# GlamDock: Development and Validation of a New Docking Tool on Several Thousand Protein−Ligand Complexes

Simon Tietze and Joannis Apostolakis*

Ludwig-Maximilians-University Munich, Institute for Informatics, TRU Bioinformatics, Amalienstr. 17, Munich, DE D-80333, Germany

Received April 5, 2007

In this study, we present GlamDock, a new docking tool for flexible ligand docking. GlamDock (version 1.0) is based on a simple Monte Carlo with minimization procedure. The main features of the method are the energy function, which is a continuously differentiable empirical potential, and the definition of the search space, which combines internal coordinates for the conformation of the ligand, with a mapping-based description of the rigid body translation and rotation. First, we validate GlamDock on a standard benchmark, a set of 100 protein−ligand complexes, which allows comparative evaluation to existing docking tools. The results on this benchmark show that GlamDock is at least comparable in efficiency and accuracy to the best existing docking tools. The main focus of this work is the validation on the scPDB database of protein−ligand complexes. The size of this data set allows a thorough analysis of the dependencies of docking accuracy on features of the protein−ligand system. In particular, it allows a two-dimensional analysis of the results, which identifies a number of interesting dependencies that are generally lost or even misinterpreted in the one-dimensional approach. The overall result that GlamDock correctly predicts the complex structure in practically half of the cases in the scPDB is important not only for screening ligands against a particular protein but even more so for inverse screening, that is, the identification of the correct targets for a particular ligand.

## INTRODUCTION

Docking of flexible ligands into rigid binding sites is a mature computational technique for the prediction of complex structures and forms the basis of structure-based virtual screening. While a number of methods[1−5] have been published in the past 10 years on docking, the validation has in general been rather limited, being performed on benchmarks that have been either completely selected or manually curated by the developers themselves. However, as both screening and inverse screening[6−8] (the search for the targets for a given ligand) are gaining in importance, and rely on the automatic preprocessing of ligands and binding sites, it is important to evaluate the performance of docking methods on automatically generated data sets.

In addition, it has previously been noted[9] that single docking tools rarely consistently outperform others in comparative studies. Rather, each program performs best on some subset of targets. In applications of docking tools, it would however be desirable to know a priori which tool can be expected to perform well for a given protein−ligand combination. One possibility to make this type of prediction is to relate prediction quality for a given docking tool to relevant descriptors of the protein and the ligand to be docked. Thus, for example, one can study the dependence of docking accuracy on the size of the ligand. In the simplest case, the results on the benchmark set would need to be split into two sets: one containing small, the other containing large ligands. This reduces the size of the sets and with that the significance of the results. In order to study higher-order dependencies, it is necessary to perform further splits in the data. For example, it is known that size and flexibility as measured by the number of torsional degrees of freedom in the molecule correlate. Therefore, any dependence on size could in principle be an indirect effect of flexibility. In order to resolve this question, it is further necessary to further split the sets obtained by splitting according to size. This further reduces the statistical significance of the results.

While the publication of high-quality and diverse manually curated benchmarks for protein−ligand docking[10,11] is generally welcome, these benchmarks are often too small to accurately characterize and compare different tools. Any attempt to investigate the relative weaknesses and strengths of the tools on the basis of properties of the protein and the ligand will be blurred even further by the small size of any subset of these benchmarks or might be impossible, as the relevant examples may have been removed during the selection of the complexes. Thus, in order to gain qualitative and quantitative insight into the expected quality of docking tools on novel targets, more extensive benchmarks should be used. The current practice of using a priori rules for selecting benchmarks from the Protein Data Bank (PDB) has the disadvantage that the selection rules have to be taken in good faith. For example, many benchmark sets have been limited to structures with a resolution better than a certain threshold; for example, the newest benchmark set from Astex[11] uses a threshold of 2.0 Å. While we basically agree with the rationale that low-resolution structures allow alternative interpretations of the atomic positions than the one presented in the PDB structure, many structures relevant in ligand design are initially of poor quality, and we would

* Corresponding author. E-mail: joannis.apostolakis@bio.ifi.lmu.de.

10.1021/ci7001236 CCC: $37.00 
Published on Web 06/23/2007

like to know how well (or poorly) docking programs fare for moderate-quality structures. Our preference is to perform the benchmark on the complete data set and report the change in accuracy on given subsets.

The main disadvantages of comprehensive benchmark sets are the questions of CPU time and of manual time for preprocessing protein structures. The first is not relevant for flexible ligand/rigid protein docking, which currently takes approximately 1 min per docking. This type of docking is developed for the purposes of screening 10 000s of compounds against one or more binding site. It is necessary to invest at least comparable computational resources in validating the tools as is required in their typical applications. The second problem is indeed more relevant, since it certainly is not easy to perform accurate manual preparation of 1000s of complexes in a benchmark set. At the same time, in real screening scenaria, the modeler usually does invest significant effort into preparing the binding site. However, this point can also be handled correctly by evaluating the difference in docking accuracy in a predefined subset before and after manual preparation. Different modelers have different levels of experience in preparation, and we expect that results are subject to variation in dependence of the individual choices of the modeler.

Here, we will focus on automatically preprocessed complexes, as we find them in the scPDB database. The scPDB database[12] consists of all pharmacologically relevant binding sites in the PDB together with their ligands, and in our opinion, it represents an ideal test case for redocking. It has a number of interesting features compared to existing redocking benchmarks:

- It has been compiled independently and not by a developer and thus is expected to be unbiased with respect to the advantages or disadvantages of individual docking tools.
- It is larger by an order of magnitude than the largest docking benchmark from the literature. Thus, the obtained statistics are expected to be more significant.
- Due to the size of the database, the single binding sites and their ligands have been prepared automatically. On the one hand, this leaves room for errors, since scripts often overlook special cases of format or chemistry; on the other hand, we can be relatively certain that no special treatment of single binding sites has been performed to improve docking results.
- Finally, while the authors expressly aimed at a pharmacologically relevant database of binding sites, the criteria have been relatively lax, and thus the database covers a large part of chemical space, which also includes ligands of general biological interest.

The scPDB as a redocking benchmark represents the largest data set reported so far and should give a realistic assessment of redocking accuracy over automatically preprocessed ligands and binding sites. Interestingly, the results we obtain on the scPDB here can also be seen as a validation of the automatic preparation procedures in large databases such as the scPDB, since poorly preprocessed binding sites and/or ligands will generally lead to errors in structure prediction. The type of validation performed here, based on the docking quality of a large number of automatically preprocessed and docked ligands can, for example, be used for evaluating the accuracy of different protonation or general ligand preparation strategies.

In this study, we present and validate GlamDock, our in-house docking tool. GlamDock (version 1.0) is based on a simple Monte Carlo (MC) with minimization (also known as basin hopping)[2,13] procedure. The main features of the method are the energy function, which is a differentiable empirical potential, and the definition of the search space, which combines internal coordinates for the conformation of the ligand, with a mapping-based description of the rigid body translation and rotation. This definition of the search space is commonly used in genetic algorithm docking (e.g., in Gold[14]) or deterministic docking algorithms.[3,5] The interaction points in the active site used by our mapping are constructed using explicit interaction geometries as in FlexX.[3] However, these points are represented by probe molecules as in Surflex,[5] with aromatic rings replacing the less specific methyl groups used by Surflex. In the context of Monte Carlo, interaction mapping allows the definition of a simple Monte Carlo move, which nevertheless has the effect of placing the ligand in a more or less reasonable position into the binding site with high probability. This initial placement of the ligand is refined using a gradient-based minimization procedure. Unlike previous approaches, the minimized complex is re-encoded in terms of the discrete interaction mapping. This can be seen as a generalization of the continuous coordinate re-encoding Lamarckian genetic algorithm used by Autodock.[15] The main novelty of our sampling strategy is the combination of interaction mapping-based placement with basin hopping. In our experience, the former leads to initial placements showing favorable interactions with the binding site, while the latter allows the resolution of clashes and the local optimization of the scoring function.

Here, we present the details of the method, validate it, and compare its results to existing docking software. First, we validate it on a standard benchmark, a set of 100 protein−ligand complexes provided by Kellenberger et al.,[10] which allows comparative evaluation to existing docking tools. The results from this benchmark show that the new method is at least comparable in efficiency and accuracy to the best existing docking tools. We then extend the validation of GlamDock to the complete scPDB database. While the results obtained on the scPDB are somewhat worse than those obtained for the smaller benchmark, they show that GlamDock predicts the correct structure in almost half of the cases. The size of the scPDB allows a thorough analysis of docking quality dependence on a number of different properties of the ligands and the binding site. In particular, a two-dimensional analysis of the effect of different features on docking accuracy is performed, which among other things demonstrates how the analysis of single-feature dependencies can be misleading.

## MATERIALS AND METHODS

**Benchmark Data.** The Kellenberger data set[10] was obtained from the authors and was used unmodified in this work. It consists of 100 diverse protein−ligand complexes and has recently been used in a comparative study of eight protein−ligand docking tools.[10]

In addition, the scPDB database[12]was used as an extensive benchmark to evaluate the performance of GlamDock in a fully automatized setting. The scPDB is a database of crystallographically resolved protein binding sites of phar-

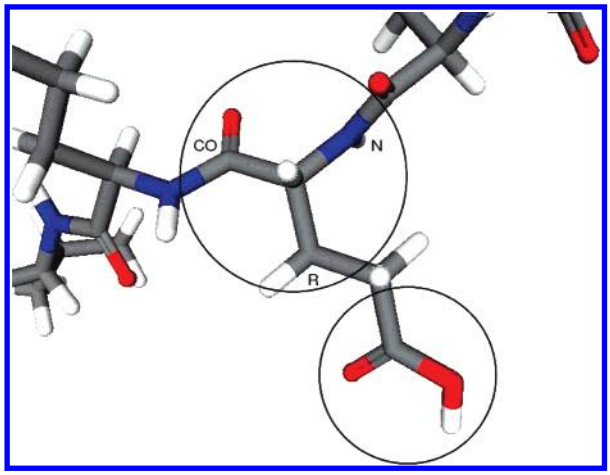


**Figure 1.** An example of a problematic ligand as contained in the scPDB (hsp90 peptide meevd, PDB 1ELR). In the center of the figure, an example of wrong chirality annotation can be seen (the marked atoms read in a clockwise direction should spell CO-R-N). In the lower-right corner, an incorrectly protonated carboxylic acid can be seen.

macologically relevant ligands containing approximately 6400 complexes crystallographically determined to a resolution better than 2.5 Å.

