JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

Statistical Potential for Modeling and Ranking of Protein-Ligand Interactions

Hao Fan, Dina Schneidman-Duhovny, John J. Irwin,
Guangqiang Dong, Brian Kenton Shoichet, and Andrej Sali

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/ci200377u • Publication Date (Web): 20 Oct 2011

Just Accepted

Journal of Chemical Information and Modeling is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Statistical Potential for Modeling and Ranking of Protein-Ligand Interactions

Hao Fan[1,2], Dina Schneidman-Duhovny[1], John J. Irwin[2], Guangqiang Dong[1], Brian K. Shoichet[2*], and Andrej Sali[1*]

[1] Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco

[2] Department of Pharmaceutical Chemistry and California Institute for Quantitative Biosciences, University of California ,San Francisco

[*]Corresponding authors:

[1] UCSF MC 2552, Byers Hall at Mission Bay, Suite 503B, University of California, San Francisco, 1700 4th Street, San Francisco, CA 94158-2330, USA

tel: +1 415 514 4227; fax: +1 415 514 4231

E-mail: sali@salilab.org

[2] UCSF MC 2550, Byers Hall at Mission Bay, Suite 508D, University of California at San Francisco, 1700 4th Street, San Francisco, CA 94158-2330, USA

tel: +1 415 514 4126; fax: +1 415 514 4260

E-mail: shoichet@cgl.ucsf.edu

Running title: a statistical potential for protein-ligand interactions

Keywords: statistical potential, reference state, binding pose, ligand enrichment

ABSTRACT

Applications in structural biology and medicinal chemistry require protein-ligand scoring functions for two distinct tasks: (i) ranking different poses of a small molecule in a protein binding site; and (ii) ranking different small molecules by their complementarity to a protein site. Using probability theory, we developed two atomic distance-dependent statistical scoring functions: PoseScore was optimized for recognizing native binding geometries of ligands from other poses and RankScore was optimized for distinguishing ligands from nonbinding molecules. Both scores are based on a set of 8,885 crystallographic structures of protein-ligand complexes, but differ in the values of three key parameters. Factors influencing the accuracy of scoring were investigated, including the maximal atomic distance and non-native ligand geometries used for scoring, as well as the use of protein models instead of crystallographic structures for training and testing the scoring function. For the test set of 19 targets, RankScore improved the ligand enrichment (logAUC) and early enrichment (EF_1) scores computed by DOCK 3.6 for 13 and 14 targets, respectively. In addition, RankScore performed better at rescoring than each of seven other scoring functions tested. Accepting both the crystal structure and decoy geometries with all-atom root-mean-square errors of up to 2 Å from the crystal structure as correct binding poses, PoseScore gave the best score to a correct binding pose among 100 decoys for 88% of all cases in a benchmark set containing 100 protein-ligand complexes. PoseScore accuracy is comparable to that of DrugScore^{CSD} and ITScore/SE, and superior to 12 other tested scoring functions. Therefore, RankScore can facilitate ligand discovery, by ranking complexes of the target with different small molecules; PoseScore can be used for protein-ligand complex structure prediction, by ranking different conformations of a given protein-ligand pair. The statistical potentials are available through the Integrative Modeling Platform (IMP) software package (http://salilab.org/imp/) and the LigScore web server (http://salilab.org/ligscore/).

INTRODUCTION

Molecular recognition between proteins and ligands plays an important role in many biological processes, such as membrane receptor signaling and enzyme catalysis. Predicting the structures of protein-ligand complexes and finding ligands by virtual screening of small molecule databases are two long-standing goals in molecular biophysics and medicinal chemistry.[1, 2] Solving both problems requires the development of an accurate and efficient scoring function to assess protein-ligand interactions.

Much effort has been devoted to developing scoring functions for modeling protein-ligand interactions.[3-12] These scoring functions can be divided into three categories[13]: potential or free energy functions based primarily on a molecular mechanics force field[14-26], knowledge-based statistical potentials based on distributions of intermolecular features in large databases of protein-ligand complex structures[27-40], and empirical-regression functions fitted to experimental binding constants of a training set of protein-ligand complexes.[41-50]

Energy functions based on molecular mechanics force field generally estimate the binding affinity by summing van der Waals, electrostatic, desolvation, and/or entropy terms. The weights for various terms are sometimes obtained by fitting the energy function to experimental binding constants for a training set of protein-ligand complexes. Because of the rugged energy landscape, minimization is often required prior to energy evaluation. The identification of the global minimum in the energy landscape generally requires extensive conformational and configurational sampling.

Statistical potentials are based on distributions of intermolecular structural features extracted from large databases, such as Protein Data Bank (PDB)[51] and Cambridge Structural Database (CSD).[52] Statistical potentials have been widely used because of their relative simplicity, accuracy, and computational efficiency.[53-98] During the last decade, several statistical potentials have been developed to describe protein-ligand interactions, such as PMF,[27] SMoG2001,[33] and DrugScore.[30, 35] Still, many aspects of statistical potentials for protein-ligand interactions have not yet been systematically explored.

Here, we are interested in the following questions. First, can a statistical potential be used for distinguishing between ligands and nonbinding molecules, in addition to recognizing native binding modes? Second, can the accuracy of a statistical potential

be improved by adding "negative" information, such as geometric decoys of the true ligands? Third, what is the accuracy of scoring complexes with modeled protein structures relative to that with crystallographic structures? Finally, what are the differences between the reference states for protein-ligand and protein-protein statistical potentials?

We describe two distance-dependent atomic statistical potentials derived from PDB - one for predicting the binding pose of a known ligand (PoseScore), and the other one for identifying ligands through virtual screening (RankScore). We proceed in three steps. First, distance distributions for the protein-ligand atom-type pairs were calculated from a sample of native complex structures (structures in the training and testing sets are excluded). Second, the distance-dependent atomic potential was derived from these distance distributions, and trained to find the optimal set of parameters for binding pose prediction (PoseScore) and ligand enrichment (RankScore), respectively. Third, PoseScore and RankScore were evaluated with the aid of two widely used docking benchmarks.[11, 99] The performance of PoseScore and RankScore was compared to that of a number of other scoring functions.

We begin by describing the theory used to derive PoseScore and RankScore, criteria to evaluate the accuracy of each statistical potential, as well as the procedures and data sets used to derive, train, and test the statistical potentials (Methods). We then describe the accuracy of PoseScore and RankScore for docking against crystal structures of proteins in comparison to 14 and 7 other scoring functions, respectively (Results). We proceed by describing (i) the effect of including both native and non-native conformations of small molecules in the derivation of the statistical potentials, (ii) the accuracy of scoring against modeled protein structures, as well as (iii) the distribution of atomic protein-ligand distances. Finally, we discuss the implications of the results, relative successes and failures, and answer the questions raised above (Discussion and Conclusions).

METHODS

Theory

From joint pdf to distance pdf

An atomic distance-dependent, statistical potential for protein-ligand complexes can be defined as the negative logarithm of the joint probability density function (pdf) of the atomic Cartesian coordinates, as suggested in a previous study[100]:

$$S_{complex} = -\ln\left(p\left(\vec{x}_1,\vec{x}_2,...,\vec{x}_m,\vec{y}_1,\vec{y}_2,...,\vec{y}_n\right)\right) \quad (1)$$

where m is the number of atoms in the protein and $\vec{x}_i$ are the Cartesian coordinates of protein atom i; n is the number of atoms in the ligand and $\vec{y}_j$ are the Cartesian coordinates of ligand atom j. $p\left(\vec{x}_1,\vec{x}_2,...,\vec{x}_m,\vec{y}_1,\vec{y}_2,...,\vec{y}_n\right)$ is the probability that the structure defined by the Cartesian coordinates $\left(\vec{x}_1,\vec{x}_2,...,\vec{x}_m,\vec{y}_1,\vec{y}_2,...,\vec{y}_n\right)$ is the native structure. The joint pdf can be approximately modelled by a normalized product of pair pdfs:

$$p\left(\vec{x}_1,\vec{x}_2,...,\vec{x}_m,\vec{y}_1,\vec{y}_2,...,\vec{y}_n\right) \simeq \frac{\prod_{i,j\in m,i\neq j} p\left(\vec{x}_i,\vec{x}_j\right) \prod_{i,j\in n,i\neq j} p\left(\vec{y}_i,\vec{y}_j\right) \prod_{i\in m,j\in n} p\left(\vec{x}_i,\vec{y}_j\right)}{\left(\prod_{i\in m} p\left(\vec{x}_i\right)\right)^{m+n-2} \left(\prod_{j\in n} p\left(\vec{y}_j\right)\right)^{m+n-2}} \quad (2)$$

where $p\left(\vec{x}_i\right)$ and $p\left(\vec{y}_j\right)$ are single-body pdfs of protein atom i and ligand atom j; a pair pdf $p\left(\vec{x}_i,\vec{y}_j\right)$ depends only on the types and positions of the two atoms.

Next, we relate the pair pdf $p\left(\vec{x}_i,\vec{y}_j\right)$ for an atom pair between a protein atom i of atom type P and a ligand atom j of atom type L to the distance pdf:

$$p\left(\vec{x}_i,\vec{y}_j\right) = p\left(\vec{x}_i\right)p\left(\vec{y}_j\right)\frac{p_{nat}\left(r_{i,j}^{P,L}\right)}{p_{ref}\left(r_{i,j}^{P,L}\right)} \quad (3)$$

where $p_{nat}\left(r_{i,j}^{P,L}\right)$ is the distance distribution for the atom-type pair $\left(P,L\right)$, derived directly from a sample of native complex structures. We define a “reference state” as uncorrelated positions of atoms of types P and L in a finite sphere of an appropriate size and centered at L. Combining Equations 1-3, we obtain:

$$S_{complex} = -\ln\left(\frac{\prod_{i,j\in m,i\neq j} p\left(\vec{x}_i,\vec{x}_j\right)}{\left(\prod_{i\in m} p\left(\vec{x}_i\right)\right)^{m-2}}\right) - \ln\left(\frac{\prod_{i,j\in n,i\neq j} p\left(\vec{y}_i,\vec{y}_j\right)}{\left(\prod_{j\in n} p\left(\vec{y}_j\right)\right)^{n-2}}\right) - \ln\left(\prod_{i\in m,j\in n}\frac{p_{nat}\left(r_{i,j}^{P,L}\right)}{p_{ref}\left(r_{i,j}^{P,L}\right)}\right) \quad (4)$$

In this formulation, the statistical potential for the protein-ligand complex is a sum of three terms, corresponding to the protein, ligand, and protein-ligand atomic distances. Here, we focus only on the protein-ligand term:

$$S_{inter} = -\sum_{i \in m, j \in n} \ln\left(\frac{p_{nat}\left(r_{i,j}^{P,L}\right)}{p_{ref}\left(r_{i,j}^{P,L}\right)}\right) \quad (5)$$

Distance distribution from native structures

Given a sample of native complex structures, the distance distribution for a pair of atom types (P,L) can be estimated as:

$$p_{nat}\left(r_{P,L}\right) \simeq \frac{N\left(r_{P,L}\right)}{\sum_{r_{\min} \le r \le r_{\max}} N\left(r_{P,L}\right)} \quad (6)$$

where $N\left(r_{P,L}\right)$ is the number of observations of the pair (P,L) in a particular distance bin $\left(r, r+\Delta r\right]$. $r_{\min}$ and $r_{\max}$ are the minimal and maximal distance bounds of the distribution, respectively. To minimize the impact of the finite sample size on the accuracy of $p_{nat}\left(r_{P,L}\right)$, we obtain a more accurate estimate of the distance distribution for a pair of atom types (P,L) by linearly combining $p_{nat}\left(r_{P,L}\right)$ with the reference state $p_{ref}\left(r_{P,L}\right)$[101, 102]:

$$p'_{nat}\left(r_{P,L}\right) = p_{nat}\left(r_{P,L}\right) \bullet \left(1 - w_{ref}\right) + p_{ref}\left(r_{P,L}\right) \bullet w_{ref} \qquad 0 \le w_{ref} < 1 \quad (7)$$

where w_{ref} is an adjustable parameter to be optimized by training. For a pair of atom types (P,L), the reference state is the distance distribution derived from all conformations of a protein-ligand complex.

