_well derived at rmsd = 2.0 Å).

Parameter Combinations and Their Discrimination Capability. Various linear combinations of docking score, energy gap, and number of local binding wells were analyzed by logistic regression with five-fold cross-validation. Models derived from full data set_1 were used to analyze the data in the independent data set (data set_2) and to judge whether the compounds were inhibitors or noninhibitors.

For data set_1, generally speaking, the performance of two-parameter combinations was better than the corresponding single-parameter models, and combination of the three parameters improved the performance further. For data set_1, the AUC values of the three-parameter combination model were 0.878 for neuraminidase (Table 3) and 0.776 for cyclooxygenase-2 (Table 4).

Considering the high similarity between the inhibitors and the decoys in the DUD database,[46,47] data set_2 presented a great challenge for scoring functions to make a discrimination. For the compounds in the DUD database, the AUC values of E_docking were only 0.648 for neuraminidase (Table 3) and 0.641 for cyclooxygenase-2 (Table 4). However, in the neuraminidase system, E_gap had a significant discrimination capability with an AUC of 0.709, and Num_well gave an AUC of 0.640. Although slightly lower than the values for data set_1, they followed a similar order. The combination of the three parameters gave an AUC

Table 3. Performance of Logistic Regression Models for the Neuraminidase System

	AUC[a]	
models using different parameter combinations[b]	cross-validation	independent test set
E_docking	0.718	0.648
E_gap	0.782	0.709
Num_well	0.721	0.640
E_docking + E_gap	0.857	0.755
E_docking + Num_well	0.837	0.686
E_gap + Num_well	0.807	0.713
E_docking + E_gap + Num_well	0.878	0.750

[a] AUC = area under the curve. [b] E_docking = docking score, E_gap = energy gap, and Num_well = number of local binding wells.

Table 4. Performance of Logistic Regression Models for the Cyclooxygenase-2 System

	AUC[a]	
models using different parameter combination[b]	cross-validation	independent test set
E_docking	0.549	0.641
E_gap	0.745	0.813
Num_well	0.684	0.821
E_docking + E_gap	0.722	0.794
E_docking + Num_well	0.691	0.825
E_gap + Num_well	0.750	0.845
E_docking + E_gap + Num_well	0.776	0.855

[a] AUC = area under the curve. [b] E_docking = docking score, E_gap = energy gap, and Num_well = number of local binding wells.

value of 0.750 for neuraminidase (Table 3 and Figure 5c). In the case of cyclooxygenase-2, the AUC value of the energy gap was 0.813, the AUC value of the number of local binding wells was 0.821, and the AUC value of the three-parameter combination model was 0.855 (Table 4 and Figure 5d).

DISCUSSION

Binding Coordinates, the Native Binding Phase, and the Clustering of Local Binding Wells. In the field of protein folding, properties derived from the folding energy landscape have been shown to depend heavily on the selection of reaction coordinate to present a "folding funnel".[48−51] Arbitrary selection could generate a misleading description of the energy landscape. Because of the objectivity and easily automated calculation, the rmsd between different binding modes is widely used to measure the similarity between two binding modes and was used to define the binding coordinates in the present study. Scales of native binding phases and standards to cluster possible binding conformations into local binding wells were set based on the rmsd.

We measured the thermodynamic stability of the native binding mode using the energy gap between the binding energy of the native binding mode and the average binding energy of all of the binding configurations in the denatured binding phase. Binding pockets in different proteins can have distinct properties, and the scale for defining the size of the