**Ligand Initialization and Atom-Type Identification in scPDB Ligands.** Ligands in the scPDB are given as SMILES strings, and in PDB format. The PDB format does not contain atom and bond types, which are necessary for the correct treatment of ligand flexibility and protein−ligand interactions. On the other hand, preliminary visual inspection of the ligand structures obtained from the SMILES strings identified problems with missing chiral annotation and incorrect protonation of the functional groups (see Figure 1). To obtain the correct chirality, we mapped, for each ligand, the molecular graph obtained from the SMILES string to the one obtained from the PDB file. The mapping was performed with the help of DIZZEE,[16] which is our in-house variant of the RASCAL algorithm for subgraph isomorphism.[17] The resulting mapping allows the assignment of bond types from the SMILES string to the corresponding ligand from the PDB. At the same time, the chiral structure of the PDB conformation is retained. Sybyl atom types were then assigned to the single atoms, according to neighbor element types and bond orders.

The ligand obtained in this way was then protonated according to simple chemical rules that identify the protonation state for each atom from its type and the bonds it forms. Primary and secondary amines as well as guanidines were modeled as charged groups [$CNH_3^+$, $CN(H_2^+)C$, and $RC^+(NH_2)NH_2$, respectively], single bonds between N and C = R, with R ∈ {N, O} as planar amide bonds. Nitrogens attached to these bonds were left uncharged. Carboxylate, phosphate, phosphonate, sulfate, and sulfonate groups were modeled in their fully deprotonated forms. All hydrogens attached to nonpolar atoms of proteins and ligands were discarded in this study. Protonation states of the proteins were taken unmodified from the original mol2 files found in the scPDB.

The reference conformation of the ligand was then modified with respect to orientation and internal degrees of freedom, to avoid any bias which could arise from using already correct input coordinates. To this end, all rotatable bonds of the ligand structure were set to a trans conformation at the beginning of the docking. In addition, the input structure was rotated by an angle of $\pi$ around the $z$ axis followed by a rotation of $\pi/2$ around the $y$ axis. In addition, the ligand structure was translated so that the center of the ligand coincides with the origin of the coordinate system of the PDB complex and not with the active site center.

**Energy Function.** The energy function used in GlamDock, called ChillScore, is based on the empirical scoring function ChemScore, described in refs 18 and 19, using the modified geometries as introduced in ref 14. ChillScore is the sum of a hydrogen bond term, a lipophilic term, an acceptor-metal interaction term, a ligand flexibility penalty, and a clash term for both ligand−protein and intraligand heavy atom overlap, each weighted by a corresponding parameter ($\Delta G$). In addition, the ChillScore function contains a term that penalizes poses in which the center of geometry of the ligand is outside of the defined binding pocket:

$$E_{\text{Chill}} = \Delta G_0 + \Delta G_{\text{HBond}} S_{\text{Hbond}} + \Delta G_{\text{Lipo}} S_{\text{Lipo}} + \Delta G_{\text{Metal}} S_{\text{Metal}} + \Delta G_{\text{Rot}} N_{\text{Rot}} + E_{\text{Clash}} + E_{\text{Pocket}}$$

The additive constant $\Delta G_0$ and the flexibility penalty $\Delta G_{\text{Rot}} N_{\text{Rot}}$ are relevant only in the context of binding affinity estimation and not for pose recognition. In contrast to ChemScore, no term for modeling ligand torsions is present in the ChillScore function. In order to facilitate the use of gradient-based minimization schemes, the interatomic terms in the above equation are based on a continuously differentiable rational sigmoid approximation $f_s(x_1, x_2, x)$ to the piecewise linear potential employed by ChemScore (see also Figure 2):

$$f_s(x_1, x_2, x) = \begin{cases} 1 & x < x_1 \\ s'(x') & x_1 \le x \le x_2 \\ 0 & x > x_2 \end{cases}$$

where $x'$ and $s'(x)$ are linearly transformed representations

$$x' = \left[x - \frac{1}{2}(x_1 + x_2)\right]\frac{s_2 - s_1}{x_2 - x_1}$$

$$s(x) = \frac{x}{1 + |x| + (1 - |x|)^{-1}}$$

$$s'(x) = \frac{-s(x)}{s(s_2) - s(s_1)} + \frac{1}{2}$$

of the input value $x$ and the underlying approximation functions $s(x)$. These transformations are used in order to keep the frame of reference regarding the geometric parameters and weighting coefficients identical to that of the original ChemScore function. In these transformations, $s_1 = -2 + \sqrt{2}$ and $s_2 = 2 - \sqrt{2}$ are the relevant roots of the approximated sigmoidal, $s(x)$.

The H-bond term is calculated for all donor (D)/acceptor (A) pairs between the protein and ligand and, in addition to the donor−acceptor distance, takes two types of angular constraints into account:

$$S_{\text{Hbond}} = \sum_{\text{D,A}} [f_s(\Delta r_{\text{hb1}}, \Delta r_{\text{hb2}}, \Delta r_{\text{DA}}) f_s(\Delta\alpha_1, \Delta\alpha_2, \Delta\alpha_{\text{AHD}}) \prod_R f_s(\Delta\beta_1, \Delta\beta_2, \Delta\beta_{\text{RAH}})]$$

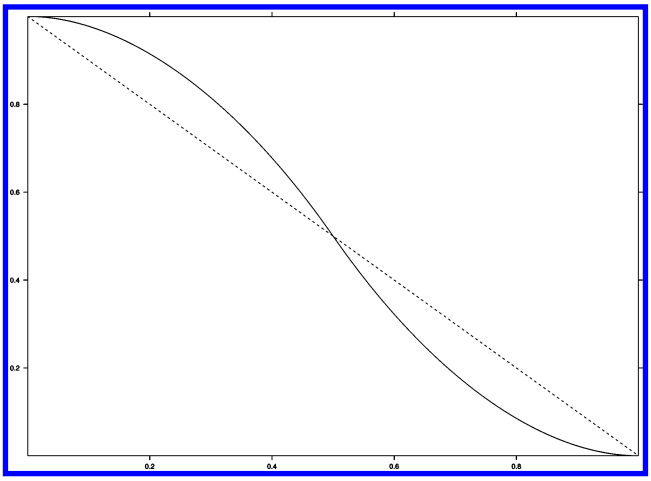

**Figure 2.** Comparison of the ramp function used in ChemScore (dashed line) and the sigmoidal function used by ChillScore (solid line).

**Table 1.** Parameters of the ChillScore Energy Function

| | | | |
|---|---|---|---|
| $\Delta G_0$ | −23.16 kcal/mol | $\Delta r_{hb1}$ | 0.25 Å |
| $\Delta G_{HBond}$ | −0.897 kcal/mol | $\Delta r_{hb2}$ | 1.15 Å |
| $\Delta G_{Lipo}$ | −0.082 kcal/mol | $r_0$ | 1.85 Å |
| $\Delta G_{Metal}$ | −2.277 kcal/mol | $r_{chb}$ | 1.60 Å |
| $\Delta G_{Rot}$ | 0.209 kcal/mol | $\Delta\alpha_1$ | 30° |
| | | $\Delta\alpha_2$ | 80° |
| $r_{l1}$ | 4.1 Å | $\alpha_0$ | 180° |
| $r_{l2}$ | 7.1 Å | $\Delta\beta_1$ | 70° |
| | | $\Delta\beta_2$ | 80° |
| $r_{m1}$ | 2.6 Å | $\beta_0$ | 180° |
| $r_{m2}$ | 3.0 Å | | |
| $r_{cm}$ | 1.4 Å | $C_m$ | 40 kcal/mol |
| $\Delta\gamma_1$ | 40° | C | 20 kcal/mol |
| $\Delta\gamma_1$ | 60° | | |
| $\gamma_0$ | 120° | | |

where $\Delta r_{hb1}$, $\Delta r_{hb2}$, $\Delta\alpha_1$, $\Delta\alpha_2$, $\Delta\beta_1$, and $\Delta\beta_2$ are the constants defined in Table 1; $\Delta r_{DA} = |r_{DA} - r_0|$ is the deviation of the donor−acceptor distance from the optimum of 1.8 Å; $\Delta\alpha_{AHD} = |\alpha_{AHD} - \alpha_0|$ is the deviation of the angle formed by donor, hydrogen, and acceptor; and $\Delta\beta_{RAH} = |\beta_{RAH} - \beta_0|$ is the deviation of the angle between any additional atoms bound to the acceptor, the acceptor itself, and the donated hydrogen atom from their optima at 180°.

The lipophilic burial term, calculated between pairs of lipohilic atoms in the receptor and ligand, is of the following form:

$$S_{Lipo} = \sum_{L,L'} f_s(r_{l1}, r_{l2}, r_{LL'})$$

where $r_{LL'}$ is the distance between the two lipophilic atoms and $r_{l1}$ and $r_{l2}$ are constants defined in Table 1. The metal−acceptor interaction term used in ChillScore differs from ChemScore by enforcing an angle constraint in addition to the distance dependency.

$$S_{Metal} = \sum_{M,A} [f_s(r_{m1}, r_{m2}, r_{MA}) \prod_{R} f_s(\Delta\gamma_1, \Delta\gamma_2, \Delta\gamma_{RAM})]$$

The clash term is composed of three differently parametrized sigmoidal functions for pairs of donor/acceptor atoms, acceptor/metal atoms, and all other heavy atom pairs.

$$E_{Clash} = C_m \sum_{M,A} f_s(0, r_{cm}, r_{MA}) + C \sum_{D,A} f_s(0, r_{chb}, r_{DA}) + C \sum_{A,B} f_s[0, r_{vdw}(A,B), r_{AB}]$$

where $r_{cm}$ and $r_{chb}$ are clash distances and $C_m$ and $C$ are clash penalties given in Table 1. The sum $\sum_{A,B}$ runs over all pairs of heavy atoms not covered by the separate metal−acceptor and donor−acceptor clash terms, and $r_{vdw}(A,B)$ is the optimal van der Waals distance of the atom pair A and B, scaled down by a factor of 0.9.