Reference state

The calculation of $p_{ref}\left(r_{P,L}\right)$ is not straightforward, because it is not possible to enumerate all conformations of the protein-ligand complex. We approximated the reference state by deleting atom type labels in native complex structures:

$$p_{ref}\left(r_{P,L}\right) = \frac{\sum_{P,L} N\left(r_{P,L}\right)}{\sum_{r_{\min} \le r \le r_{\max}} \sum_{P,L} N\left(r_{P,L}\right)} \quad (8)$$

where $\sum_{P,L} N\left(r_{P,L}\right)$ is the number of all pairs of atom types in a particular distance bin $\left(r, r+\Delta r\right]$. Because the reference state in Equation 8 is based on the native

complex structures, it may still be different from the distribution for the native and all non-native structures. Therefore, we also adjust the reference state by linearly combining it with the uniform distribution:

$$p'_{ref}\left(r_{P,L}\right) = p_{ref}\left(r_{P,L}\right) \bullet \left(1 - w_{uni}\right) + \frac{1}{N_{bin}} \bullet w_{uni} \qquad 0 \le w_{uni} < 1 \qquad (9)$$

where w_{uni} is an adjustable parameter to be optimized by training, and N_{bin} is the total number of distance bins between $r_{\min}$ and $r_{\max}$. Considering Equation 5, 7, and 9, the statistical potential for protein-ligand interactions is:

$$S_{inter} = -\sum_{i \in m, j \in n} \frac{p_{nat}\left(r_{i,j}^{P,L}\right)'}{p_{ref}\left(r_{i,j}^{P,L}\right)'} \qquad (10)$$

Parameters of the statistical potential

The bin size Δr is set to 0.1 Å. The minimal distance boundary $r_{\min}$ is set to 2 Å. The protein-ligand score for distances of less than 2 Å is calculated by a linear interpolation between $S_{\max} = 20.0$ and $S\left(r_{P,L}\right)$ in the distance bin $\left(2.0, 2.1\right]$. $r_{\max}$ is an adjustable parameter to be optimized by training.

Atom types for proteins and ligands

The protein atom types were adapted from the DOPE scoring function,[100] resulting in 158 residue-dependent atom types for non-hydrogen atoms. 26 atom types were used to represent non-hydrogen atoms in small molecules, derived from the SYBYL software (Tripos, Inc.) (Table 1).

Assessment of the scoring accuracy

Ligand pose

The geometric accuracy of a ligand pose was measured by its all-atom root-mean-square-deviation (RMSD) from the crystal structure. The correct binding pose of a ligand was considered successfully recognized if the all-atom RMSD value of the best-scored pose is less than 2 Å.[11, 30, 35, 39, 40]

Ligand rank

The accuracy of ranking ligands in molecular docking screens was evaluated by the enrichment for the known ligands among the entire docking library, as quantified by the area under the curve (logAUC) of the enrichment plot.[103] A random selection of

compounds from the mixture of actual ligands and decoys yields a logAUC of 14.5; a mediocre selection that picks twice as many ligands as a random selection has logAUC of 24.5; a highly accurate enrichment that produces ten times as many ligands than a random selection has logAUC of 47.7. For each compound, the DOCK-produced complex model[104, 105] was re-ranked by the tested scoring function. The ligand enrichment was quantified using the area under the enrichment curve with the x-axis on the logarithmic scale (logAUC),[103, 106] For each protein target, the ligand enrichment for the DOCK-produced ranking was compared to that generated by re-ranking the DOCK list with the statistical potential. A difference larger than 3 logAUC units between the two enrichment values was defined to be significant; otherwise the enrichment values were considered to be comparable. The value for this significance cutoff was chosen subjectively, based on a previous study.[103]

PoseScore and RankScore

Native complex structures

8,885 X-ray structures of protein-ligand complexes used for the calculation of PoseScore and RankScore were selected from the dataset used in previous automated docking screens.[51, 106] Five conditions were applied: 1) only crystallographically determined complexes with a resolution better than or equal to 2.5 Å were used; 2) the protein receptor had to contain more than 50 non-hydrogen atoms; 3) the ligand had to contain at least one carbon or nitrogen atom; 4) at least one pair of protein-ligand non-hydrogen atoms had to have a distance between 2.0 and 4.0 Å; 5) no overlap between the complex structures (PDB entries)in the training and testing sets was allowed.

Training and testing of PoseScore

The training set used for PoseScore was constructed from the Astex diverse set that contains 85 crystallographically determined protein-ligand complexes.[107] DOCK was employed to generate ligand poses for all complexes. In 70 out of the 85 cases, 100 poses were generated for the ligand, containing at least one pose with an all atom RMSD error of less than 2.0 Å (near-native solution). The training set (Astex_DOCK, Table S1) was formed by these 70 complexes. For each complex, Astex_DOCK included the crystal structure of the protein, the crystal binding pose of the ligand, and the 100 docking poses of the ligand (geometric decoys).

The Astex_DOCK training set was used to find optimal values for the three adjustable parameters in PoseScore, including the maximal distance boundary $r_{\max}$, and the two smoothing parameters w_{ref} and w_{uni}. $r_{\max}$ was optimized over 6 discrete values, including 4, 6, 8, 10, 12, and 14 Å. The search for optimal values of w_{ref} and w_{uni} ranged from 0.0 to 0.9 with an increment of 0.1. First, we fixed w_{uni} at 0.0 and scanned $r_{\max}$ and w_{ref}. The statistical potential was most accurate when $r_{\max}$ was 6 Å and w_{ref} was 0.4 (Figure 1a). The correct binding pose of the ligand, either the crystal structure or a docking pose with an all-atom RMSD error ≤ 2.0 Å, was detected for 64 (91%) targets. When the crystal structures of ligands were excluded from the training set, we were able to identify the correct binding pose for 53 (76%) targets. Second, the value of $r_{\max}$ was set to the optimized value of 6 Å, while the values of w_{ref} and w_{uni} were explored. The statistical potential was most accurate when w_{ref} and w_{uni} were both 0.3 (PoseScore). The correct binding pose was detected for 67 (96%) and 56 (80%) targets when the crystal structures of ligands were included and omitted, respectively.

The performance of the trained PoseScore was tested using the previously constructed data set of 100 protein-ligand complexes[11] (Wang_AutoDock). For each complex, 100 docked conformations were generated using AutoDock.[26] In 91 out of 100 cases, there is at least one near-native solution. The accuracy of PoseScore on Wang_AutoDock was compared to accuracies of 14 other scoring functions that were previously tested using the same data set [11, 35, 38-40],(Table 2).

Training and testing of RankScore

38 crystallographically determined protein-ligand complexes were taken from "A Directory of Useful Decoys" (DUD) benchmark set[99] and divided into two equally sized subsets. 19 complexes (DUD-1) were used in the training of RankScore (Table S2a), while the rest (DUD-2) were used to test RankScore (Table S2b). All compounds in DUD (annotated ligands and screening decoys) were screened against the 38 holo X-ray structures.[103] The generated docking poses in DUD-1 and DUD-2 were used to train and test RankScore, respectively.

The DUD-1 training set was used to find optimal values for $r_{\max}$, w_{ref}, and w_{uni}. The training process used for PoseScore was employed, except that 3 values were explored for $r_{\max}$, including 6, 10, and 14 Å. The statistical potential showed the highest accuracy in the rescoring of the DUD-1 training set when $r_{\max}$, w_{ref}, and w_{uni} were 6 Å, 0.4, and 0.0 (RankScore), respectively (Figure 1b). For 14 targets, enrichment against the entire DUD library (logAUC) was improved by rescoring, compared to the original enrichment by DOCK. For 1 target, the rescoring enrichment was comparable to that by DOCK. For the remaining 4 targets, lower enrichment was obtained after the rescoring procedure. Rescoring improved the average logAUC by 6.9.

The accuracy of the trained RankScore was tested using the DUD-2 set. We also rescored DUD-2 with 7 other scoring functions, including ITScore[38, 39], DrugScore^{PDB}[30], FlexX[108], PMF[27], PLP[109], ScreenScore[32], and the all-atom energy function in PLOP[30, 32, 38, 39, 110]. We did not test ITScore/SE and DrugScore^{CSD} because they are not publicly available. However, ITScore and DrugScore^{PDB} should perform similarly as ITScore/SE and DrugScore^{CSD}, in terms of ligand enrichment (personal communication with XQ. Zou and G. Klebe, respectively). FlexX, PMF, PLP, and ScreenScore were calculated by a re-implementation of the original scoring functions (kindly provided by M. Stahl).[32]

Statistical potentials computed from docking-produced ligand poses

For each of the 8,885 native complex structures, the crystal ligands were docked to the binding site using DOCK 3.6.[104, 105] Overall, ligand docking poses were generated for 7,215 targets (the remaining 1670 targets did not produce any ligand docking pose during the fully automated docking). Out of these 7,215 cases, 4,059 produced at least one near-native solution, while 6,895 produced poses with all-atom RMSD errors of more than 2 Å (random solutions).

The influence of incorporating geometric decoys into deriving a protein-ligand statistical potential[111-113] was investigated (Table S3). The near-native complex structures and the random complex structures were considered during the evaluation of Eqn. 5, in three different ways: (i) keeping p_{ref} unchanged, the near-native structures were combined with the native structures and used to calculate p_{nat}; (ii)

keeping p_{nat} unchanged, the random structures were used to represent the reference state and used to calculated p_{ref}; and (iii) using p_{nat} from (i) and p_{ref} from (ii). Each resulting statistical potential was subjected to the same training process that was used for PoseScore and RankScore.

Statistical potentials applied to modeled structures

PoseScore and RankScore were derived, trained, and tested for target protein structures determined by crystallography. In realistic applications, comparative models are often used to represent the receptor structure in both docking and virtual screening.[114-117] Thus, we also investigated the accuracy of statistical potentials on docking and screening against models of target proteins (Table S3).

Training and testing sets

The 170 protein structures from Astex_DOCK and Wang_AutoDock were structurally perturbed using MODELLER[118] in the absence of the crystal ligand. For each protein, binding site residues that have atoms within 10 Å from the bound ligand were simulated by 100 steps of molecular dynamics with simulated annealing (MD-SA) in which the temperature was reduced from 400 K to 100 K, and 100 steps of conjugate gradient minimization (CG). All-atom RMSD errors of the resulting models ranged from 0.5 to 1.5 Å. These perturbed structures have been used in the past as proxies for comparative models.[119] Next, ligand poses were generated in the modeled binding site for each of the 170 structures. Of the 70 structures in Astex_DOCK and the 100 structures in Wang_AutoDock, 60 (Astex_DOCK^{model}) and 85 (Wang_AutoDock^{model}) produced 100 ligand docking poses including at least one near-native solution, respectively. Astex_DOCK^{model} and Wang_AutoDock^{model} were used as the training set and the testing set, respectively. The optimal values of the three adjustable parameters $r_{\max}$, w_{ref}, and w_{uni} were determined as 6 Å, 0.7, and 0.1, with the aid of the Astex_DOCK^{model} training set.

The 38 protein structures from DUD-1 and DUD-2 were perturbed using the same approach as described above (DUD-1^{model} and DUD-2^{model}, respectively). All-atom RMSD errors of 38 resulting models also ranged from 0.5 to 1.5 Å. All compounds in the DUD library were docked against each of the 38 models. DUD-1^{model} and DUD-2^{model} were used as the training set and testing set, respectively. The optimal values

of $r_{\max}$, w_{ref}, and w_{uni} were determined as 6 Å, 0.5, and 0, respectively, with the aid of the DUD-1[model] training set.

Reference state

To investigate the reference state for a protein-ligand statistical potential, an additional protein-ligand statistical potential was derived from the same sample of complex structures that were used for PoseScore, employing the formula described for the DFIRE potential[83]:

$$\bar{u}(i,j,r) = -\eta RT \ln \frac{N_{obs}(i,j,r)}{\left(\frac{r}{r_{cut}}\right)^{\alpha} \frac{\Delta r}{\Delta r_{cut}} N_{obs}(i,j,r_{cut})} \qquad r \le r_{\max} \qquad (11)$$

where $N_{obs}(i,j,r)$ is the number of (i,j) pairs within the distance shell $(r, r+\Delta r]$ observed in the X-ray structures used to generate PoseScore. Two approximations similar to those in DFIRE were made. First, the number of pairs of ideal gas points in a finite protein-ligand sphere is proportional to r^{α} in Eqn. 11. Second, the potential has a finite interaction range r_{cut} fixed at 14 Å. That is, for $r > r_{cut}$, $\bar{u}(i,j,r) \equiv 0$. Differently from DFIRE, we set the bin width Δr to 0.1 Å. We then generated a statistical potential with distinct values for the exponent α including 1, 2, 3, 4, 5, and, 6. To exclude the influence of distance boundary, for each α value, 5 different values (6, 8, 10, 12, 14 Å) for the maximal boundary $r_{\max}$, beyond which atom pairs were not considered during scoring, were tested on the Astex_DOCK set.