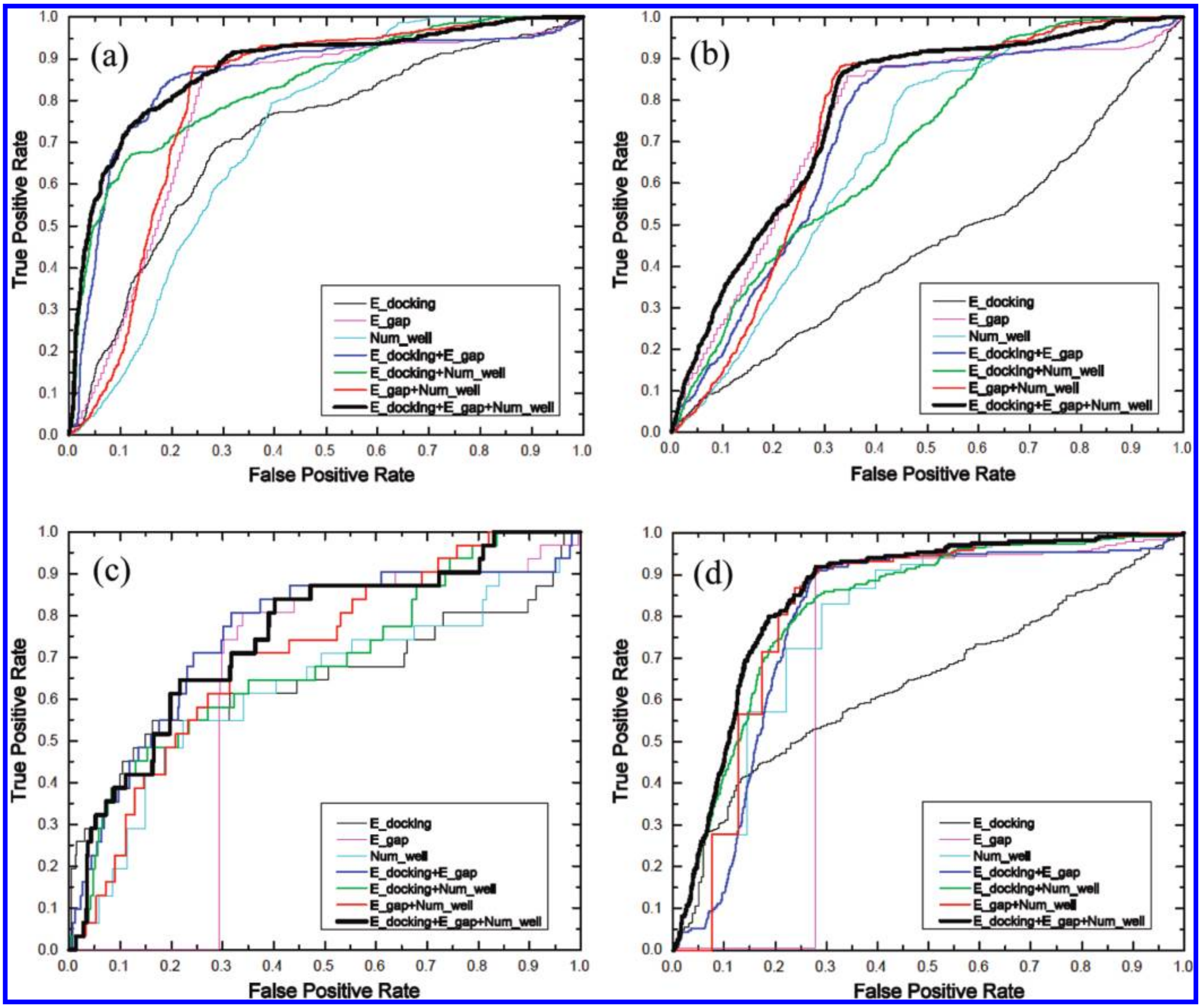

Figure 5. ROC curves for docking score, energy gap, number of local binding wells, and their combination. In the ROC curve, the x axis is the false positive rate, and the y axis is the true positive rate. The area under the curve characterizes the classification capability. (a) Data set_1 of neuraminidase, (b) data set_1 of cyclooxygenase-2, (c) data set_2 of neuraminidase, (d) data set_2 of cyclooxygenase-2.