The docking-sphere-limiting term is implemented as a harmonic potential on the distance of the center of geometry of the ligand to the center of the active site. This term only contributes a penalty to the ChillScore function if this distance exceeds the docking sphere radius ($R_{Site}$ = 9 Å in this work) and is defined as follows:

$$E_{Site} = \begin{cases} [d(\bar{x} - c_{Site}) - R_{Site}]^2 & d(\bar{x} - c_{Site}) > R_{Site} \\ 0 & \text{else} \end{cases}$$

where $\bar{x}$ is the center of geometry of the ligand and $d(\bar{x} - c_{Site})$ is its Euclidian distance to the center of the binding size. We note that this term allows any ligand atom to move outside of the docking sphere as long as the center of geometry remains inside.

The $\Delta G$ coefficients of the ChillScore energy function were fit by a least-squares procedure to the binding affinities reported for 800 protein−ligand complexes in the PDBbind database release of 2004.[20] To this end, the complexes were prepared by performing a minimization using a ChillScore function using the parameters of the modified ChemScore function given in ref 14. On this training set, a correlation coefficient of $R$ = 0.53 is achieved by the ChillScore function, which is consistent with the quality reported in ref 21 for the original Chemscore function. In addition, we compared the ChillScore function to the original ChemScore described in ref 19. A total of 76 of the 112 complexes used in the publication of the original ChemScore function are contained in the PDBbind database. Fitting the parameters solely on these complexes achieves a squared correlation coefficient of $R^2$ = 0.76 with respect to the values reported in ref 19. This correlation is remarkably high, as the ChillScore function is based on the modified ChemScore function described in ref 18 and has been further modified with respect to the ramp functions, dihedral strain, and entropy loss terms used. Further factors influencing the achievable correlation may be missing water molecules and different preparation and minimization schemes used in preparing the complex structures. Verdonk et al.[14] report a correlation of $R^2$ = 0.84 for their reimplementation of ChemScore and cite a personal communication with Sandretto et al., who are reported to have achieved an $R^2$ = 0.72. These results imply that our implementation is in many ways highly similar to the original function or at least as similar as reimplementations by other groups.

**Search Space Definition.** In this work, we have assumed a rigid binding site. The conformational search space thus includes the ligand rigid body translation and rotation, and the ligand torsional degrees of freedom. So-called ring flips are currently not explored by GlamDock. All noncyclic single bonds are treated as rotatable, and each torsion angle *i* is

**Table 2.** Protomol

| protein interaction centers | probe | ligand group | $N_{local}$ |
|---|---|---|---|
| donor | C=O | acceptor | 15 |
| acceptor | NH | donor | 15 |
| lipohilic atoms, aromatic ring centers | benzene | 5−7-member cycles, lipophilic atoms | 2 |

modeled as discrete variables $\phi(i)$ which can assume 128 values representing the interval $[-\pi, \pi]$. The torsional degrees of freedom are applied to the ligand using a quaternion-based tree structure as described in ref 22.

The translational and rotational degrees of freedom of the ligand are implicitly encoded using an approach similar to the interaction mapping scheme used by Gold.[4,23] Interaction centers on the ligand are mapped to precalculated interaction sites in the protein binding site. The binding site is defined on the basis of a docking sphere of radius $R_{site} = 9$ Å centered on the geometric center of the ligand reference pose. This definition of the docking sphere is used for the harmonic out-of-pocket penalty of the ChillScore function, which penalizes structures where the ligand center of geometry was placed outside of the binding site. As part of protein preparation, a discretized model of high-scoring regions of the protein active site for three different types of probe molecules is created. This model, called protomol in analogy to refs 5 and 24, is constructed as follows: First, a grid describing partially buried volumes on the protein surface is placed into the binding site. The grid used is cubic, with a side length of $R_{site} + 6$ Å, and has a resolution of 0.5 Å. The probe radius for the identification of buried volumes (cavities) on the protein surface using an approach similar to the one desribed in ref 25 is 2.8 Å. The grid cells in buried volumes are annotated accordingly. Complementary probe "molecules" are then placed on 200 evenly distributed points on a sphere centered on all protein interaction centers (see Table 2) found in an extended binding site defined as the docking sphere using a slightly enlarged radius of $1.2R_{site}$. The radius of the spheres and the orientation of the probe molecules are chosen to be optimal according to the ChillScore energy function. For all probes close to a grid cell marked as part of a cavity, the ChillScore energy between the probe and the receptor is calculated. Probes with a non-negative energy are discarded. The remaining probes belonging to each protein interaction center are locally clustered into 60 clusters using the *K*-means algorithm, and the best scoring $N_{local}$ representatives are kept. The probe candidates of all protein interaction centers are then merged into global lists, one for lipophilic probes and one for both types of polar interactions. These lists are sorted by interaction energy, and a maximum of 35 lipophilic and 150 polar probes are kept. An example of a protomol is shown in Figure 3.

The retained probes are indexed. For each interaction group $j$ on the ligand, an integer degree of freedom $m(j)$ is generated which can take on the value of any of the indices of complementary probes in the binding site, indicating that the ligand is to be placed in a way that the corresponding interaction site overlaps with the group corresponding to that index. A value of $-1$ indicates no mapping for the particular interaction site.

All mapped ligand groups contribute with equal weight to a rigid least-squares superposition onto their assigned

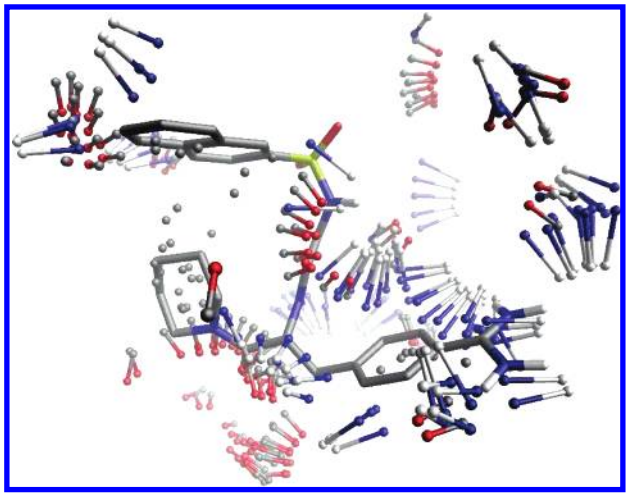

**Figure 3.** Protomol for the binding site of thrombin (thin sticks). The protomol shows good overlap with key groups of the bound ligand (NAPAP, thick sticks). The presence of probes showing no overlap with the reference ligand indicates that our protomol and active-site definitions are quite general.

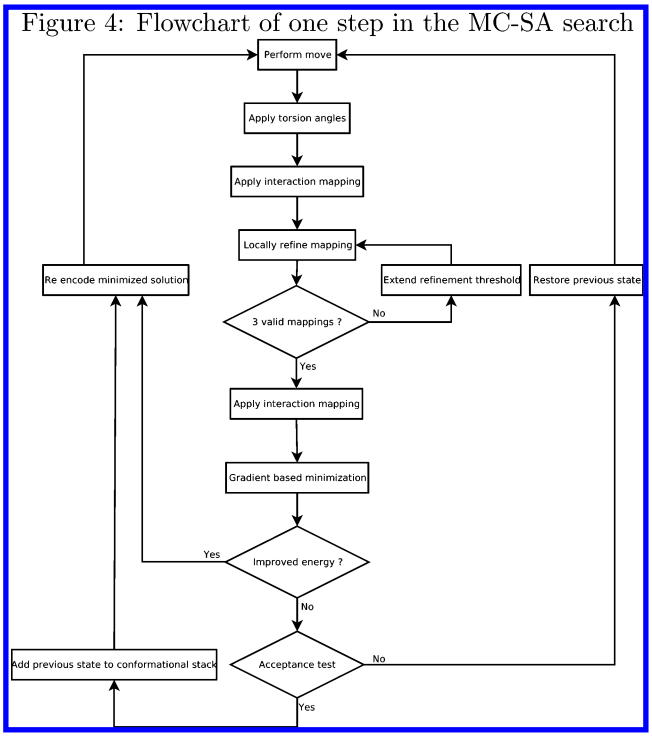


**Figure 4.** Flowchart of one step in the MC-SA search.

probes using the Kabsch algorithm.[26,27] After this initial superposition, a new mapping is formed solely on the basis of the spatial distance between ligand groups and probe molecules, ignoring any previous assignment. For each ligand group, the closest complementary probe is assigned if its distance is smaller than $(1 + 0.5k)1.5$ Å, where $k \in [1,6]$ is the current expansion cycle. Expansions are performed until at least three mappings could be assigned, as otherwise no unique solution for the Kabsch rotation exists. After the new mapping is formed, the resulting Kabsch superposition is applied to the ligand. If no valid mapping has been identified after six expansions, the ligand orientation is identical to the initialized ligand structure, which, as explained at the beginning of the methods section, is not identical to the X-ray structure.

**Table 3.** Parameter Settings

| parameter | fast screen | screen | dock |
|---|---|---|---|
| $N_{runs}$ | 2 | 3 | 5 |
| $N_{steps}$ | 200 | 275 | 300 |
| $N_{poses}$ | 30 | | |
| $N_{min,loc}$ | 8 | 10 | 15 |
| $N_{min,post}$ | 30 | 60 | 80 |
| $T_0$ | 20 | | |

**Search Strategy.** The search algorithm used in this work is based on standard Monte Carlo/simulated annealing (MC-SA),[28] augmented with local minimization[13] and a conformational stack (see Figure 4). In each step of the search, two variables are modified. For each move to be performed, either the set of conformational state variables or rigid body placement variables is chosen with equal probability. For the conformational variables, each move consists of adding a random uniformly distributed increment from the interval $[-\pi, \pi]$ discretized to steps of $1/3\pi$. Torsion angles around peptide bonds and between aromatic rings, while being sampled in steps of $\pi$, are free to assume any intermediate value in the following torsion space minimization described below.

For the move applied to the rigid body placement variable, the transition probabilities between the protomol point assigned to each ligand group are biased toward small, local exchanges. To this end, transition probabilities from a currently assigned probe $m(j) \in T$, where $T$ denotes a set of compatible probe positions for a specific ligand group, to all replacement groups $m'(j) \in T$ are defined as:

$$P[m'(j)|m(j)] = \frac{\frac{1}{d[m(j),m'(j)]} + |T|^{-1}}{\sum_{l \in T[\backslash]m(j)} \frac{1}{d[m(j),l]} + 1}$$

where $d[m(j),m'(j)]$ is the Euclidian distance between probes $m(j)$ and $m'(j)$. The transition probabilities from a mapped state to the unmapped state as well as all inverse transitions are defined as $p[-1|m(j)] = p[m'(j)| - 1] = |T|^{-1}$. All self-transitions $p[m(j)|m(j)]$ are assigned a probability of zero.