Results

PoseScore and RankScore

Ligand pose

The trained PoseScore in which $r_{\max}$, w_{ref}, and w_{uni} were set to 6 Å, 0.3, and 0.3, respectively, was assessed by the Wang_AutoDock testing set of 100 protein-ligand complexes. The correct binding pose was detected for 88 (88%) targets, of which 70 were crystal structures (Table 2). Furthermore, a correct binding pose was ranked the best, top 5, and top 10 by PoseScore for 88 (88%), 97(97%), and 99 (99%) targets in the testing set, respectively (Table 3). To mimic realistic applications, only

the geometric decoys of the ligands were scored, resulting in the correct binding pose identification for 63 (69%) targets.

Among all the scoring functions under test, ITScore/SE, PoseScore, and DrugScore^{CSD} performed better than other scoring functions, showing the success rate of 91%, 88%, and 87% in the identification of the correct binding pose, respectively. When the crystal structures of ligands were excluded from the test, PLP, PoseScore, and F-Score were the three best performing scoring functions, with the success rate of 70%, 69%, and 68%, respectively.

Ligand pose example

PoseScore detected the correct binding pose for 88 (88%) targets in the testing set: four examples are shown in Figure 2. The 12 failures were investigated in detail. Out of the 12, 9 and 11 cases had correct binding poses ranked in top 5 and top 10, respectively (Table 3). The 12 targets can be divided into four classes: (i) water molecules play an important role in ligand binding, including five targets 1cla, 3cla, 4cla, 1rgl, and 3tmn; (ii) the ligand and the receptor forms a transition-state complex, including three targets 1tlp, 1zzz, and 2sns; (iii) the ligand is located in the neighborhood of a cofactor or another ligand, including two targets 1dr1 and 1tha; and (iv) no particular feature in the binding site is found to be responsible for the failure, including two targets 1tni and 1tnj. Next, we discuss one example from each class.

1cla is the crystal structure of chloramphenicol acetyltransferase, determined at the resolution of 2.34 Å.[120, 121] The binding pocket accommodates the substrate chloramphenicol and several ordered water molecules (Figure 3). The residues lining the pocket are predominantly hydrophobic. The substrate adopts an eclipsed conformation and forms direct hydrogen bonds only with the water molecules. These water molecules were not included during the generation of docking solutions of the substrate. As a consequence, the crystal structure of chloramphenicol was only ranked 9. The best ranking pose is in a staggered conformation and has an all-atom RMSD error of 10.3 Å.

1zzz is the crystal structure of trypsin with a peptidyl aldehyde inhibitor CVS1694, determined at the resolution of 1.90 Å.[122] In the crystal complex, the guanidinopiperidyl group of CVS1694 makes water-bridged hydrogen bonds with Asp189 and Gly219. The carbonyl oxygen of the aldehyde group is hydrogen

bonding with Gly193N and Ser195N, while the carbonyl carbon forms a tetrahedral intermediate with Ser195OG. A consequence of the latter interaction is the covalent-bonding distance for C-Ser195OG (1.8 Å). The glycine residue and the six-member lactam ring of the inhibitor make hydrogen bonds with Ser214-Gly216, holding this part of the inhibitor close to trypsin. This crystal structure of CVS1694 was ranked 2. The best ranking pose has an all-atom RMSD error of 3.10 Å, deviating from the crystal structure in its aldehyde and lactam groups.

1dr1 is the crystal structure of dihydrofolate reductase (DHFR), solved as a complex with NADP+ and biopterin at the resolution of 2.20 Å.[123] The closely packed cofactor and the substrate interact at an angle of 45º. Water molecules also hydrogen bond to both hydroxyls in the dihydroxypropyl and the pteridine group of biopterin. The cofactor and these water molecules were not included during the generation of docking solutions of the substrate. A near-native solution of biopterin (RMSD 1.02 Å) was ranked 2, while the crystal structure was ranked 3. The best ranking pose has an all-atom RMSD error of 4.17 Å. The pteridine ring of the best ranking pose connects the substrate and the cofactor binding sites, and is perpendicular to the plane of the pteridine ring in the crystal structure. The 2-amino group of the pteridine ring still hydrogen bonds to Glu30, as it does in the crystal structure.

1tni and 1tnj are crystal structures of bovine trypsin, solved with phenylbutylamine (PBA) and phenylethylamine (PEA) at the resolution of 1.90 Å and 1.80 Å, respectively.[124]. The binding affinity of PEA to trypsin (Ki 11.0 mM) is higher than that of PBA (Ki 20.0 mM). In the crystal complexes, the amine group in both inhibitors forms a salt-bridge with Asp189 and a hydrogen bond to Gly219. The difference in the binding affinity between the two compounds could be due to the position of the benzene ring, which is more solvent exposed in PBA than in PEA. In the case of 1tni, the crystal structure was ranked 22. The best ranking pose has an all-atom RMSD error of 5.72 Å. The benzene ring of the best ranking pose was close to the equivalent group of PEA in 1tnj structure. However, the amine group in the best ranking pose pointed in the direction opposite to the crystal structure and failed to form a hydrogen bond to Gly219. Interestingly, in the case of 1tnj, the crystal structure was ranked 2 while the all-atom RMSD errors of the 1st and 3rd ranking poses are 2.29 and 1.51 Å, respectively (poses not shown). The best ranking pose was in the same orientation as in the crystal structure and formed a hydrogen bond to Gly219.

Ligand rank

The trained RankScore was assessed by the DUD-2 testing set of 19 targets. For 13 (68%) targets, the rescoring enhanced the enrichment against the entire DUD library (logAUC), with respect to the original enrichment by DOCK (Table 4). Rescoring improved the average logAUC by 6.8. Another enrichment indicator - the enrichment factor at 1% of the ranked docking library (EF_1)[99] - was also measured because the early enrichment is particularly important in realistic applications. RankScore significantly improved EF_1 for 14 (74%) targets (of which 12 had an increased logAUC, Table 5). In particular, the rank of the best-scored ligand was enhanced by RankScore for 16 (84%) targets.

RankScore was more accurate in the rescoring than 7 other tested scoring functions (Table 4). Of these 7 scoring functions, only rescoring by ITScore and FlexX improved the enrichment relative to DOCK (the average logAUC increased by 3.6 and 0.9, respectively). For the 19 targets, increased enrichment was observed in 9, 8, 6, 9, 7, 7, and 8 cases with ITScore, $DrugScore^{PDB}$, FlexX, PMF, PLP, ScreenScore, and PLOP, respectively.

Ligand rank example

Rescoring by RankScore worsened both logAUC and EF_1 for 4 targets, including thrombin, cyclooxygenase-2 (COX-2), AmpC β-lactamase (AmpC) and hydroxymethylglutaryl-CoA reductase (HMGR). In 3 Of the 4 cases, including thrombin, AmpC and HMGR, the best rank of ligands produced by RankScore is better than that by DOCK. For COX2, the crystal structure at 3 Å resolution (PDB code: 1cx2) was determined with SC-558, a selective COX-2 inhibitor. However, the sulfonamide group in SC-558 is only 1.3 Å away from the guanidinium ion of Arg513. Many docking poses of ligands were close to that of the crystal ligand but deviated from the position of the sulfonamide. For the other three targets, we discuss AmpC in detail.

In the DUD benchmark, the A chain in the dimeric structure of AmpC (1xgj) was used.[125, 126] Ligand enrichments in logAUC are 47.4 and 10.3 units by DOCK and the rescoring method proposed here, respectively. However, the H10 helix and the α-helix close to the binding site are unstructured in the A chain, potentially explaining poor accuracy of RankScore in this case. We tested this hypothesis by using the B chain structure, in which these two helices are well defined. As a result, both the

enrichment by DOCK and that by rescoring were improved, reflected in logAUC of 53.8 and 19.3, respectively. The B chain structure was solved as a complex with a thiophene-carboxylate derivative (HTC). The docking pose of HTC was ranked 628 and 14 by DOCK and by RankScore, respectively. Another ligand CTC differs from HTC in the 4-carboxylate benzene ring on which HTC has a 2-hydroxyl group but CTC has a 3-chloride group (Figure 4a). The binding affinity of HTC to AmpC (Ki 1.0 μM) is higher than that of CTC (Ki 1.9 μM). As shown in Figure 4b, the thiophene-carboxylate in HTC hydrogen bonded to Thr316, Asn346 and Arg349 in the primary carboxylate binding site. The sulfonamide oxygen hydrogen bonded to Ser64, Tyr150, and Ala318 in the oxyanion hole and the hydroxyl binding site. The hydroxyl group on the benzene-carboxylate ring hydrogen bonded to other active site residues (Lys67 and Asn152). However, another AmpC ligand CTC with similar chemical structure and docking pose as HTC, was ranked 786 by DOCK but only 52,978 by RankScore. In CTC, the 2-hydroxyl on the carboxylate benzene ring of HTC was replaced by the 3-chloride, which forced this part of CTC to move away from the neighboring residue Tyr221, resulting in weaker thiophene-carboxylate and sulfonamide hydrogen bonds.

Statistical potentials computed from docking-produced ligand poses

The near-native and random solutions were considered in the derivation of protein-ligand statistical potential. The resulting statistical potentials were subjected to the same training process as PoseScore and RankScore.

For ligand enrichment, the trained potentials still improved the enrichment with respect to the original result by DOCK, but performed worse than RankScore (the average logAUC improved by 6.8): (i) keeping p_{ref} unchanged but incorporating the near-native structures in p_{nat}, the trained potential with the three adjustable parameters $r_{\max}$, w_{ref}, and w_{uni} optimized to be 6 Å, 0.2, and 0 improved the average logAUC by 6.0; (ii) keeping p_{nat} unchanged but computing p_{ref} from random structures, the trained potential with $r_{\max}$, w_{ref}, and w_{uni} optimized to be 6 Å, 0, and 0 improved the average logAUC by 2.2; and (iii) using p_{nat} from (i) and p_{ref} from (ii), the trained potential with $r_{\max}$, w_{ref}, and w_{uni} optimized to be 6 Å, 0, and 0 improved the average logAUC by 2.8.

In contrast, for ligand pose prediction, the trained potential (PoseScore^{dock}) that was computed with the same adjustable parameters as PoseScore but incorporated random docking solutions in p_{ref}, showed better accuracy than PoseScore. Of the 100 proteins in the testing set, PoseScore^{dock} detected the correct binding pose for 90 (90%) targets, among which 70 (70%) are crystal structures (88% and 70% by PoseScore, respectively). When the crystal structures were omitted, the near-native solution was identified for 67 (74%) targets (69% by PoseScore). The results of PoseScore and PoseScore^{dock} are probably robust because both potentials were trained and assessed based on decoys constructed by different docking programs.

Statistical potentials applied to modeled structures

The trained potential (PoseScore^{model}) detected the near-native solution for 66 (78%) targets in the Wang_AutoDock^{model} testing set (69% by PoseScore for Wang_AutoDock). In comparison to Wang_AutoDock, more near-native solutions were included for 60 targets, 48 of which (80%) had the correct binding pose detected. Equal and less near-native solutions were included for 2 and 23 targets, respectively, among which the correct binding pose was detected for 18 (72%) targets. Clearly, the improvement in the accuracy is partially due to a better sampling of near-native solutions. Meanwhile, the perturbation applied to the original X-ray structures also contributed to the higher success rate. As an example, in both Wang_AutoDock and Wang_AutoDock^{model} testing sets, the target 1bra contained 15 near-native solutions among 100 geometric decoys with the minimal RMSD error of 0.6 and 1.2 Å, respectively. In the crystal structure, the acetamidine group in the ligand benzamidine hydrogen bonded with Ser190 backbone, Gly219 backbone, and Asp226 sidechain. A docking pose of 1.8 Å RMSD was ranked the best when scoring against the crystal structure of 1bra, because many near-native poses were in different orientations favoring hydrogen bonds with Asp226 and/or Ser190. In the modeled structure of 1bra, the Asp226 sidechain became closer to the binding site, associated with the Ser190 backbone moving away. The ligand could form hydrogen bonding network with this binding site configuration, in an orientation close to that in the crystal structure. The docking pose with the minimal RMSD of 1.2 Å was selected.