Table 5. Coefficients from the Normalized Logistic Regression Models for Neuraminidase

model	intercept	E_docking	E_gap	Num_well
E_docking + E_gap	−3.01 ± 0.26	−0.59 ± 0.12	−1.58 ± 0.21	
E_docking + Num_well	−3.36 ± 0.32	−1.16 ± 0.14		−2.90 ± 0.48
E_gap + Num_well	−2.92 ± 0.25		−1.20 ± 0.19	−0.98 ± 0.35
E_docking + E_gap + Num_well	−3.48 ± 0.32	−1.02 ± 0.15	−1.06 ± 0.21	−1.91 ± 0.43

Table 6. Coefficients from the Normalized Logistic Regression Models for Cyclooxygenase-2

model	intercept	E_docking	E_gap	Num_well
E_docking + E_gap	−2.68 ± 0.15	−0.18 ± 0.12	−0.91 ± 0.14	
E_docking + Num_well	−3.01 ± 0.22	−0.34 ± 0.14		−1.94 ± 0.38
E_gap + Num_well	−3.08 ± 0.21		−1.05 ± 0.18	−1.00 ± 0.23
E_docking + E_gap + Num_well	−3.13 ± 0.22	−0.23 ± 0.14	−1.05 ± 0.18	−1.16 ± 0.28

binding phases depends on the properties of specific systems. Verkhivker et al. investigated the thermodynamic and kinetic aspects of molecular recognition for the methotrexate−dihydrofolate reductase system using the binding energy landscape approach.[35] They found that the native binding phase of the system extends to nearly 5.0 Å rmsd from the native structure, that no significant free energy barriers exist within this region, and that the binding domain could extend to 7.0 Å with only moderate barriers. The definition of the native binding phases can depend on human experience. For neuraminidase, when the rmsd cutoff value was in the range of 6.0−8.0 Å, the energy gap parameter could differentiate between the inhibitors and the high-scoring noninhibitors, and for cyclooxygenase-2, the valid energy gap could be obtained when the rmsd cutoffs were larger than or equal to 7.5 Å. When the boundaries of native binding phases were set at rmsd = 6.5 Å for neuraminidase and rmsd = 8.0 Å for cyclooxygenase-2, the energy gap showed the greatest capability to discriminate the inhibitors from the high-scoring decoys. For proteins for which no inhibitors have been reported, we suggest that the size of the native binding phase be set at rmsd = 6−8 Å and that the rmsd values used to cluster binding wells be set at 1−3 Å.

In the present study, the energy gap was calculated as the energy difference between the binding energy of the native binding mode and the average binding energy of all of the binding configurations in the denatured binding phase. As the native binding state is not a single binding conformation, but a cluster of similar binding conformations, it would be more representative to calculate the energy gap by using the average binding energy of the whole cluster instead of the binding energy of one single binding conformation. For these two systems, we attempted to use the average binding energy of binding conformations in the native binding phase as the binding energy of the native binding state to calculate the energy gap. The discrimination capabilities of the energy gaps derived at different rmsd values defining the size of native binding phase were calculated and are reported in Tables S4 and S5 in the Supporting Information. The AUC values of the newly calculated energy gaps were similar to the values that used the binding energy of the assumed native binding conformation as the binding energy of native binding state. Considering the conventions in molecular docking studies and the convenience for calculation, we still used the binding energy of the assumed native conformation as the binding energy of the native binding state.

Contributions of Different Parameters in the Regression Models. To investigate the contributions of different parameters, the parameters were normalized, and their combinations were analyzed by logistic regression (Tables 5 and 6). The negative coefficients of energy gap showed that compounds with larger energy gaps were more likely to be inhibitors, and the negative coefficients of the number of local binding wells suggested that less competitive binding modes (intermediates) benefit binding. The coefficient of E_docking in the three-parameter model of neuraminidase (−1.02) and in the model for cyclooxygenase-2 (−0.23) suggested that the docking score played a more important role in the neuraminidase system than in the cyclooxygenase-2 system. This is because the docking score in the neuraminidase system had a certain discrimination capability for the inhibitors and the high-scoring decoys, which was not the case for cyclooxygenase-2. The coefficient ratio between the number of local binding wells and the energy gap in the neuraminidase system was greater than the coefficient ratio for cycloxygenase-2, which shows that kinetic accessibility plays a more important role for the inhibition of neuraminidase than for the inhibition of cycloxygenase-2.