The poses generated by applying the internal degrees of freedom and the transformation derived from the interaction mapping superposition are subject to a local minimization. This minimization is performed with a simple Levenberg−Marquardt algorithm in torsion space[16] using $N_{min,loc}$ steps.[29,30]

The energy of the minimized poses is then used in the Metropolis acceptance criterion $R < e^{-\Delta E/T(i)}$, where $R \in [0,1]$ is a random number from a uniform distribution and $\Delta E$ is the difference between the energies of the last accepted and the currently proposed pose. $T(i) = 0.995^i T_0$ is the virtual temperature factor at step $i$. The conformation resulting from the local minimization is re-encoded into the state variables by directly transferring the current torsion angles. In addition, the rigid body parameters are converted into interaction probe mappings by following the same procedure described above as the second part of state variable decoding.

In addition, the last pose before an acceptance of an increase of the current energy due to the stochastic element of MC-SA is saved in a list of poses, called the conformational stack. If the length of the conformational stack exceeds a given threshold ($N_{poses}$ in Table 3), it is reduced using a pose clustering procedure. Pose clustering is implemented by sorting the conformations contained in the conformational stack in ascending order of their energies. This list is then traversed starting with the lowest-energy pose, and a new reduced list is formed by only adding poses with a root-mean-square deviation (RMSD) > 1.5 Å to all poses already part of the new list.

One docking consists of several ($N_{runs}$) independent runs of the MC-SA search. The conformational stacks of these runs are finally merged into one list which is then subjected to the pose clustering procedure described above, keeping at most $N_{poses} = 30$ poses in this study. All remaining poses are used as input for a longer gradient-based minimization to ensure convergence to their respective local minima, using $N_{min,post}$ steps. The parameter settings used in this study are given in Table 3.

**Statistical Significance of Docking Results.** In order to assess the influence of protein and ligand properties on docking success rates, we treat the success rate on subsets of a benchmark formed by a condition $X$ as the conditional probability $P(S|X)$, where $S$ is the event indicating docking successes (RMSD < 2.0 Å). The significance of the difference between a subset's estimated probability of success $P(S|X)$ and the overall probability of success $P(S)$ can be assessed by Z scores on the basis of the expected distribution of success probabilities when sampling random subsets of size $n$ from a complete benchmark. The Z scores are calculated as follows:

$$Z[P(S|X),n] = \frac{P(S|X) - P(S)}{\mathrm{dev}(n)}$$

where dev($n$) is the expected sample deviation of the percentage of successes when drawing $n$ samples from a binomial distribution.

$$\mathrm{dev}(n) = \sqrt{\frac{P(S)[1 - P(S)]}{n}}$$

**Identification of Dependencies between Descriptor Effects.** The descriptors used to form subsets may be dependent on each other. Such dependencies between two descriptors $X$ and $Y$ may obscure the true effect of a descriptor; therefore, we estimate the nonadditivity of the effects of descriptor pairs by calculating the Kullback−Leibler divergence[31] between the observed joint distribution $P(S,X,Y) = P(S|X,Y)\ P(X,Y)$ and a model distribution assuming independence of the descriptors' effects. In order to derive the independency model distribution, we write $P(S|X,Y)$ as

$$P(S|X,Y) = \frac{P(S|X)\ P(S|Y)}{P(S)} \cdot \frac{P(X,Y|S)}{P(X|S)\ P(Y|S)} \cdot \frac{P(X)\ P(Y)}{P(X,Y)}$$

Assuming that conditioning on docking success does not influence the statistical dependence of the descriptors $X$ and $Y$:

$$\frac{P(X,Y|S)}{P(X|S)\ P(Y|S)} \cdot \frac{P(X)\ P(Y)}{P(X,Y)} = 1$$

**Table 4.** Results for FlexX, Gold, Surflex, Glide, and Dock Read from Figure 2 from the Kellenberger Study[10] [a]

| setting | top < 1.5 Å | top < 2.0 Å | any < 1.5 Å | any < 2.0 Å | runtime [s] |
|---|---|---|---|---|---|
| GlamDock | | | | | |
| fast screen | 44% (3%) | 51% (2%) | 63% (%3) | 73% (3%) | 11 (68) |
| screen | 52% (2%) | 59% (2%) | 72% (%3) | 81% (3%) | 21(130) |
| docking | 55% (2%) | 62% (1%) | 79% (%3) | 85% (2%) | 38 (235) |
| Flexx | 43% | 51% | 62% | 66% | (67) |
| Gold | 51% | 57% | 78% | 82% | (137) |
| Surflex | 45% | 56% | 69% | 78% | (135) |
| Glide | 41% | 54% | 78% | 85% | (234) |
| Dock | 34% | 40% | 45% | 54% | (46) |

[a] Standard deviations over 10 repeats are given in parentheses for the GlamDock results.

Using this assumption, we define the model distribution as

$$P'(S,X,Y) = P'(S|X,Y)\,P(X,Y) = \frac{M(S,X,Y)}{\sum_{s\in S,x\in X,y\in Y} M(s,x,y)}$$

with

$$M(S,X,Y) = \frac{P(S|X)\,P(S|Y)}{P(S)}P(X,Y)$$

The Kullback−Leibler divergence KL($P$,$P'$) is a fundamental information theoretic measure of the difference between two distributions. Higher values of the Kullback−Leibler divergence between $P$ and $P'$ indicate stronger dependence effects of the descriptors $X$ and $Y$ on the probability of success. We calculate the Kullback−Leibler divergence—measured in bits—as

$$\mathrm{KL}(P,P') = \sum_{s\in S,x\in X,y\in Y} P(s,x,y)\log_2\frac{P(s,x,y)}{P'(s,x,y)}$$

## RESULTS AND DISCUSSION

**Kellenberger.** The results for redocking on the Kellenberger data set are shown in Table 4. In the publication by Kellenberger et al.,[10] eight different docking tools (Dock, FlexX, FRED, Glide, Gold, Slide, Surflex, and QXP) have been compared against each other with respect to redocking accuracy. The docking tools were run with standard parameters as suggested by the developers and took between 18 and 240 s on an SGI (MIPS R12000) CPU. The different docking algorithms were compared with respect to sampling and scoring accuracy. Sampling quality was measured as the percentage of cases for which there was at least one good pose among all suggested solutions. At most, $N_{\mathrm{poses}} = 30$ candidate solutions were used in this study. Good poses were defined as those that show a RMSD of the ligand heavy atoms below 2 Å to the crystal structure. Scoring accuracy was defined as the percentage of cases where the top-ranked pose was a good pose. The best performing dockers (FlexX, Glide, Gold, and Surflex) had a scoring accuracy of 51−55%.[10] Further, there seemed to be a clear tradeoff between efficiency and accuracy, as the best dockers were also, in general, slower.

GlamDock with its slowest settings leads to very good sampling already at the 1.5 Å level (where good structures are defined as structures that show a RMSD of up to 1.5 Å to the crystal structure of the complex) and even better at the 2.0 Å level (79% and 85%, respectively). This is again seen at the scoring accuracy, where GlamDock finds a structure with a good pose at the first rank in 55% or 62% of cases at the 1.5 and 2.0 Å level, respectively.

For our calculations, a 2.8 GHz Xeon CPU was used. In order to compare efficiencies quantitatively, we estimated the relative efficiency of the CPU used in the Kellenberger study and the 2.8 GHz Xeon used here. For the estimation, we used the Surflex program as a reference, as it was part of the Kellenberger study and timings of its efficiency on a CPU identical to the one used here are available in ref 32. For Surflex, we have the time requirements under identical settings, once given by Jain on a 2.8 GHz Xeon CPU (as 3 s per free torsion in the ligand) and once given for the SGI CPU used by Kellenberger et al., who report 135 s per run on average. As the ligands in the Kellenberger data set contain on average 7.3 free torsion angles, we calculate an average of 22 s per ligand on the Xeon, which gives a relative factor of approximately 6 between the two architectures. Taking the factor of 6 into account, the slowest settings for GlamDock ("docking" in Table 4) correspond to the slowest docker in the Kellenberger study, namely, Glide, which took on average 240 s for a single docking. In a recent paper, Jain suggested that the Surflex version used in the Kellenberger study had not been fully optimized for speed, so that the ratio derived here may in reality be even lower, making the GlamDock settings appear slower in the comparison. Nevertheless, the approximate magnitude of the times are expected to be correct.

From Table 4, it is seen that in the fastest setting ("fast screening") GlamDock can compete with programs known for their good runtime efficiency, namely, Dock and FlexX. At the medium setting ("screening"), GlamDock is comparable to the best established docking tools, with respect to sampling and scoring accuracy. In the slowest setting ("docking"), which corresponds to the efficiency of GLIDE in the comparison by Kellenberger, the structure at the best rank is, in 62% of the cases, within 2 Å from the crystal structure, which corresponds to a 5% improvement over the best docking tool (Gold) in the study by Kellenberger et al. However, due to the small size of this benchmark, this result is statistically not highly significant with a Z score of 1.03 assuming that 62% represents the true probability of success for GlamDock (see the methods section). In other words, this difference is approximately one standard deviation. Interestingly enough the efficiency-to-accuracy tradeoff observed among the different dockers can be reproduced with GlamDock by these different settings for sampling (see Table 4, and see Figure 5).

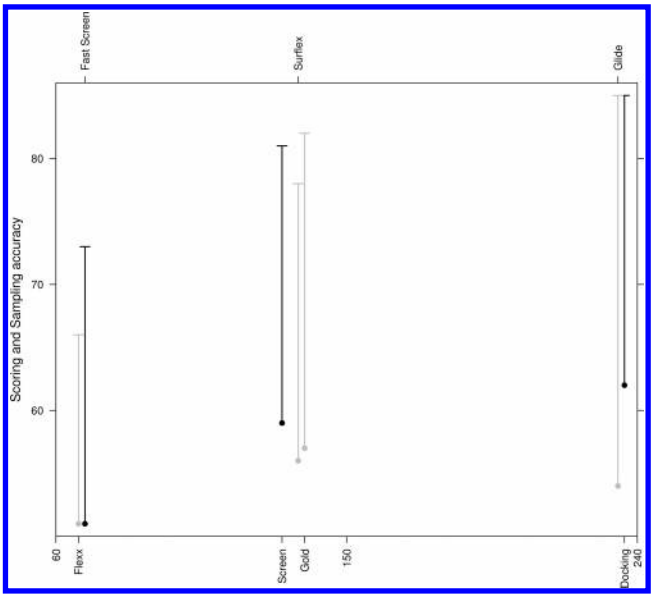


**Figure 5.** Comparsion of scoring and sampling quality of the best tools studied (gray lines)[10] and three parameter settings for GlamDock (black lines). The average runtime per complex is shown on the horizontal axis (in seconds). The vertical axis shows the percentage of complexes with the RMSD of the top-ranked pose below 2 Å (filled circles) and with any pose below 2 Å (horizontal bars).