An improved enrichment (logAUC) was observed for 10 targets in the DUD-2^{model} testing set. For 3 and 6 targets, the rescoring enrichment was comparable to and lower than that by DOCK, respectively. Rescoring by RankScore^{model} improved the

average logAUC by 1.6. Interestingly, for three targets, dihydrofolate reductase (DHFR), glycinamide ribonucleotide transformylase (GART), and thrombin, the logAUC by DOCK using the modeled structure was significantly higher by 40.1, 21.5, and 24.1 than that using the ligand-bound crystal structure, respectively.

Reference state

The protein-ligand statistical potential with the DFIRE reference state showed the best accuracy when r_{max} was 6 Å and α was 3 (Figure 5). The correct binding pose of the ligand, either the crystal structure or a docking pose with an all-atom RMSD error ≤ 2.0 Å, was detected for 54 (77%) targets. When the crystal structures of ligands were excluded from the training set, the correct binding pose was detected for 49 (70%) targets. For the 5 r_{max} values tested, the statistical potential always showed the maximal accuracy with an α value of 3 and/or 4.

Discussion

Two key results emerge from this study. First, two different statistical potentials can be derived by statistical analysis of a database of known protein-ligand complex structures: PoseScore for ligand pose prediction, and RankScore for ligand discovery in virtual screeing. Second, PoseScore is as accurate as DrugScore^{CSD} and ITScore/SE in detecting the native structure of a protein-ligand complex, and superior to 12 other scoring functions tested; RankScore is more accurate than 7 other scoring functions in discriminating between true ligands and nonbinders.

We address three points here. First, we compare the distance distributions of atom pairs in protein-ligand complexes to those in proteins, and discuss the dataset used to derive the reference state. Second, we discuss the optimal parameters used in PoseScore and RankScore, including the maximal distance boundary r_{max} and the atom types. Finally, we suggest possible improvements of PoseScore and RankScore.

Reference state

Proteins are finite systems. For a folded protein in the ligand-free state, the reference distribution should not increase in r^2 as in an infinite system, but in r^{α}, where α is

smaller than 2. The optimal value of α was found empirically to be approximately 1.6 for protein structures.[83] In protein-ligand complexes, however, each ligand atom is partly surrounded by other ligand atoms. As a result, the protein-ligand distance distributions are expected to be different from those for protein structures alone. In Figure 6, the protein is approximated as the outer sphere (solid line) and the ligand is completely embedded inside the protein as the inner sphere with a radius r. All protein-ligand pairwise interactions within the maximal distance R_{cutoff} of the ligand atom only occur inside the protein sphere. Thus, for the ligand atom positioned at distance d from the ligand center, the number of protein-ligand atom pairs within distance R ($R \le R_{cutoff}$) is proportional to the partial volume V of the sphere with a radius R centered on the ligand atom that is not taken by the ligand:

$$V = \frac{\pi\left(3R^4 + 8dR^3 + 6d^2R^2 - 6r^2R^2 + 6d^2r^2 - 8dr^3 + 3r^4 - d^4\right)}{12d} \qquad R \le d + r$$

$$V = \frac{4\pi}{3}\left(R^3 - r^3\right) \qquad R \ge d + r \quad (12)$$

The expected number of protein-ligand atom pairs in the distance bin $\left(R, R + \Delta R\right]$ is proportional to the derivative of V with respect to R. For distance R not larger than $d + r$, the distribution of atom pairs increases faster than R^3. Clearly, there are additional geometrical arrangements, depending on the size of the protein, the size of the ligand, and the location of the ligand relative to the protein. Furthermore, the shapes of both protein and ligand are generally not ideal spheres. Nevertheless, this simplified example suggests that the number of protein-ligand atom pairs could increase much faster with distance R than the number of protein atom pairs does. The reference distributions observed in the X-ray and docking-generated structures of protein-ligand complexes (Figure 7) are better approximated by the distribution in eqn. 11 with an α value of 3 (not 2), thus supporting our hypothesis. Our approach to defining a reference state might be applicable for other multi-component systems, such as protein-peptide, protein-nucleotide, and even protein-protein interfaces.

In general, the goal of scoring functions for structure prediction is to distinguish the native from non-native states. The reference state used in a statistical potential should maximize this discrimination. Therefore, the choice of the optimal reference state depends on the native and non-native states. In particular, we hypothesize that a protein-ligand statistical potential for distinguishing ligand poses with an RMSD error of less than 2 Å (correct binding pose) from other poses (random solutions)

should depend on a reference state that is maximally different from the native state and maximally similar to random solutions. For ligand pose prediction, a reference state computed from random solutions should improve the accuracy of the statistical potential, in comparison to the reference state computed from a sample of native complex structures with the same approach. PoseScore^{dock} is more accurate than PoseScore, thus supporting our hypothesis. Similarly, for ligand discovery by virtual screening, a reference state including information about non-binders should improve a statistical potential relative to the reference state from true ligands alone. Therefore, it was not surprising that RankScore^{dock} using the reference state derived from docking poses of true ligands did not show better accuracy than RankScore.

Parameters of the statistical potentials

Different values had been used for the maximal distance boundary in previously developed protein-ligand statistical potentials. A distance bound of 12 Å was used in the potential of mean force (PMF) compiled from 697 complex structures in PDB, in conjunction with a correction term regarding the volume occupied by the ligand to incorporate solvent effects in the pairwise potential.[27] In comparison to PMF, a shorter distance cutoff of 6 Å was used in several recently developed protein-ligand statistical potentials, including DrugScore^{PDB} [30], DrugScore^{CSD} [35], ITScore[38] and ITScore/SE[40] that were based on distinct structure samples and solvation models. In this study, the distance cutoff was subjected to the optimization with the aid of a training set, for both ligand pose prediction and virtual screening. In both cases, the optimal value was found to be 6 Å. This distance cutoff is much smaller than those used in protein statistical potentials (e.g., 20 Å for RAPDF[70], 14.5 Å for DFIRE[83], and 15 Å for DOPE[100]). This result may indicate that non-covalent protein-ligand interactions are dominated by the specific interactions formed between a ligand and the first shell of the protein residues outlining the binding site pocket; examples of such interactions include hydrogen bonds and salt bridges.[127]

An atom type classification is needed for generating atom pair distributions. 16 protein atom types and 34 ligand atom types were employed in PMF. 17 and 18 atom types for both protein and ligand were employed in DrugScore^{PDB} and DrugScore^{CSD}, respectively. 26 and 27 atom types for both protein and ligand were employed in ITScore and ITScore/SE, respectively. These atom types reflect both the type of an atom as well as its environment, similar to the Sybyl atom type.

For proteins, 158 residue-dependent atom types that were used in DOPE[100]

statistical potential were tested in this study. In addition, we explored combining the 158 protein atom types, to obtain a smaller number of protein atom classes and thus improve the statistical robustness (and therefore accuracy) of the resulting statistical potential. Two protein atom types could be combined without loss of information if each one of the 26 protein-ligand atom distance distribution comparisons had a sufficiently small χ^2 (tail probability $p = 0.05$).[128] Surprisingly, for the maximal distance bound of 6 Å, only 16 pairs of protein atom types were nearly identical to each other (Table S4). Therefore, we did not combine the individual protein atom types. For ligands, 26 atom types were used, based on the Sybyl atom type classification (Table 1). Although the chemical space of small molecules is much more diverse than that of the 20 standard amino acid residues, we did not explore a more fine-grained atom type classification for practical reasons.

Future improvements

For 12 of the 100 benchmark targets, a correct binding pose was not identified by PoseScore as the best scoring pose. For 10 of the 12 targets, the assessment of binding between protein and ligand was affected by the lack of considering crystal water molecules, a cofactor, or the transition state. This observation indicates that the current protein-ligand statistical potential could be improved by considering ligand-water interactions, ligand-cofactor interactions, and transition states. Currently, the derivation of statistical potentials that explicitly consider ligand-water and ligand-cofactor interactions is limited by the size and accuracy of the sample of known complex structures. The accuracy of a statistical potential that is derived from a sample of experimental structures clearly depends on the resolution of the experimental structures. More accurate structures result in a more accurate statistical potential, all other things being equal. In this study, we derived the statistical potentials from 2.5 Å structures. We also tested a resolution cutoff of 2 Å. In comparison to PoseScore and RankScore, the number of available structures decreased from 8885 to 5353, and the newly derived statistical potentials performed worse (data not shown). Clearly, the disadvantage of a smaller sample outweighed the advantage of more accurate structures in the sample, a problem that would only be larger for statistical potentials for ligand-water and ligand-cofactor interactions. This dilemma may be at least partially solved by deriving a statistical potential from small molecule crystals in CSD, which provides a rich source of interaction geometries for small molecules at high resolution (292,539 structures with R-factor < 0.050). One successful example of such a statistical potential is DrugScore^{CSD} that

showed a substantial improvement compared to DrugScore^{PDB} in the recognition of near-native ligand binding poses.[35]

As shown in eqn. 4, the statistical potential for the protein-ligand complex should be a sum of three terms, including those for the intramolecular protein interactions, intramolecular ligand interactions, and intermolecular protein-ligand interactions. In this study, we only developed statistical potentials for the protein-ligand intermolecular interactions. More accurate scoring could be achieved by combining PoseScore/RankScore with other scoring functions such as the DOPE potential that measure the intramolecular protein interactions.

Conclusions

We conclude by returning to the four questions posed in Introduction.

Can a statistical potential be used for distinguishing between ligands and nonbinding molecules, in addition to recognizing native binding modes?

Yes. We developed two statistical potentials from a sample of X-ray structures of protein-ligand complexes: PoseScore was optimized for distinguish correct binding poses from geometric decoys and RankScore was optimized for distinguishing ligands from screening decoys. The reference states of PoseScore and RankScore are different, because of the differences in the values of three adjustable parameters. PoseScore scored a correct binding pose best among 100 decoys for 88% of all cases in the benchmark set containing 100 protein-ligand complexes. Furthermore, a correct binding pose was ranked the best, top 5, and top 10 by PoseScore for 88 (88%), 97 (97%), and 99 (99%) targets in the testing set, respectively. RankScore improved the ligand enrichment (logAUC) and early enrichment (EF_1) by rescoring the results by DOCK for 13 and 14 targets, respectively. Furthermore, RankScore ranked at least one annotated ligand within the top 500 scored compounds for all targets.

Can the accuracy of a statistical potential be improved by adding "negative" information, such as geometric decoys of the true ligands?

Yes. $PoseScore^{dock}$ in which the reference state was computed from geometric decoys that had all-atom RMSD errors of more than 2 Å from crystal binding poses (random solution) showed higher accuracy (74%) in detecting near-native solutions from geometric decoys, in comparison to PoseScore (69%). However, the performance of RankScore was not improved by including the geometric decoys.

What is the accuracy of scoring complexes with modeled protein structures relative to that with crystallographic structures?

For ligand pose prediction, $PoseScore^{model}$ showed higher accuracy (78%) in detecting near-native solutions from geometric decoys, in comparison to PoseScore (69%) and $PoseScore^{dock}$ (74%). For ligand enrichment, $RankScore^{model}$ also improved the average logAUC by 1.6, relative to DOCK. However, this improvement was less than that by RankScore (6.8) applied to complexes with crystal structures.

What are the differences between the reference states for protein-ligand and protein-protein statistical potentials?

Proteins are finite systems. Therefore, for a folded protein in the ligand-free state, the reference distribution should not increase in r^2 as in an infinite system, but in r^{α} where α is smaller than < 2. The optimal value of α was found empirically to be approximately 1.6. In contrast, in protein-ligand complexes, each ligand atom is partly surrounded by other ligand atoms. As a result, the number of protein-ligand atom pairs should increase faster with the distance r than the number of protein atom pairs. Correspondingly, the optimal value of α was found empirically to be around 3.