Possible Applications in Virtual Screening. To evaluate the performance of the parameters in practical screening, we checked the number of inhibitors that were ranked in the top 10% of compounds chosen by the five-fold cross-validation of the scoring function and three-parameter combination models. In data set_1 for neuraminidase, there were 16 inhibitors in top 10% of compounds chosen by the scoring function. According to the formula

$$\text{enrichment ratio} = (\text{Hits}_{\text{sampled}}/N_{\text{sampled}})/(\text{Hits}_{\text{total}}/N_{\text{total}})$$

the enrichment ratio of the scoring function was 2.05, which means that the success ratio was 2.05 times that of random screening. In comparison, 41 inhibitors were chosen by the three-parameter combination model, and the enrichment ratio was 5.26, more than twice that of the scoring function. In the case of cyclooxygenase-2, among the top 10% of compounds chosen by the docking score for data set_1, there were 11 inhibitors, and the enrichment ratio was 1.41. Twenty-one inhibitors were found in 10% of the compounds chosen by the three-parameter combination model, and the enrichment ratio was 2.69. The enrichment of the three-parameter combination models was 200−300% of the enrichment of the scoring function. Thus, binding energy landscape analysis should greatly benefit the process of virtual screening.

In the present study, we used AutoDock, a popular molecular docking program, to sample the possible binding conformations of small molecules with proteins. It is clear that this approach cannot fully sample the whole binding energy landscape, and different docking programs and scoring functions could give discrepant results. However, useful information reflecting the binding process can be obtained and applied in discriminating true inhibitors from high-scoring decoys, as shown in the case of neuraminidase and COX-2 using our strategy. Exhaustive sampling methods might provide more information, but might not be practicable for computer-aided drug design applications because of computational cost.

To test the influences of different scoring functions, we also employed an empirical scoring function, SCORE 2.0,[52] to rescore the sampled conformations from AutoDock. Similar results were obtained (see Tables S6−S9 in the Supporting Information for details).

CONCLUSIONS

One of the bottlenecks for virtual screening by molecular docking is determining how to correctly choose true hits from among high-scoring decoys. Inspired by the success of energy landscape analysis in protein folding and biomolecular binding studies, we selected two parameters reflecting the thermodynamic stability and the kinetic accessibility of the binding energy landscape and successfully applied them in discriminating inhibitors from high-scoring decoys for neuraminidase and cyclooxygenase-2. The success was due to exploration of the binding process characterization by energy landscape analysis, instead of using only the final binding conformation. Our study confirmed that binding is determined not just by the native structure of the complex, but rather by the general properties of the binding process. We expect that the success rate can be further increased by incorporating more details about the binding mechanism into virtual screening.

Based on our analysis, we suggest that it is better to do a binding energy landscape analysis after a large scale of virtual screening. When there are existing inhibitors or other known ligands, a procedure similar to that described in the present study can be used. When no ligands have been reported for the target protein, the native binding phase can be defined in the range of 6−8 Å, and the rmsd used to cluster the conformations to obtain the number of local binding wells can be set in the range of 1.0−3.0 Å. Compounds with large energy gaps and small numbers of local binding wells have a greater probability of being true binders.

Abbreviations: COX-2, cyclooxygenase-2; AUC, area under the receiver operator characteristic curve; ROC, receiver operator characteristic.

ACKNOWLEDGMENT

This project was supported in part by the Ministry of Science and Technology of China and the National Natural Science Foundation of China. The authors thank Professor Zhirong Liu and other members of the Center for Theoretical Biology, Peking University, for helpful discussions and Professor Iain Bruce for his generous help in language improvement.

Supporting Information Available: Collected structures for the inhibitors of neuraminidase and cyclooxygenase-2, results calculated for the number of local binding wells obtained at other different rmsd thresholds for data set_1, results obtained using the average binding energy of the conformations in the native binding phase as the binding energy of the native binding state for data set_1, and rescore analysis using SCORE 2.0 for the compounds in data set_1 and data set_2. This material is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006**, *11*, 580–594.

(2) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.

(3) Tame, J. R. H. Scoring functions−The first 100 years. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 445–451.

(4) Krumrine, J. R.; Maynard, A. T.; Lerman, C. L. Statistical tools for virtual screening. *J. Med. Chem.* **2005**, *48*, 7477–7481.