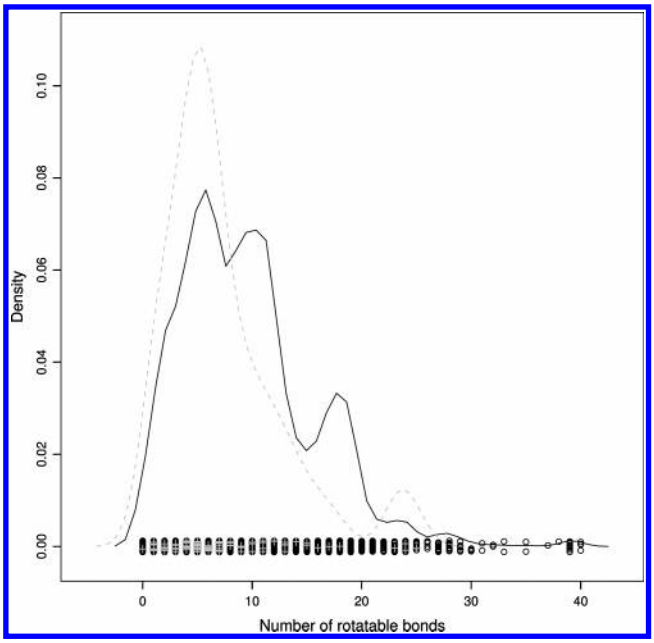


**Figure 6.** Distribution of the number of rotatable bonds in the Kellenberger benchmark (gray dashed curve) and the SC-PDB data set (black curve).

**Table 5.** Scoring Errors (as Percent of Benchmark Size)

| | Kellenberger $n = 100$ | SCPDB $n = 5681$ |
|---|---|---|
| errors | 38% | 53% |
| scoring errors | 26% | 36% |
| sampling errors | 12% | 17% |
| avg. ref RMSD | 0.5 Å | 0.5 Å |

In comparison to the results reported in the original publication, it is important to note that the results presented here are obtained by the developers (the authors) while the results reported in the Kellenberger study are independent results obtained by the same users for the different docking programs, and by setting the parameters according to standard protocols suggested by the developers. Nevertheless, while a true comparative evaluation of GlamDock against other dockers can only be achieved by such independent comparisons, this is the best type of data available to us for validating GlamDock at the present time, and from the current results, it appears that GlamDock is of at least comparable accuracy and efficiency to the best existing docking software.

**SC-PDB.** Of the 6415 complexes in the scPDB, 6130 were present in both the complexed and protein-only files downloaded from the scPDB Web site. Of these, 5681 complexes were succesfully parsed and protonated by our system. In 417 cases, the SMILES strings provided by the scPDB could not be exhaustively assigned to the ligand molecules; the remaining errors are caused by unsupported elements, mainly cobalt. For the benchmark, we used the "docking" settings from the previous section, as they are still fast enough to allow large-scale screening and appear to be at least as accurate as standard docking protocols. On the scPDB, docking with these settings required approximately 46 s per complex. We performed a single docking for every complex, as the overall statistics are significant due to the size of the benchmark set.

In the results, we find that with respect to both sampling and scoring the accuracy in the scPDB benchmark is significantly lower than that for the Kellenberger data set: the former decreases from 85% to 77%, while the latter decreases from 62% to 47%.

In order to better discriminate scoring from sampling problems, we refine the definition of scoring errors as those unsuccesful dockings (RMSD of the top ranked pose > 2.0 Å) where the score of the highest ranked conformation is better than that of the crystal conformation. For this comparison, the crystal structure is minimized in torsion space for 80 steps to find the next local optimum of our scoring function. This definition allows a clearer attribution of the cause of errors.

In Table 5, the analysis of the results based on this definition of scoring error is shown. The comparison to the results on the benchmark by Kellenberger et al. will be based on one arbitrarily selected run out of the 10 repeats in the previous section, which in this section is evaluated according to the refined definiton of scoring errors. For the Kellenberger benchmark and the scPDB data set, the ratio of unsuccessful dockings (RMSD of best rank) was 38% and 53%. In respectively 33% and 46% of the complexes, the docking finds a structure with a RMSD > 2.0 Å and a better score than the crystal structure. When the comparison is performed with the score of the minimized crystal structure, the number of scoring errors decreases to 26% and 36% for the two benchmarks. This leaves 12% and 17%, which fail due to poor sampling. The higher incidence of sampling errors in the scPDB set could in principle be due to the difference in the composition of the data sets. As is shown in Figure 6, the scPDB contains more flexible ligands. Higher flexibility as measured by the number of free torsions makes sampling more difficult, as the size of the sampling space increases with each degree of freedom. The influence of this and other descriptors on docking accuracy will be examined in detail in the next section.

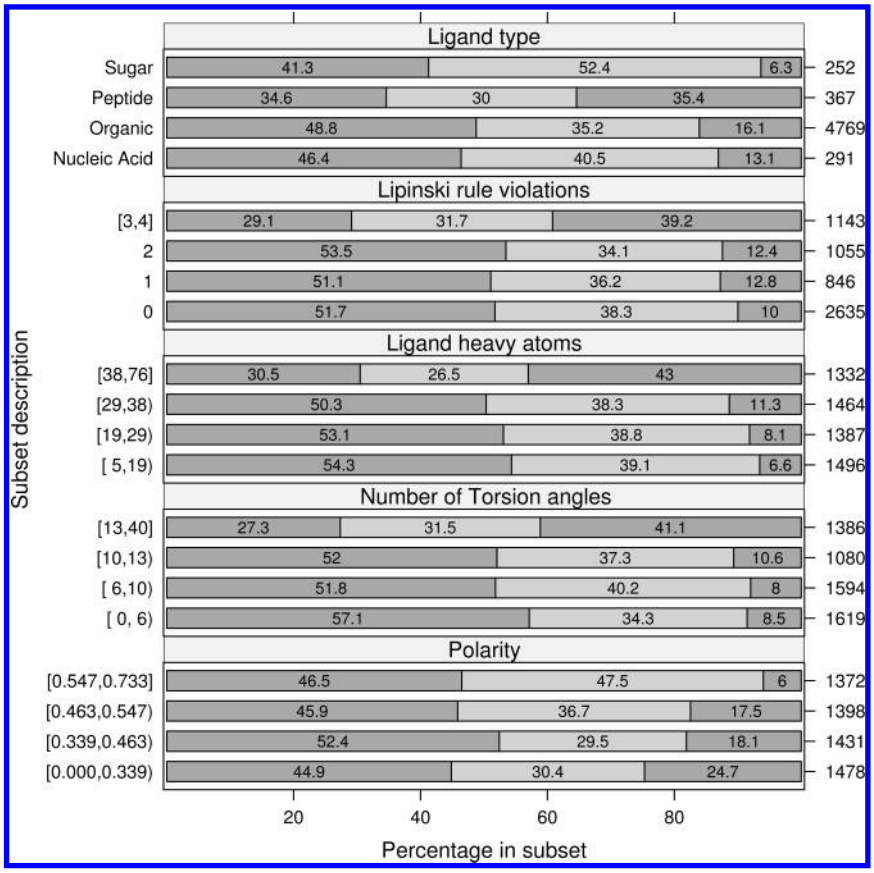


**Figure 7.** Average success rates (top-scored RMSD < 2.0 Å, left dark bar), scoring errors (middle light bar), and sampling errors (right dark bar) for several subsets of the SC-PDB data set. The left vertical axis shows the values or intervals used to select the subsets; the right vertical axis shows the number of instances in each subset.

**Influence of Ligand Properties on Docking Quality.** To further test the effect of several descriptors on the quality of docking solutions generated by GlamDock, we calculated the scoring and sampling error statistics for several subsets of the scPDB. We only performed this analysis on the scPDB, as the number of complexes in the Kellenberger benchmark is too small for such an analysis, as the expected standard deviation of the success probability for the complete Kellenberger benchmark is already 5%. The results are shown in Figure 7. Differences in the statistics of different subsets indicate an influence of the subset-determining descriptor on docking quality. The results in this section are all statistically significant as measured by Z scores unless noted otherwise.

The number of rotatable bonds directly affects the dimensionality of the search space of the docking problem and as such is expected to strongly affect the sampling performance of docking tools. As expected, the impact of increased flexibility on the percentage of sampling errors is strongly pronounced, with approximately 10% of the ligands with less than 13 rotatable bonds showing sampling errors, while sampling fails for more than 40% of ligands that have 13 or more rotatable bonds. The overall success rate is above average for up to 12 rotatable bonds. Figure 8 shows an example of a highly flexible ligand (FAD in a mutant of NADP+ reductase, PDB 1QH0) that was not sampled correctly. Shown is the best solution found, with a RMSD of 2.42 Å, close to our threshold of 2 Å. However, the energy of the top solution (RMSD 3.2 Å) was −20.5 kcal/mol, which is higher than the fitness of the minimized reference structure at −21.6 kcal/mol.

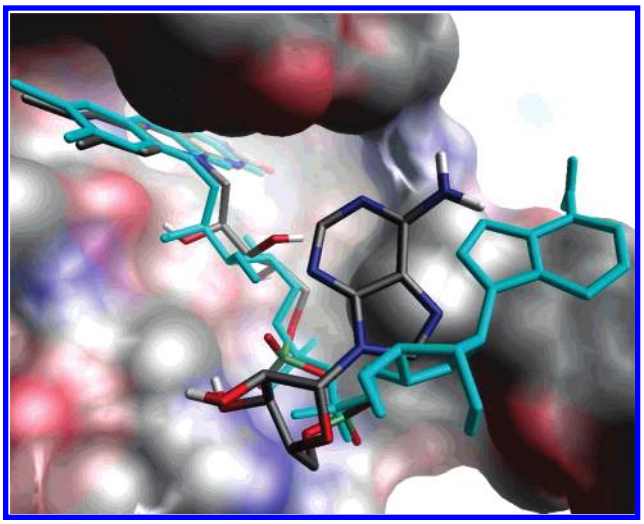

**Figure 8.** Examples of a sampling error caused by a highly flexible ligand (FAD in a mutant of NADP+ reductase, PDB 1QH0). The reference structure (light blue sticks) scores better than the top-ranked solution (atom-colored sticks) found and was not sampled during docking.