Acknowledgements

We thank Dr. Shaomeng Wang for the geometry test set, Dr. Xiaoqin Zou for ITScore, Dr. Gerhard Klebe for DrugScore^{PDB}, Michael Mysinger and Yiqun Cao for discussions about the DUD database and the automated docking pipeline, Dr. Min-yi Shen and Dr. Jiang Zhu for discussion about the theory used for computing statistical potentials, and Dr. Daniel Russel and Keren Lasker for implementation of the scoring function in IMP. We are supported by National Institutes of Health grants (GM71790, GM54762, and GM093342 to AS, GM59957 and GM71896 to BKS). We are also grateful to Ron Conway, Mike Homer, Hewlett-Packard, IBM, NetApp, and Intel for computing hardware gifts.

Supporting information

The list of PDB codes of 70 protein-ligand X-ray complexes used to train the PoseScore, the list of PDB codes of 19 protein-ligand X-ray complexes in DUD benchmark (DUD-1) used to train the RankScore, and the list of PDB codes of 19 protein-ligand X-ray complexes in DUD benchmark (DUD-2) used to test the RankScore. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Brooijmans, N.; Kuntz, I. D., Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003,** 32, 335-373.
2. Leach, A. R.; Shoichet, B. K.; Peishoff, C. E., Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps. *J. Med. Chem.* **2006,** 49, (20), 5851-5855.
3. Klebe, G., Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006,** 11, (13-14), 580-594.
4. Gilson, M. K.; Zhou, H. X., Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007,** 36, 21-42.
5. Jain, A. N., Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci.* **2006,** 7, (5), 407-420.
6. de Azevedo, W. F.; Dias, R., Evaluation of ligand-binding affinity using polynomial empirical scoring functions. *Bioorg. Med. Chem.* **2008,** 16, (20), 9378-9382.
7. Dominy, B. N., Molecular recognition and binding free energy calculations in drug development. *Curr. Pharm. Biotechnol.* **2008,** 9, (2), 87-95.
8. Rester, U., Dock around the clock - Current status of small molecule docking and scoring. *QSAR Comb. Sci.* **2006,** 25, (7), 605-615.
9. Coupez, B.; Lewis, R. A., Docking and scoring - Theoretically easy, practically impossible? *Curr. Med. Chem.* **2006,** 13, (25), 2995-3003.
10. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004,** 47, (12), 3032-3047.
11. Wang, R. X.; Lu, Y. P.; Wang, S. M., Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003,** 46, (12), 2287-2303.
12. Gohlke, H.; Klebe, G., Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie-International Edition* **2002,** 41, (15), 2645-2676.
13. Shoichet, B. K.; McGovern, S. L.; Wei, B. Q.; Irwin, J. J., Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002,** 6, (4), 439-446.
14. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.* **1995,** 117, (19), 5179-5197.
15. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996,** 118, (45), 11225-11236.
16. Daura, X.; Mark, A. E.; van Gunsteren, W. F., Parametrization of aliphatic CHn united atoms of GROMOS96 force field. *J. Comput. Chem.* **1998,** 19, (5), 535-547.
17. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004,** 47, (7), 1739-1749.

18. Halgren, T. A., Merck molecular force field .5. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **1996,** 17, (5-6), 616-641.
19. Meng, E. C.; Shoichet, B. K.; Kuntz, I. D., Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992,** 13, (4), 505-524.
20. 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, (35), 8033-8043.
21. Nevins, N.; Lii, J. H.; Allinger, N. L., Molecular mechanics (MM4) calculations on conjugated hydrocarbons. *J. Comput. Chem.* **1996,** 17, (5-6), 695-729.
22. Clark, M.; Cramer, R. D.; Vanopdenbosch, N., Validation of the General-Purpose Tripos 5.2 Force-Field. *J. Comput. Chem.* **1989,** 10, (8), 982-1012.
23. Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F., A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004,** 25, (13), 1656-1676.
24. Shoichet, B. K.; Leach, A. R.; Kuntz, I. D., Ligand solvation in molecular docking. *Proteins-Structure Function and Genetics* **1999,** 34, (1), 4-16.
25. Mysinger, M. M.; Shoichet, B. K., Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *J. Chem. Inf. Model.* **2010,** 50, (9), 1561-1573.
26. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998,** 19, (14), 1639-1662.
27. Muegge, I.; Martin, Y. C., A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999,** 42, (5), 791-804.
28. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M., BLEEP - Potential of mean force describing protein-ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.* **1999,** 20, (11), 1177-1185.
29. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M., BLEEP - Potential of mean force describing protein-ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999,** 20, (11), 1165-1176.
30. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000,** 295, (2), 337-356.
31. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T., Molecular Recognition of the Inhibitor Ag-1343 by Hiv-1 Protease - Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* **1995,** 2, (5), 317-324.
32. Stahl, M.; Rarey, M., Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001,** 44, (7), 1035-1042.
33. Ishchenko, A. V.; Shakhnovich, E. I., SMall molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions. *J. Med. Chem.* **2002,** 45, (13), 2770-2780.
34. Ozrin, V. D.; Subbotin, M. V.; Nikitin, S. M., PLASS: Protein-ligand affinity statistical score - a knowledge-based force-field model of interaction derived from the PDB. *J. Comput. Aided Mol. Des.* **2004,** 18, (4), 261-270.
35. Velec, H. F. G.; Gohlke, H.; Klebe, G., DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition

rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005,** 48, (20), 6296-6303.
36. Kirtay, C. K.; Mitchell, J. B. O.; Lumley, J. A., Knowledge based potentials: the reverse Boltzmann methodology, virtual screening and molecular weight dependence. *QSAR Comb. Sci.* **2005,** 24, (4), 527-536.
37. Catana, C.; Stouten, P. F. W., Novel, customizable scoring functions, parameterized using N-PLS, for structure-based drug discovery. *J. Chem. Inf. Model.* **2007,** 47, (1), 85-91.
38. Huang, S. Y.; Zou, X. Q., An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006,** 27, (15), 1866-1875.
39. Huang, S. Y.; Zou, X. Q., An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006,** 27, (15), 1876-1882.
40. Huang, S. Y.; Zou, X. Q., Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2010,** 50, (2), 262-273.
41. Murray, C. W.; Auton, T. R.; Eldridge, M. D., Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput. Aided Mol. Des.* **1998,** 12, (5), 503-519.
42. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions .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, (5), 425-445.
43. Olson, A. J.; Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998,** 19, (14), 1639-1662.
44. Wang, S. M.; Wang, R. X.; Lai, L. H., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.* **2002,** 16, (1), 11-26.
45. Hu, X.; Balaz, S.; Shelver, W. H., A practical approach to docking of zinc metalloproteinase inhibitors. *J. Mol. Graph. Model.* **2004,** 22, (4), 293-307.
46. Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M., LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graph. Model.* **2005,** 23, (5), 395-407.
47. Antes, I.; Merkwirth, C.; Lengauer, T., POEM: Parameter optimization using ensemble methods: Application to target specific scoring functions. *J. Chem. Inf. Model.* **2005,** 45, (5), 1291-1302.
48. 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, (22), 6821-6831.
49. Linusson, A.; Lindstrom, A.; Pettersson, F.; Almqvist, F.; Berglund, A.; Kihlberg, J., Hierarchical PLS modeling for predicting the binding of a comprehensive set of structurally diverse protein-ligand complexes. *J. Chem. Inf. Model.* **2006,** 46, (3), 1154-1167.
50. Klebe, G.; Sotriffer, C. A.; Sanschagrin, P.; Matter, H., SFCscore: Scoring functions for affinity prediction of protein-ligand complexes. *Proteins-Structure Function and Bioinformatics* **2008,** 73, (2), 395-419.

51. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000,** 28, (1), 235-242.
52. Allen, F. H., The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B-Structural Science* **2002,** 58, 380-388.
53. Hendlich, M.; Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J., Identification of Native Protein Folds Amongst a Large Number of Incorrect Models - the Calculation of Low-Energy Conformations from Potentials of Mean Force. *J. Mol. Biol.* **1990,** 216, (1), 167-180.
54. Colovos, C.; Yeates, T. O., Verification of Protein Structures - Patterns of Nonbonded Atomic Interactions. *Protein Sci.* **1993,** 2, (9), 1511-1519.
55. Sippl, M. J., Boltzmann Principle, Knowledge-Based Mean Fields and Protein-Folding - an Approach to the Computational Determination of Protein Structures. *J. Comput. Aided Mol. Des.* **1993,** 7, (4), 473-501.
56. Furuichi, E.; Koehl, P., Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins-Structure Function and Bioinformatics* **1998,** 31, (2), 139-149.
57. Kocher, J. P. A.; Rooman, M. J.; Wodak, S. J., Factors Influencing the Ability of Knowledge-Based Potentials to Identify Native Sequence-Structure Matches. *J. Mol. Biol.* **1994,** 235, (5), 1598-1613.
58. Huang, E. S.; Subbiah, S.; Levitt, M., Recognizing Native Folds by the Arrangement of Hydrophobic and Polar Residues. *J. Mol. Biol.* **1995,** 252, (5), 709-720.
59. Rooman, M. J.; Wodak, S. J., Are database-derived potentials valid for scoring both forward and inverted protein folding? *Protein Eng.* **1995,** 8, (9), 849-858.
60. Jernigan, R. L.; Bahar, I., Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **1996,** 6, (2), 195-209.
61. Jones, D. T.; Thornton, J. M., Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **1996,** 6, (2), 210-216.
62. Miyazawa, S.; Jernigan, R. L., Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **1996,** 256, (3), 623-644.
63. Moult, J., Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* **1997,** 7, (2), 194-199.
64. Park, B.; Levitt, M., Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **1996,** 258, (2), 367-92.
65. Park, B. H.; Huang, E. S.; Levitt, M., Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **1997,** 266, (4), 831-46.
66. Reva, B. A.; Finkelstein, A. V.; Sanner, M. F.; Olson, A. J., Residue-residue mean-force potentials for protein structure recognition. *Protein Eng.* **1997,** 10, (8), 865-876.
67. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D., Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997,** 268, (1), 209-25.
68. Vajda, S.; Sippl, M.; Novotny, J., Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol.* **1997,** 7, (2), 222-228.
69. Gilis, D.; Rooman, M., Predicting protein stability changes upon mutation using database-derived potentials: Solvent accessibility determines the importance of

local versus non-local interactions along the sequence. *J. Mol. Biol.* **1997,** 272, (2), 276-290.
70. Samudrala, R.; Moult, J., An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **1998,** 275, (5), 895-916.
71. Betancourt, M. R.; Thirumalai, D., Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* **1999,** 8, (2), 361-369.
72. Jones, D. T., GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **1999,** 287, (4), 797-815.
73. Rojnuckarin, A.; Subramaniam, S., Knowledge-based interaction potentials for proteins. *Proteins-Structure Function and Genetics* **1999,** 36, (1), 54-67.
74. Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D., Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins-Structure Function and Genetics* **1999,** 34, (1), 82-95.
75. Bastolla, U.; Vendruscolo, M.; Knapp, E. W., A statistical mechanical method to optimize energy functions for protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **2000,** 97, (8), 3977-3981.
76. Chiu, T. L.; Goldstein, R. A., How to generate improved potentials for protein tertiary structure prediction: a lattice model study. *Proteins* **2000,** 41, (2), 157-63.
77. Gatchell, D. W.; Dennis, S.; Vajda, S., Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins-Structure Function and Genetics* **2000,** 41, (4), 518-534.
78. Lazaridis, T.; Karplus, M., Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000,** 10, (2), 139-145.
79. Vendruscolo, M.; Najmanovich, R.; Domany, E., Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins-Structure Function and Genetics* **2000,** 38, (2), 134-148.
80. Lu, H.; Skolnick, J., A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **2001,** 44, (3), 223-32.
81. Melo, F.; Sanchez, R.; Sali, A., Statistical potentials for fold assessment. *Protein Sci.* **2002,** 11, (2), 430-48.
82. Keasar, C.; Levitt, M., A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.* **2003,** 329, (1), 159-74.
83. Zhou, H.; Zhou, Y., Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **2002,** 11, (11), 2714-26.
84. Zhou, H.; Zhou, Y., Erratum: Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **2003,** 12, (9), 2121.
85. Betancourt, M. R.; Skolnick, J., Local propensities and statistical potentials of backbone dihedral angles in proteins. *J. Mol. Biol.* **2004,** 342, (2), 635-649.
86. Buchete, N. V.; Straub, J. E.; Thirumalai, D., Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.* **2004,** 14, (2), 225-232.
87. Buchete, N. V.; Straub, J. E.; Thirumalai, D., Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci.* **2004,** 13, (4), 862-874.