(5) Ruvinsky, A. M.; Kozintsev, A. V. New and fast statistical-thermodynamic method for computation of protein−ligand binding entropy substantially improves docking accuracy. *J. Comput. Chem.* **2005**, *26*, 1089–1095.

(6) Ruvinsky, A. M. Role of binding entropy in the refinement of protein−ligand docking predictions: Analysis based on the use of 11 scoring functions. *J. Comput. Chem.* **2007**, *28*, 1364–1372.

(7) Ruvinsky, A. M. Calculations of protein−ligand binding entropy of relative and overall molecular motions. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 361–370.

(8) Tao, P.; Lai, L. H. Protein ligand docking based on empirical method for binding affinity estimation. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 429–446.

(9) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein−ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.

(10) Smith, R.; Hubbard, R. E.; Gschwend, D. A.; Leach, A. R.; Good, A. C. Analysis and optimization of structure-based virtual screening protocols (3). New methods and old problems in scoring function design. *J. Mol. Graphics Modell.* **2003**, *22*, 41–53.

(11) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. Improving database enrichment through ensemble docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621–627.

(12) 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.

(13) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the essential features of a protein surface for improving protein−ligand docking, scoring, and virtual screening. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 883–902.

(14) Butini, S.; Campiani, G.; Borriello, M.; Gemma, S.; Panico, A.; Persico, M.; Catalanotti, B.; Ros, S.; Brindisi, M.; Agnusdei, M.; Fiorini, I.; Nacci, V.; Novellino, E.; Belinskaya, T.; Saxena, A.; Fattorusso, C. Exploiting protein fluctuations at the active-site gorge of human cholinesterases: Further optimization of the design strategy to develop extremely potent inhibitors. *J. Med. Chem.* **2008**, *51*, 3154–3170.

(15) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jorgensen, F. S. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein−ligand binding affinities. *J. Med. Chem.* **2001**, *44*, 2333–2343.

(16) Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* **2006**, *26*, 531–568.

(17) Steinbrecher, T.; Case, D. A.; Labahn, A. A multistep approach to structure-based drug design: Studying ligand binding at the human neutrophil elastase. *J. Med. Chem.* **2006**, *49*, 1837–1844.

(18) Beautrait, A.; Leroux, V.; Chavent, M.; Ghemtio, L.; Devignes, M. D.; Smaiel-Tabbone, M.; Cai, W.; Shao, X.; Moreau, G.; Bladon, P.; Yao, J.; Maigret, B. Multiple-step virtual screening using VSM-G: Overview and validation of fast geometrical matching enrichment. *J. Mol. Model. Rev.* **2008**, *14*, 135–148.

(19) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.

(20) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.

(21) Wolber, G.; Langer, T. LigandScout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160–169.

(22) Springer, C.; Adalsteinsson, H.; Young, M. M.; Kegelmeyer, P. W.; Roe, D. C. PostDOCK: A structural, empirical approach to scoring protein ligand complexes. *J. Med. Chem.* **2005**, *48*, 6821–6831.

(23) Garmendia-Doval, A. B.; Morley, S. D.; Juhos, S. Post Docking Filtering Using Cartesian Genetic Programming. In *Artificial Evolution*; Springer: Berlin, 2004; Vol. 2936, Chapter 16, pp 189−200.

(24) Kollman, P. Free-Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395–2417.

(25) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Freer, S. T.; Rose, P. W. Complexity and simplicity of ligand−macromolecule interactions: The energy landscape perspective. *Curr. Opin. Struct. Biol.* **2002**, *12*, 197–203.

(26) Shakhnovich, E.; Farztdinov, G.; Gutin, A. M.; Karplus, M. Protein Folding Bottlenecks: A Lattice Monte Carlo Simulation. *Phys. Rev. Lett.* **1991**, *67*, 1665–1668.

(27) Shakhnovich, E. I.; Gutin, A. M. Engineering of Stable and Fast-Folding Sequences of Model Proteins. *Proc. Natl. Acad. Sci. U S A.* **1993**, *90*, 7195–7199.