The number of ligand atoms has an impact on docking quality similar to that for the number of torsion angles. Docking quality varies from 54% to 30% for the smallest and the largest subset, respectively. The reduced accuracy on larger ligands is caused by an increase in the percentage of sampling errrors.

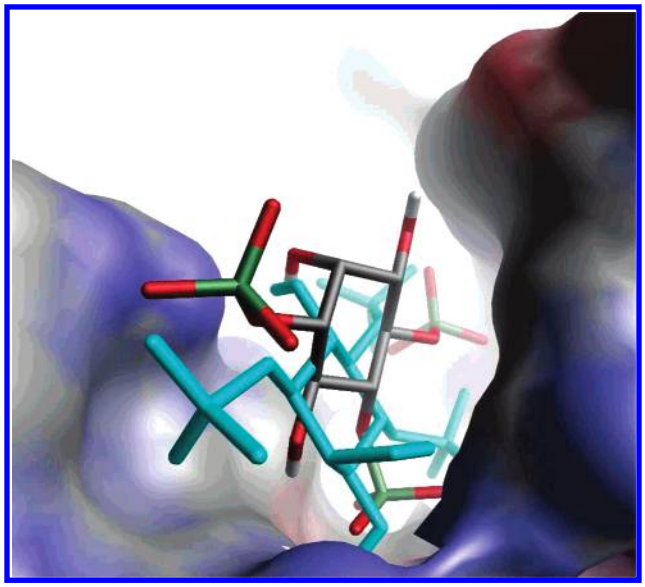

**Figure 9.** Example of a scoring error caused by a highly polar ligand (IP3 into the inositol 1,4,5-trisphosphate receptor; PDB 1N4K). Even though a solution close to the reference structure (light blue sticks) was found, a pose with a RMSD of 3.8 Å (atom-colored sticks) was ranked as the best scoring pose.

The relative dependence of successful and incorrectly scored or sampled complexes on ligand polarity is shown in Figure 7. Polarity is simply measured as the number of H-bond donors and acceptors divided by the number of ligand heavy atoms. Interestingly, the most hydrophobic subset (polarity < 33.9%) shows an increased occurrence of sampling errors. As ligand polarity increases, sampling errors become rarer. However, this increase in sampling quality is barely reflected in an increased success rate, except for the group of ligands in the polarity interval between 34% and 46%. Rather, the percentage of scoring errors increases, indicating that our scoring function is not able to accurately detect the correct pose for highly polar ligands forming a large number of putative H bonds. An example of such a ligand is shown in Figure 9. Shown are the results of docking IP3 into the inositol 1,4,5-trisphosphate receptor (PDB 1N4K). While a pose with very good agreement with the crystal structure (RMSD = 1.4 Å) was identified, a structurally incorrect solution with a RMSD of 3.8 Å was ranked at the top position with an energy of −11 kcal/mol as compared to the energy of the minimized crystal structure of −9.8 kcal/mol. As most of the best docking tools use similar approaches for the placement of ligands (hydrogen-bond mapping in Gold,[14] interaction surfaces in FlexX,[3] protomol in Surflex[5]), we expect that a dependence of sampling accuracy on ligand hydrophobicity is a general phenomenon and not a peculiarity of GlamDock.

We also investigated the influence of drug-likeness, as described by the Lipinski rules, on docking quality. Docking quality is stable and slightly above average (success rates > 51%) for ligands with up to two Lipinski rule violations. Ligands with more than three violations, however, are likely to show incorrect binding mode predictions with a success rate of approximately 29%. This decrease in quality is caused by a high incidence of sampling errors (39%).

A different view on the distribution of errors is afforded by the ligand-type annotations in the scPDB, which classify the ligands as nucleic acids, peptides, sugars, lipids, or general organic compounds. The lipid category was excluded from this analysis as only four lipids were found in our data set. Perhaps the most striking result is the extremely high incidence of scoring errors among the ligands in the sugar category, where only 6% fail due to sampling errors while 52% percent of the dockings fail due to scoring errors. This effect is most probably closely tied to the high number of H-bond donors and acceptors in carbohydrates, indicated by an average polarity of 61% in the sugar subset versus 40% over all other ligands and the high self-similarities (approximate symmetries). Self-similarity can lead to poses that deviate significantly from the crystal structure on a RMSD basis, showing however very similar interactions with the crystal structure.

The ligands classified as peptides show the opposite behavior, with a high incidence of sampling errors (35%) leading to a very low succes rate of 35%. This result is most likely caused by the high number of rotatable bonds (average 18.3) in the peptide subset when compared to all other compounds (average 8.8).

**Influence of Protein Properties on Docking Quality.** In addition to properties of the complexed ligand, the protein binding sites themselves carry information relevant to the expected docking success rate. Of these, the resolution of the X-ray structure has often been used as a criterion to exclude structures from docking benchmarks and was also analyzed here. We do observe a drop in success rates from 52% to 43% driven by a loss of scoring accurarcy, where the best performing subset contains all structures with a resolution of better than 1.81 Å, and the worst subset consists of structures with a resolution of 2.21 Å and above. Interestingly, the average active site B-factor, as annotated in the scPDB database, displays a better discriminative power than the overall resolution of a PDB entry. Subsetting on the active site B-factor shows a drop from 53% to 39% in success rates, based on the difference of the two marginal subsets with B-factors below 14.3 and above 27.25 (see Figure 10).

The correct parametrization of metal interactions in protein−ligand docking is known to be difficult. Interestingly, the success rate of 46% on active sites containing metal ions is only slightly and insignificantly lower than that of 48% on active sites that do not. The lower accuracy is caused by an increase in scoring errors (from 35% to 41%), while the incidence of sampling errors decreases from 18% to 13%, and further analysis suggests that the metal interaction weight derived from fitting on binding affinities is too strong for accurate pose recognition: for the subset of metal-ion-containing sites (18% of the complete data set) whose reference structures do not indicate any ligand−metal interaction (7% of the complete data set), the percentage of scoring errors increases to 54%, yielding a succress rate of only 36%. An example of such a case (PDB 1KTG) is shown in Figure 11. The crystal structure of a complex between the diadenosine tetraphosphate hydrolase of *C. elegans* and AMP shows no close interactions between the ligand and the metal ion coordinated by the protein. The top scoring solution by GlamDock (RMSD 4.8 Å) however shows acceptor−metal contacts and is scored with −7.2 kcal/mol as compared to −5.2 kcal/mol for the minimized reference structure. The success rate for metal-containing binding sites that do show interactions between metal ion and ligand in the reference structure of 53% is higher than average.

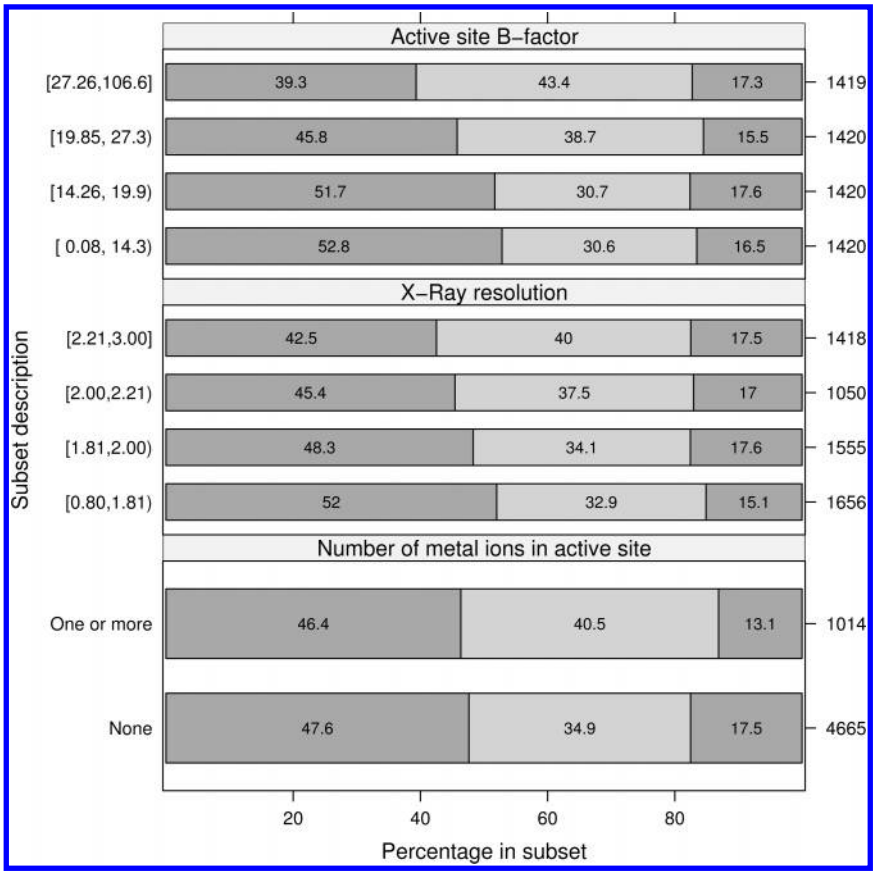


**Figure 10.** Average success rates (top-scored RMSD < 2.0 Å, left dark bar), scoring errors (middle light bar), and sampling errors (right dark bar) for several subsets of the SC-PDB data set. The left vertical axis shows the values or intervals used to select the subsets; the right vertical axis shows the number of instances in each subset.