88. Wang, K.; Fain, B.; Levitt, M.; Samudrala, R., Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct. Biol.* **2004,** 4, (1), 8.
89. Zhang, C.; Liu, S.; Zhou, H.; Zhou, Y., The dependence of all-atom statistical potentials on structural training database. *Biophys. J.* **2004,** 86, (6), 3349-58.
90. Chen, W. W.; Shakhnovich, E. I., Lessons from the design of a novel atomic potential for protein folding. *Protein Sci.* **2005,** 14, (7), 1741-1752.
91. Fang, Q. J.; Shortle, D., A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. *Proteins-Structure Function and Bioinformatics* **2005,** 60, (1), 90-96.
92. Qiu, J.; Elber, R., Atomically detailed potentials to recognize native and approximate protein structures. *Proteins-Structure Function and Bioinformatics* **2005,** 61, (1), 44-55.
93. Summa, C. M.; Levitt, M.; DeGrado, W. F., An atomic environment potential for use in protein structure prediction. *J. Mol. Biol.* **2005,** 352, (4), 986-1001.
94. Dehouck, Y.; Gilis, D.; Rooman, M., A new generation of statistical potentials for proteins. *Biophys. J.* **2006,** 90, (11), 4010-4017.
95. Eramian, D.; Shen, M. Y.; Devos, D.; Melo, F.; Sali, A.; Marti-Renom, M. A., A composite score for predicting errors in protein structure models. *Protein Sci.* **2006**.
96. Lu, M. Y.; Dousis, A. D.; Ma, J. P., OPUS-PSP: An orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.* **2008,** 376, (1), 288-301.
97. Wu, Y. H.; Lu, M. Y.; Chen, M. Z.; Li, J. L.; Ma, J. P., OPUS-Ca: A knowledge-based potential function requiring only C alpha positions. *Protein Sci.* **2007,** 16, (7), 1449-1463.
98. Rykunov, D.; Fiser, A., New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics* **2010,** 11, -.
99. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *J. Med. Chem.* **2006,** 49, (23), 6789-6801.
100. Shen, M. Y.; Sali, A., Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006,** 15, (11), 2507-2524.
101. Sippl, M. J., Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **1990,** 213, (4), 859-83.
102. Sali, A.; Blundell, T. L., Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **1993,** 234, (3), 779-815.
103. Fan, H.; Irwin, J. J.; Webb, B. M.; Klebe, G.; Shoichet, B. K.; Sali, A., Molecular Docking Screens Using Comparative Models of Proteins. *J. Chem. Inf. Model.* **2009,** 49, (11), 2512-2527.
104. Lorber, D. M.; Shoichet, B. K., Flexible ligand docking using conformational ensembles. *Protein Sci.* **1998,** 7, (4), 938-950.
105. Lorber, D. M.; Shoichet, B. K., Hierarchical docking of databases of multiple ligand conformations. *Curr. Top. Med. Chem.* **2005,** 5, (8), 739-749.
106. Irwin, J. J.; Shoichet, B. K.; Mysinger, M. M.; Huang, N.; Colizzi, F.; Wassam, P.; Cao, Y. Q., Automated Docking Screens: A Feasibility Study. *J. Med. Chem.* **2009,** 52, (18), 5712-5720.
107. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W., Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007,** 50, (4), 726-741.

108. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996,** 261, (3), 470-489.
109. 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, (8), 731-751.
110. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R., Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006,** 49, (2), 534-553.
111. Graves, A. P.; Brenk, R.; Shoichet, B. K., Decoys for docking. *J. Med. Chem.* **2005,** 48, (11), 3714-3728.
112. Pham, T. A.; Jain, A. N., Customizing scoring functions for docking. *J. Comput. Aided Mol. Des.* **2008,** 22, (5), 269-286.
113. Pham, T. A.; Jain, A. N., Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006,** 49, (20), 5856-5868.
114. Kamat, S. S.; Fan, H.; Sauder, J. M.; Burley, S. K.; Shoichet, B. K.; Sali, A.; Raushel, F. M., Enzymatic Deamination of the Epigenetic Base N-6-Methyladenine. *J. Am. Chem. Soc.* **2011,** 133, (7), 2080-2083.
115. Schlessingera, A.; Geiera, E.; Fan, H.; Irwin, J. J.; Shoichet, B. K.; Giacominia, K. M.; Sali, A., Structure-based discovery of prescription drugs that interact with the norepinephrine transporter, NET. *Proceedings of the National Academy of Sciences USA* **2011,** 108, (38), 15810-15815.
116. Carlsson, J.; Coleman, R. G.; Setola, V.; Irwin, J. J.; Fan, H.; Schlessinger, A.; Sali, A.; Roth, B. L.; Shoichet, B. K., Structure-based Ligand Discovery Against a Homology Model and X-ray Structure of the Dopamine D3 Receptor *Nat. Chem. Biol.* **2011,** doi:10.1038/nchembio.662.
117. Goble, A. M.; Fan, H.; Sali, A.; Raushel, F. M., The Discovery of New Enzymes: Cytokinin Deaminase. *ACS Chem. Biol.* **2011, doi:** 10.1021/cb200198c.
118. Sali, A.; Blundell, T. L., Comparative Protein Modeling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993,** 234, (3), 779-815.
119. Fiser, A.; Do, R. K. G.; Sali, A., Modeling of loops in protein structures. *Protein Sci.* **2000,** 9, (9), 1753-1773.
120. Leslie, A. G. W.; Moody, P. C. E.; Shaw, W. V., Structure of Chloramphenicol Acetyltransferase at 1.75-a Resolution. *Proc. Natl. Acad. Sci. U. S. A.* **1988,** 85, (12), 4133-4137.
121. Lewendon, A.; Murray, I. A.; Shaw, W. V.; Gibbs, M. R.; Leslie, A. G. W., Evidence for Transition-State Stabilization by Serine-148 in the Catalytic Mechanism of Chloramphenicol Acetyltransferase. *Biochemistry (Mosc.)* **1990,** 29, (8), 2075-2080.
122. Krishnan, R.; Zhang, E.; Hakansson, K.; Arni, R. K.; Tulinsky, A.; Lim-Wilby, M. S. L.; Levy, O. E.; Semple, J. E.; Brunck, T. K., Highly selective mechanism-based thrombin inhibitors: Structures of thrombin and trypsin inhibited with rigid peptidyl aldehydes. *Biochemistry-Us* **1998,** 37, (35), 12094-12103.
123. Mctigue, M. A.; Davies, J. F.; Kaufman, B. T.; Kraut, J., Crystal-Structure of Chicken Liver Dihydrofolate-Reductase Complexed with Nadp+ and Biopterin. *Biochemistry (Mosc.)* **1992,** 31, (32), 7264-7273.
124. Kurinov, I. V.; Harrison, R. W., Prediction of New Serine Proteinase-Inhibitors. *Nat. Struct. Biol.* **1994,** 1, (10), 735-743.

125. Tondi, D.; Morandi, F.; Bonnet, R.; Costi, M. P.; Shoichet, B. K., Structure-based optimization of a non-beta-lactam lead results in inhibitors that do not up-regulate beta-lactamase expression in cell culture. *J. Am. Chem. Soc.* **2005,** 127, (13), 4632-4639.
126. Powers, R. A.; Shoichet, B. K., Structure-based approach for binding site identification on AmpC beta-lactamase. *J. Med. Chem.* **2002,** 45, (15), 3222-3234.
127. Xu, D.; Tsai, C. J.; Nussinov, R., Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.* **1997,** 10, (9), 999-1012.
128. Cramer, H., Mathematical methods of statistics. *Princeton University Press* **1946**, 575.

Tables

Table 1. List of 26 atom types in small molecules	
Atom type	**Description**
C1	carbon sp
C2	carbon sp2
C3	carbon sp3
Car	carbon aromatic
Ccat	carbocation (C^+) used only in a guanidinium group
N1	nitrogen sp
N2	nitrogen sp2
N3	nitrogen sp3
N4	nitrogen sp3 positively charged
Nar	nitrogen aromatic
Nam	nitrogen amide
Npl3	nitrogen trigonal planar
O2	oxygen sp2
O3	oxygen sp3
Oco2	oxygen in carboxylate and phosphate groups
Oar	oxygen aromatic
S2	sulfur sp2
S3	sulfur sp3
So	sulfoxide sulfur
So2	sulfone sulfur
Sar	sulfur aromatic
P3	phosphorous sp3
F	fluorine
Cl	chlorine
Br	bromine
I	iodine

Table 2. Success rates of scoring functions on the test set containing 100 protein-ligand complexes

Scoring function	Success rate (%)[a]		
	Native	RMSD < 2 Å (Native / Decoy)	RMSD < 2 Å (Decoy)[b]
PoseScore	70	88	69
PoseScore^{dock}	70	90	74
ITScore/SE[c]		91	
ITScore		82	
DrugScore^{CSD}	77	87	66
DrugScore^{PDB}	49	72	65
Cerius2/PMF	32	52	48
Cerius2/PLP	52	76	70
Cerius2/LigScore	48	74	60
Cerius2/LUDI	23	67	64
AutoDock	8	62	66
X-Score	25	65	64
SYBYL/F-Score	38	74	68
SYBYL/G-Score	13	42	43
SYBYL/D-Score	3	26	29
SYBYL/ChemScore	7	35	34

[a]The success rate of each scoring function was evaluated by three different criteria and given as a percent with respect to the complexes analyzed. **Native**, the crystal conformation of the ligand received a better score than the 100 docking solutions of the ligand. **RMSD < 2 Å (Native / Decoy)**, the crystal conformation of the ligand or a docking solution with RMSD < 2 Å from the crystal conformation received the best rank. **RMSD < 2 Å (Decoy)**, a docking solution with RMSD < 2 Å from the crystal conformation received the best rank among the 100 docking solutions. [b]There is one or multiple docking solutions with RMSD < 2 Å from the crystal conformation for 91 complexes. [c]There are no data in literature for ITScore/SE and ITScore for the criteria **Native** and **RMSD < 2 Å (Decoy)**.

Table 3. Failures in ligand binding pose identification

PDB code	Resolution (Å)	Protein description	Ligand binding in crystal[a]	The best scored correct binding pose[b]	
				Rank	RMSD (Å)
1cla	2.34	chloramphenicol acetyltransferase	Hydrogen binding only to water	9	0
3cla	1.75	chloramphenicol acetyltransferase	Hydrogen binding only to water	4	0
4cla	2.00	chloramphenicol acetyltransferase	Hydrogen binding only to water	2	0
1rgl	2.00	ribonuclease T1	Hydrogen binding to water	5	0.89
3tmn	1.70	thermolysin	Hydrogen binding to water	7	0
1tlp	2.30	thermolysin	Transition state, with metal center	5	0
1zzz	1.90	trypsin	Transition state, with Ser195	2	0
2sns	1.50	staphylococcal nuclease	Transition state, with Arg35 and metal center	3	1.44
1tha	2.00	transthyretin	Two identical ligands	2	0
1dr1	2.20	dihydrofolate reductase	Ligand close to cofactor NADP+	2	1.02
1tni	1.90	trypsin		20	1.73
1tnj	1.80	trypsin		2	0

[a]The characterization of ligand binding in the crystal structure. [b]**correct binding** pose: ligand pose with RMSD < 2 Å from the ligand conformation in the crystal structure.