(28) Shakhnovich, E. I. Proteins with Selected Sequences Fold into Unique Native Conformation. *Phys. Rev. Lett.* **1994**, *72*, 3907–3910.

(29) Sali, A.; Shakhnovich, E.; Karplus, M. Kinetics of Protein Folding. A Lattice Model Study of the Requirements for Folding to the Native State. *J. Mol. Biol.* **1994**, *235*, 1614–1636.

(30) Sali, A.; Shakhnovich, E.; Karplus, M. How Does A Protein Fold. *Nature* **1994**, *369*, 248–251.

(31) Ebeling, M.; Nadler, W. On Constructing Folding Heteropolymers. *Proc. Natl. Acad. Sci. U S A.* **1995**, *92*, 8798–8802.

(32) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins* **1995**, *21*, 167–195.

(33) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand−protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731–751.

(34) Bouzida, D.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Gehlhaar, D. K.; Larson, V.; Luty, B. A.; Rejto, P. A.; Rose, P. W.; Verkhivker, G. M. Thermodynamics and kinetics of ligand−protein binding studied with the weighted histogram analysis method and simulated annealing. *Pac. Symp. Biocomput. '99* **1999**, *4*, 426–437.

(35) Verkhivker, G. M.; Rejto, P. A.; Bouzida, D.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Gehlhaar, D. K.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Towards understanding the mechanisms of molecular recognition by computer simulations of ligand−protein interactions. *J. Mol. Recognit.* **1999**, *12*, 371–389.

(36) Verkhivker, G. M.; Rejto, P. A.; Gehlhaar, D. K.; Freer, S. T. Exploring the energy landscapes of molecular recognition by a genetic algorithm: Analysis of the requirements for robust docking of HIV-1 protease and FKRP-12 complexes. *Proteins* **1996**, *25*, 342–353.

(37) Rejto, P. A.; Verkhivker, G. M. Unraveling principles of lead discovery: From unfrustrated energy landscapes to novel molecular anchors. *Proc. Natl. Acad. Sci. U S A.* **1996**, *93*, 8945–8950.

(38) Verkhivker, G. M.; Rejto, P. A. A mean field model of ligand protein interactions: Implications for the structural assessment of human immunodeficiency virus type 1 protease complexes and receptor-specific binding. *Proc. Natl. Acad. Sci. U S A.* **1996**, *93*, 60–64.

(39) Wang, J.; Verkhivker, G. M. Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Phys. Rev. Lett.* **2003**, *90*, 188101–188104.

(40) Wang, J.; Zheng, X.; Yang, Y.; Drueckhammer, D.; Yang, W.; Verkhivker, G.; Wang, E. Quantifying intrinsic specificity: A potential complement to affinity in drug screening. *Phys. Rev. Lett.* **2007**, *99*, 198101–198104.

(41) Kallblad, P.; Mancera, R. L.; Todorov, N. P. Assessment of multiple binding modes in ligand−protein docking. *J. Med. Chem.* **2004**, *47*, 3334–3337.

(42) Chan, H. S.; Shimizu, S.; Kaya, H. Cooperativity principles in protein folding. *Methods Enzymol.* **2004**, *380*, 350–379.

(43) Jackson, S. E.; Fersht, A. R. Folding of Chymotrypsin Inhibitor-2. 1. Evidence for a Two-State Transition. *Biochemistry* **1991**, *30*, 10428–10435.

(44) Baker, D. A surprising simplicity to protein folding. *Nature* **2000**, *405*, 39–42.

(45) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection−What can we learn from earlier mistakes. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.

(46) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(47) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.

(48) Ozkan, S. B.; Dill, K. A.; Bahar, I. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci.* **2002**, *11*, 1958–1970.

(49) Cho, S. S.; Levy, Y.; Wolynes, P. G. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. U S A.* **2006**, *103*, 586–591.

(50) Krivov, S. V.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U S A.* **2004**, *101*, 14766–14770.

(51) Shakhnovich, E. Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chem. Rev.* **2006**, *106*, 1559–1588.

(52) Wang, R. X.; Liu, L.; Lai, L. H.; Tang, Y. Q. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379–394.

CI900463U