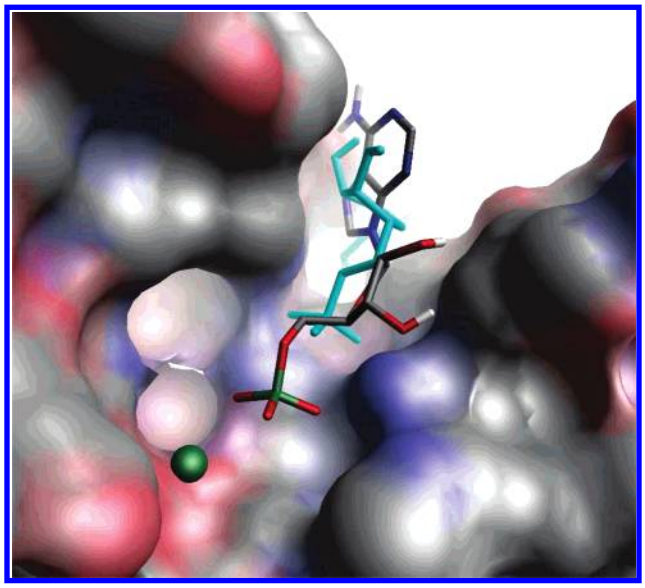

**Figure 11.** Example of a scoring error caused by an overestimation of ligand−metal ion interactions (AMP in the diadenosine tetraphosphate hydrolase of *C. elegans* and AMP; PDB 1KTG). The reference structure (thin blue sticks) shows ligand−metal ion interactions. ChilScore however chooses a top-ranked pose, forming interactions with the metal ion (atom-colored sticks).

**Influence of Scoring Function Characteristics on Docking Quality.** Table 6 shows the results of a subset analysis based on properties derived from the ChillScore scoring function. The first analysis shown in this figure separates the complexes on the basis of the reference ligand RMSD during the minimization of the crystal structure. This

**Table 6.** Docking Quality and Error Distribution on the High-Quality Subset and Using a Reduced Binding Site Definition

| | success | scoring errors | sampling errors | *n* |
|---|---|---|---|---|
| complete | 47.4% | 35.8% | 16.7% | 5681 |
| complete (6 Å) | 53.9% | 30.5% | 15.5% | |
| HQ subset | 59.6% | 31.4% | 9.0% | 2161 |
| HQ subset (6 Å) | 64.3% | 28.7% | 6.9% | |

descriptor indicates how close the next local optimum identified by the ChillScore function is to the true structure. Interestingly, this local accuracy translates well to the accuracy of the global optimum idendtified by ChillScore, as indicated by an increase in the percentage of scoring errors from 25% to 50% when comparing the subsets with a RMSD of less than 0.3 Å to those with a RMSD above 0.5 Å.

When grouped by the sum of all bump terms for the top ranked pose, the intuition that successful dockings rarely show protein−ligand overlap is strongly supported. Success rates drop from 60% for nearly overlap-free poses to 31% for the group showing the strongest overlap (Figure 12).

The saturations of polar and lipohilic groups, here defined as the sum of the unweighted H-bond terms divided by the number of acceptors and donors and as the sum of the unweighted lipophilic terms divided by the number of lipophilic atoms, respectively, are highly powerful indicators of accurate dockings. Top scoring poses with the highest saturation of lipohilic interactions are within 2 Å in 60% of the cases. Success rates drop continuously, reaching 34% for those poses in the lowest quantile shown. This effect is even more strongly pronounced for polar interactions, where

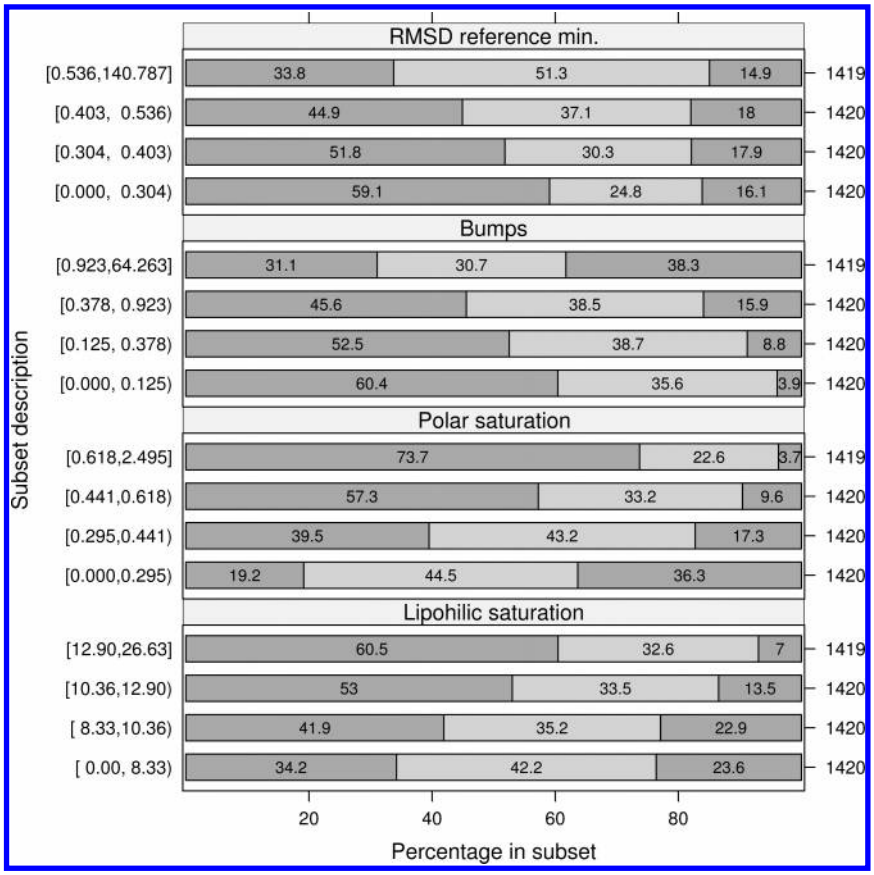


**Figure 12.** Average success rates (top-scored RMSD < 2.0 Å, left dark bar), scoring errors (middle light bar), and sampling errors (right dark bar) for several subsets of the SC-PDB data set. The left vertical axis shows the values or intervals used to select the subsets; the right vertical axis shows the number of instances in each subset.

success rates reach 73% for the group with the highest saturation and only 19% for poses with the lowest number of possible hydrogen bonds formed.

**Second-Order Analysis of Properties.** Due to strong dependencies between several of the descriptors discussed in the previous section, results based on isolated analysis may yield misleading results. We investigated dependencies among the different system features and their combined effect on success rates with the help of a two-dimensional analysis of the results. In order to identify pairs of descriptors that show interesting, nonadditive behavior if analyzed in combination, we calculated the Kullback−Leibler divergence of the observed joint distribution of the probability of successful dockings and two descriptors and a model distribution based on the assumption of independence (see methods section). Figure 13 shows the results of this analysis with darker areas in the plot corresponding to a higher Kullback−Leibler divergence between the true distribution and the independency model.

Figure 14 shows a 2D histogram with bins formed by the quantiles of ligand torsions (horizontal axis) and ligand atoms (vertical axis). This combination of descriptors is assigned a Kullback−Leibler divergence of 0.022 bits, indicating strong nonadditive effects. The area of each square in this plot is proportional to the number of complexes belonging to the combination of intervals shown on both axes. Squares containing fewer than 50 instances are not shown. The plotted numbers show the success rate over the subset of complexes represented by each respective square region. Red squares show an increase in success rates when compared to the average over the complete scPDB (47%). Blue is used to mark areas of decreased success rates. The saturation of the red and blue colors used to fill the squares is based on the statistical significance as measured by Z scores, with higher saturation indicating stronger significance (see methods section).

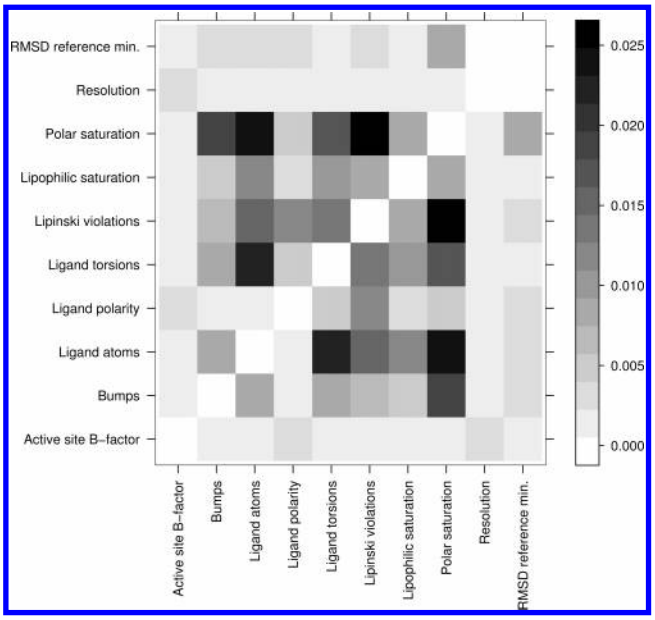


**Figure 13.** Matrix plot of the Kullback−Leibler divergence between the observed distribution and a model assuming independence between desciptor effects for all descriptor pairs. The shades of gray in the off-diagonal elements encode the KL divergence (in bits), with lighter shades corresponding to stronger independence.

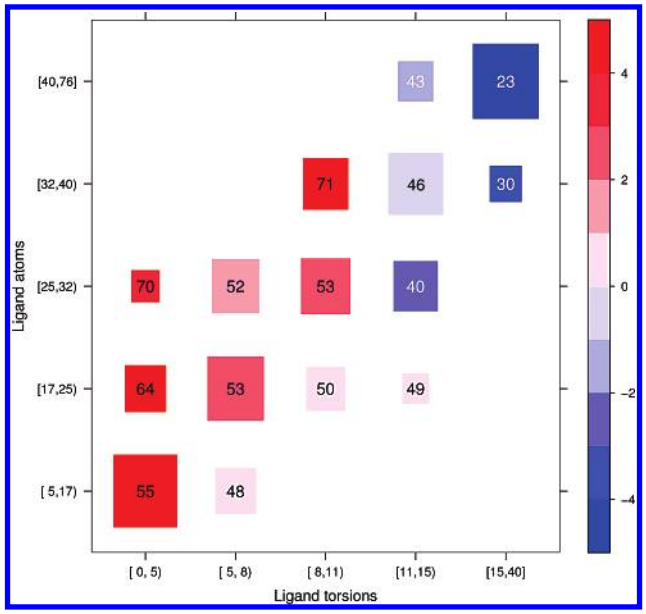


**Figure 14.** 2D histogram of ligand rotatable bonds and ligand atoms (KL divergence = 0.022 bits). The area of the squares is proportional to the number of complexes falling into the intervals listed on the horizontal and vertical axes. The numbers plotted in each square show the success rate in the subset represented by each square region. Red and blue squares indicate above- and below-average success rates, respectively. The saturation of the colors used represents the statistical significance of each area's performance as modeled by Z scores. The saturations are mapped to Z scores as indicated by the color key on the left of the plot.