Table 4. Ligand enrichment (logAUC) by rescoring DOCK-generated docking poses

Protein targets	DOCK	Rescoring							
		Rank Score	ITScore	Drug Score ^{PDB}	FlexX	PMF	PLP	Screen Score	PLOP
ADA	22.7	**45.8**	*21.9*	9.3	**44.4**	**34.1**	15.4	**29.9**	**52.2**
DHFR	18.9	**62.0**	**76.6**	**48.0**	**86.7**	**82.9**	**74.7**	**86.2**	**45.7**
GART	35.3	**40.0**	**61.2**	**63.3**	**39.6**	**49.5**	**43.0**	**46.7**	15.6
Thrombin	29.4	22.1	*31.2*	25.2	17.9	22.1	21.4	19.4	16.7
AChE	38.5	*39.8*	12.1	29.8	15.8	18.1	24.3	18.3	**56.7**
AmpC (A chain)	47.4	10.3	11.4	5.1	13.9	7.3	8.6	11.1	15.4
AmpC (B chain)	53.8	19.3	19.8	6.6	11.3	8.6	9.5	9.8	7.5
COX-2	40.8	19.2	**56.8**	26.5	*42.5*	24.9	**47.1**	**46.9**	31.7
HIVPR	11.9	**33.2**	**24.0**	**44.6**	*9.4*	*13.8*	**16.4**	*12.8*	*11.8*
HMGR	40.9	35.3	30.6	27.8	7.0	14.8	11.1	8.2	10.8
NA	47.6	**58.4**	*50.4*	14.1	10.8	*44.6*	12.4	11.4	18.3
PARP	8.2	**40.7**	**25.6**	**13.1**	**64.7**	**32.3**	**39.4**	**53.2**	**12.4**
HSP90	24.6	**29.6**	**30.3**	18.6	*25.4*	**30.2**	*24.4*	*26.1*	15.0
EGFr	21.5	17.0	*21.6*	12.0	*24.1*	*20.3*	17.3	*21.7*	**27.3**
SRC	9.5	**26.6**	**17.5**	**16.2**	**32.8**	**19.7**	**16.4**	**26.6**	**13.4**
TK	63.5	**75.4**	48.3	28.0	*64.6*	33.9	50.5	56.3	**75.1**
$ER_{agonist}$	55.4	**61.9**	50.6	23.6	42.2	34.1	42.6	39.3	36.6
GR	20.5	**28.2**	*23.1*	**26.1**	**26.6**	**32.3**	**30.9**	**28.0**	*22.0*
PPARg	4.4	**17.6**	**11.7**	**37.6**	*5.5*	**10.3**	*6.7*	*6.0*	**10.9**
RXRa	37.9	**45.1**	**43.1**	**52.2**	23.3	**42.5**	*40.0*	30.7	23.8
Average	30.5	37.3	34.1	27.4	31.4	29.9	28.6	30.5	26.9

The ligand enrichment is represented by logAUC. **DOCK**, the ligand enrichment is calculated by ranking all the docked compounds with their DOCK scores. **Rescoring**, the DOCK-generated docking poses were rescored by an alternative scoring function and the ligand enrichment is calculated by reranking all the docked compounds with their new scores. The enrichment from virtual screens was used as control. Rescoring is defined to be improved if the ligand enrichment is larger than the control by 3 or more in the logAUC units (bold); otherwise, the enrichment values were considered to be comparable (italic). $DrugScore^{PDB}$ was computed with and without the solvent-accessible surface-dependent term. The one without the surface contribution actually performed better and was present here. The ligand binding site in the chain A of AmpC β-lactamase is distorted, in contrast to that in the chain B.

Table 5. The rank of best scored ligand by DOCK and RankScore				
Protein targets	**Ligand screens by DOCK**		**Ligand rescoring by RankScore**	
	EF_1[a]	**Rank of the best scored ligand**	**EF_1**	**Rank of the best scored ligand**
ADA	0.0	2989	**34.8**	**74**
DHFR	1.5	166	**47.3**	**2**
GART	9.5	123	**23.8**	**72**
Thrombin	15.4	21	6.2	**1**
AChE	6.7	304	**10.5**	**107**
AmpC (A chain)	0.0	1098	0.0	27680
AmpC (B chain)	19.0	628	4.8	**14**
COX-2	22.1	12	0.9	74
HIVPR	0.0	5200	**13.2**	**24**
HMGR	34.3	19	17.1	**3**
NA	40.8	15	**42.9**	**1**
PARP	0.0	15976	**12.1**	**292**
HSP90	0.0	2967	**4.2**	**108**
EGFr	1.1	257	**2.9**	**103**
SRC	0.0	7536	**5.6**	**2**
TK	50.0	319	**54.6**	**40**
$ER_{agonist}$	28.4	3	**41.8**	4
GR	7.7	9	**10.3**	**1**
PPARg	0.0	16898	**1.2**	**462**
RXRa	30.0	1	30.0	5

[a]The enrichment factor EF_1 is the percent of actual ligands found in the top 1% ranked subset of all database compounds. [b]The ligand binding site in chain A structure of AmpC β-lactamase is partially broken while fine in the chain B structure.

Figures

Figure 1. ***The performance of the statistical potential affected by the distance cut-off, showed on the training sets.*** (a) Two parameters of the potential were fixed ($w_{ref} = 0.4$, $w_{uni} = 0$), the potential showed the highest accuracy in ligand pose detection when the other parameter r_{max} is set to 6 Å, selecting correct binding mode for 64 (91%) targets in the training set of 70 proteins. (b) Two parameters of the potential were fixed ($w_{ref} = 0.4$, $w_{uni} = 0$) the potential showed the highest accuracy in the rescoring when the other parameter r_{max} is set to 6 Å, improving enrichment (logAUC) for 14 targets in the DUD-1 training set.

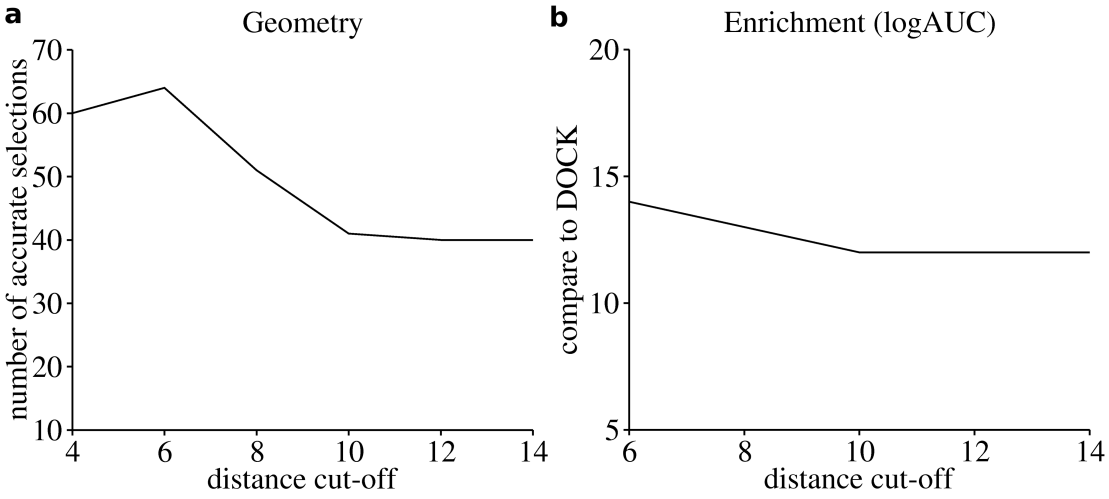


Figure 2. ***Four examples of accurate ligand pose prediction from the PoseScore test set.*** For each target, the crystal structure of the protein binding site and the co-crystallized ligand (solid stick, green) as well as the best-ranked ligand geometric decoy (solid stick, yellow) are shown. (a) Thrombin (1a46). The crystal structure of the ligand was ranked 1. A geometric decoy with the 1.39 Å RMSD error was ranked 2. (b) Carbonic anhydrase I (1bzm). The crystal structure of the ligand was ranked 3. A geometric decoy with the 1.65 Å RMSD error was ranked 1. (c) Elastase (1ela). The crystal structure of the ligand was ranked 1. A geometric decoy with the 1.37 Å RMSD error was ranked 2. (d) Streptavidin (1sre). The crystal structure of the ligand was ranked 5. A geometric decoy with the 1.39 Å RMSD error was ranked 1.

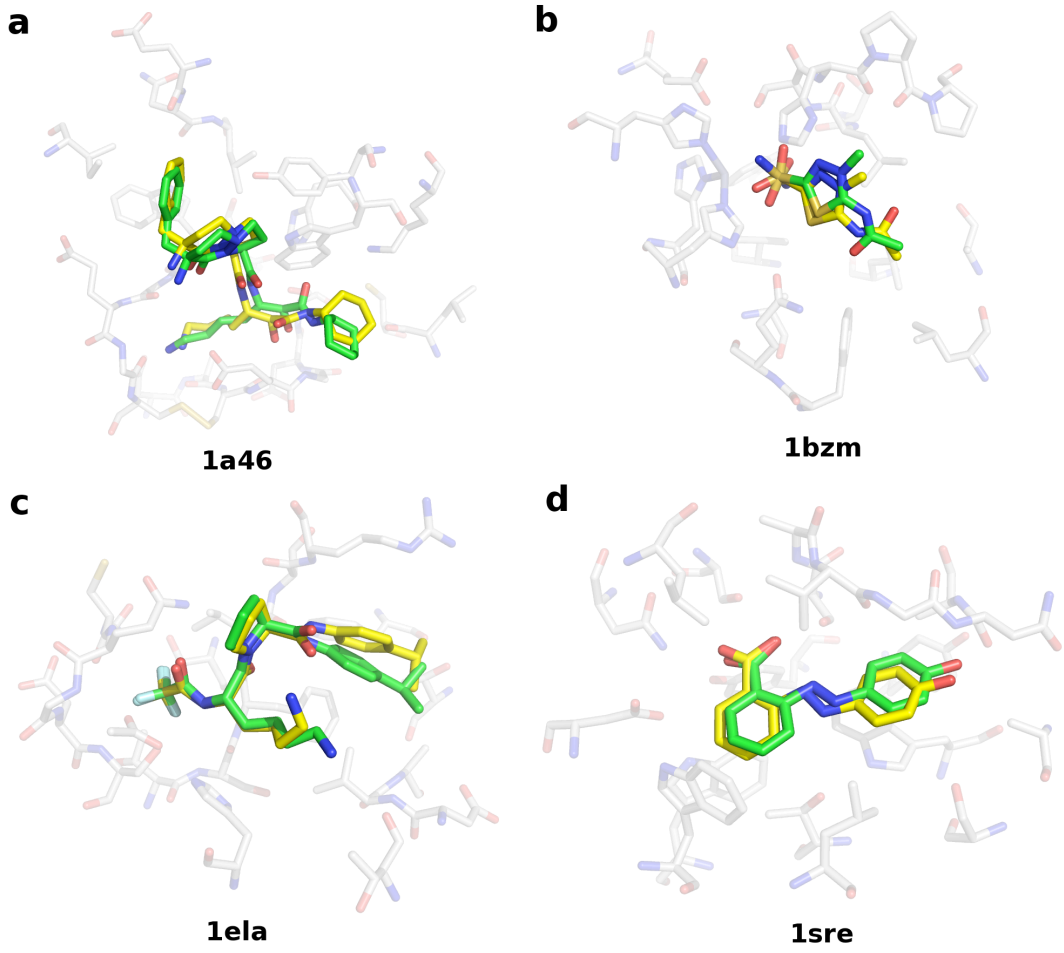


Figure 3. ***Four examples of inaccurate ligand pose prediction from the PoseScore test set.*** For each target, the crystal structure of the protein binding site, the co-crystallized ligand, and the highest ranking geometric decoy of the ligand are presented as in Figure 2. See Results for more detail.

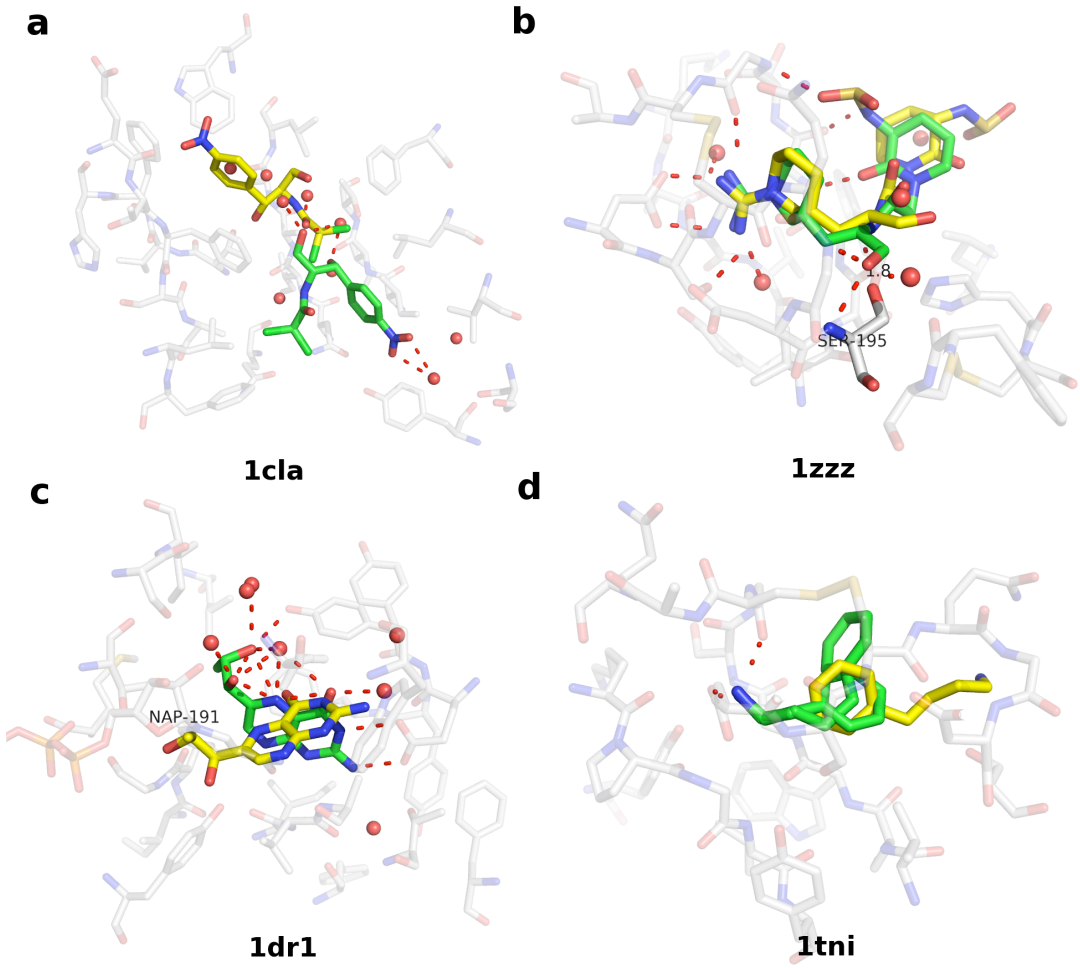


Figure 4. ***Ligand poses of Amp****C* ***β-lactamase from the test set of RankScore.*** (a) 2D images of AmpC ligands HTC and CTC (b) Docking poses of HTC (yellow stick) and CTC (blue stick) generated by screening against the B chain of AmpC structure (PDB code: 1xgj).

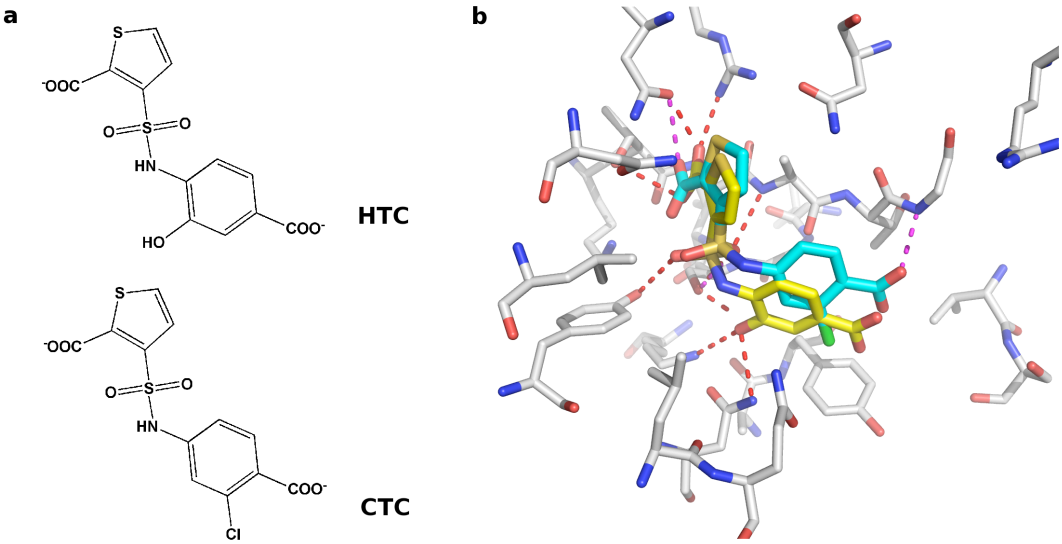


Figure 5. ***The effect of the parameter α on the performance of statistical potential derived using DFIRE formula, showed on the training set.*** α value was set to 1, 2, 3, 4, 5 and 6 in the calculation of the potential independently. For each α value, 5 different values were chosen for the maximal boundary r_{max} including 6 Å (black solid line), 8 Å (black dotted line), 10 Å (red solid line), 12 Å (red dotted line), 14 Å (blue solid line) respectively. The generated potentials were tested on the training set containing 70 proteins. The potential was the most accurate when α was set to 3 and r_{max} set to 6 Å.

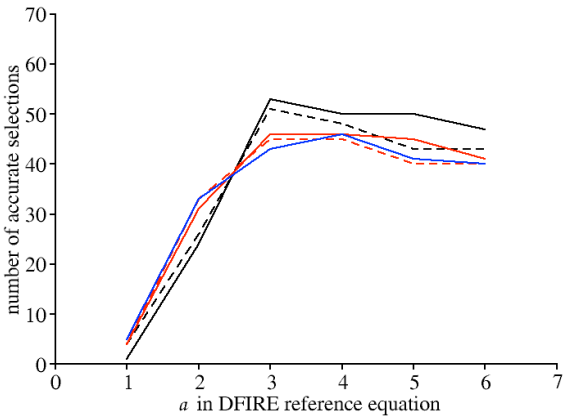


Figure 6. ***Schematic presentation of a protein-ligand complex.*** The protein is approximated as the outer sphere (solid line) and the ligand is completely embedded inside the protein as the inner sphere with a radius r. For the ligand atom positioned at a distance of d to the ligand center, the amount of protein-ligand atom pairs within certain distance R ($R \le R_{cutoff}$) is calculated by equation 12.

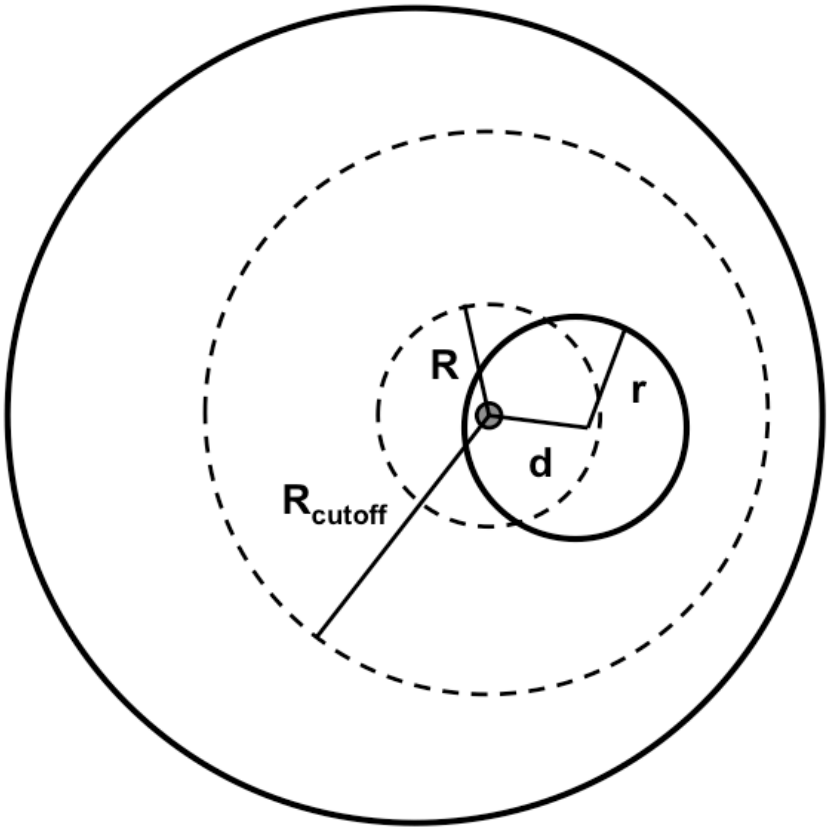
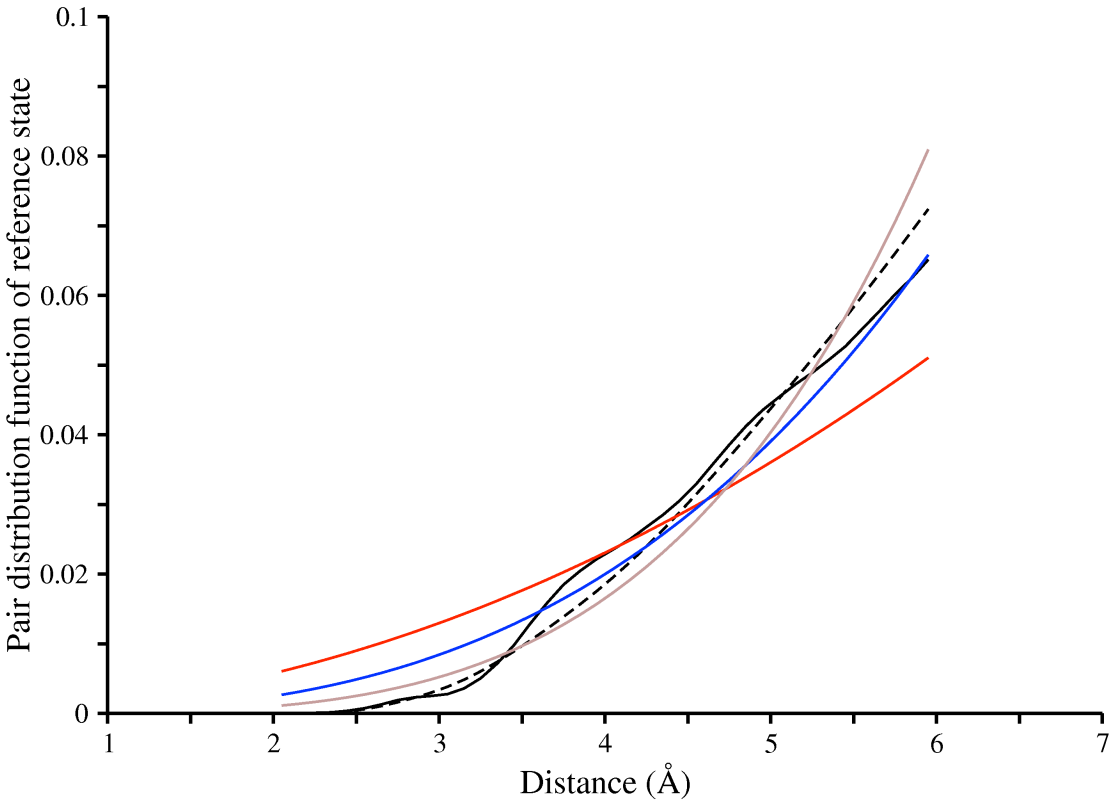


Figure 7. ***The probability distribution of protein-ligand atom pairs, assuming no difference between atom types.*** Five distributions are plotted. First, the distribution derived using eqn. 8, from the sample of X-ray structures of protein-ligand complexes (black solid line). Second, the distribution derived using eqn. 8, from the sample of docking poses that had RMSD error of larger than 2 Å with respect to the X-ray structures (black dashed line). Third, the distribution derived using eqn. 11 in which the parameter α was set to 2 (red solid line). Fourth, the distribution derived using eqn. 11 in which the parameter α was set to 3 (blue solid line). Fifth, the distribution derived using eqn. 11 in which the parameter α was set to 4 (brown solid line).



For Table of Contents use only

“Statistical Potential for Modeling and Ranking of Protein-Ligand Interactions”

Hao Fan, Dina Schneidman-Duhovny, John J. Irwin, Guangqiang Dong, Brian K. Shoichet, and Andrej Sali