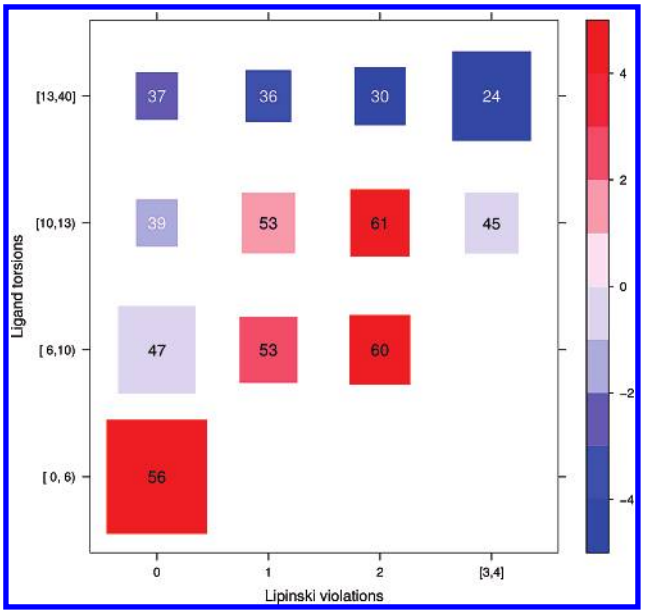


**Figure 15.** 2D histogram of Lipinski rule violations and ligand torsions (KL divergence = 0.014 bits). See the caption of Figure 14 for an explanation of this plot type.

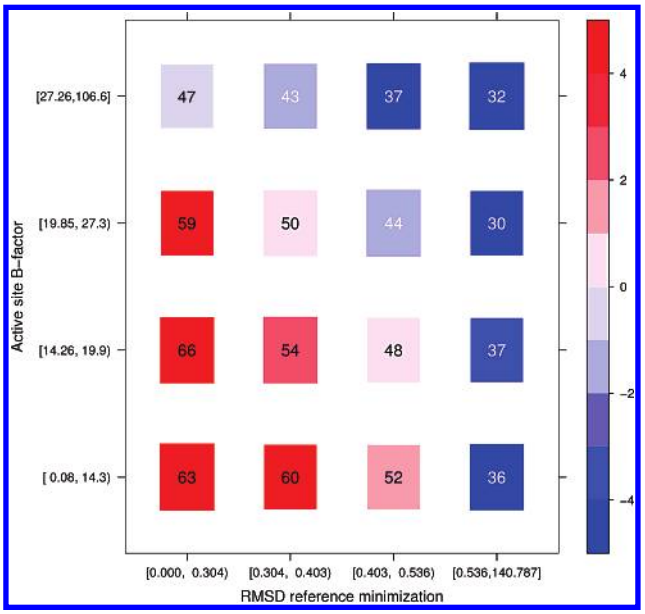


**Figure 16.** 2D histogram of reference minimization RMSD and the active site B factor (KL divergence = 0.002 bits). See the caption of Figure 14 for an explanation of this plot type.

The large size of the squares on the diagonal and the absence of data in the upper-left and lower-right corners indicate the strong linear correlation between the number of rotatable bonds and the number of atoms. When the plot in Figure 14 is read from from left to right, the expected decrease in docking quality due to the increase in the number of rotatable bonds is directly reflected by continuously decreasing success rates and the change of predominately red to blue squares in the two right-most columns. However, no such detrimental effect can be observed when reading the columns in an upward direction, from low to high atom counts. Rather, at least for the columns in the left part of the plot, increasing the ligand size tends to improve docking accuracy. This stands in stark contrast to the result from the one-dimensional analysis in Figure 7. There, the dependence of the success rate on ligand size suggests a significant negative effect of the latter on the docking quality.

The one-dimensional analysis indicated a stong negative effect of more than two Lipinski rule violations on the docking quality. Figure 15 shows a 2D histogram of the number of Lipinski rule violations and the number of rotatable bonds. A relatively high degree of nonadditive interactions is indicated by a KL divergence of 0.014 bits. The distribution of the square sizes indicates no linear correlation between the two descriptors. Rather, the number of Lipinski violations is shown to be a lower bound on the number of rotatable bonds. For ligands with less than 13 rotatable bonds, increasing the number of Lipinski rule violations up to two has a more pronounced beneficial effect on the docking quality compared to the analysis stratified by Lipinski rules alone. Only those ligands with more than two rule violations combined with high flexibility in the upper-right corner are responsible for the low success rates attributed to the rule violations alone in the one-dimensional analysis.

Figure 16 shows the 2D histogram combining the RMSD of the X-ray ligand structure during minimization with the ChillScore function and the average active site B-factor. The uniform distribution of the square sizes indicates a surprising independence between the minimization RMSD and the structure quality as described by the B-factor. In addition, the detrimental effects of both desciptors on docking appear additive. This interesting independence of the effects of structure quality and the RMSD upon minimization is supported by the low KL divergence of 0.002 bits.

Figure 17, showing a histogram of polar and lipophilic saturation, defined as in the one-dimensional analysis, gives another example of two relatively independent descriptors. The positive effect of both descriptors on docking quality is

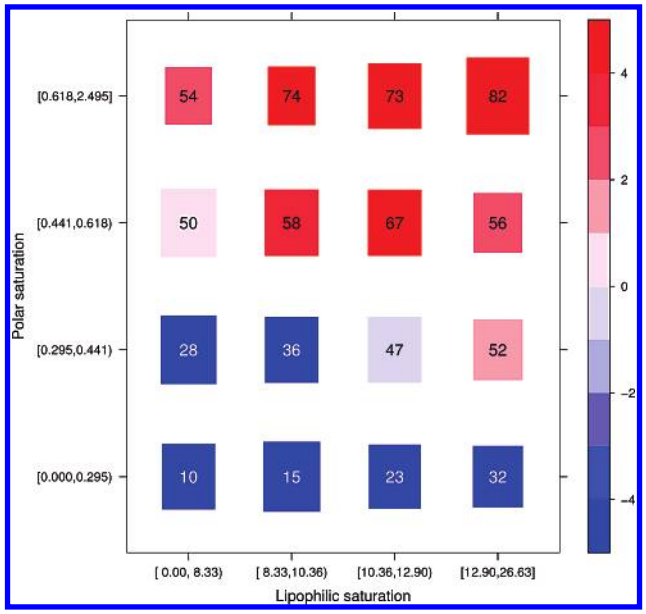


**Figure 17.** 2D histogram of polar and lipophilic saturation (KL divergence = 0.008 bits). See the caption of Figure 14 for an explanation of this plot type.

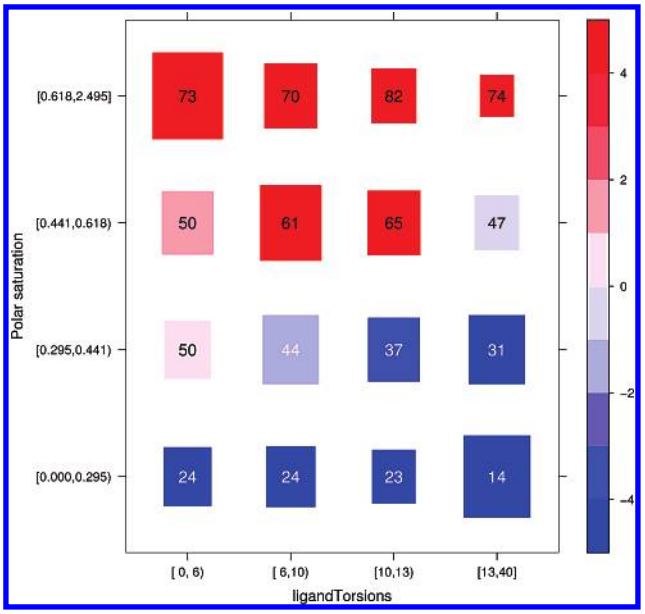


**Figure 18.** 2D histogram of the number of rotatable bonds and polar saturation (KL divergence = 0.017). See the caption of Figure 14 for an explanation of this plot type.

almost additive (KL divergence = 0.008 bits), with higher saturation in both dimensions leading to better accuracy. The prevalence of red squares, showing improved docking quality, in the upper-left corner and blue squares in the lower-right corner shows that high polar saturation alone indicates good dockings while the lipohilic saturation is only informative if it co-occurs together with sufficient polar saturation. For this plot, we explicitly chose the polar and lipophilic saturation of the ligand in the best ranked pose, as these are measures available also in the real-life scenario and allow an assessment of docking confidence in real applications. It is in principle possible that the dependency of accuracy on interaction saturation is an indirect effect of poor sampling: Whenever the best ranked pose is not saturated with respect to its interaction potential, the structure prediction is poor. This would leave the possibility open that all or at least most of the ligands found in the scPDB are saturated in the native structure and that the docker simply does not always find this perfect pose. To eliminate this possibility, we produced the same plot with the crystal structures of the complexes. The result dependencies are very similar (not shown). Thus, we conclude that ligands, forming only few interactions with the protein (compared to what they theoretically could), do exist (in high abundance) and are more difficult to dock. This of course is not a surprising result. Ligands binding in shallow pockets or having a significant part of their structure solvated are known to be hard cases for docking.

Several other descriptors show high KL-divergence values when paired against the polar saturation, namely, the numbers of Lipinski rule violations (0.025 bits), ligand rotatable bonds (0.017 bits), and ligand atoms (0.025 bits). Figure 18 shows the plot for polar saturation and ligand torsions; the other histograms for these pairs show a similar pattern (not shown). The size distribution in this plot indicates that highly saturated poses are rarely identified for highly flexible ligands while they are commonly found for rigid ligands, as seen by the smaller squares in the lower-left and upper-right corners. Interestingly, an increase in the number of rotatable bonds in the highly saturated upper two bands is associated with an unexpected stability of the probability of success. These results suggest that identifying a pose with highly saturated polar interactions for a flexible ligand suggests that sampling was successful and as such has a higher probability of success than expected for highly flexible dockings in general. A similar effect is seen in the histogram of polar saturation against the sum of the bump term, where an increase in the number of protein−ligand clashes in the highly saturated bands is not associated with a drop in docking quality (KL divergence = 0.019 bits, not shown).

**A "Sweet Spot" Benchmark for Docking.** In order to emphasize the scope of the validation performed in this work, we constructed a reasonable subset of the scPDB based on rules that, in similar form, have been employed in other docking validation studies. On the basis of the analysis shown above, we only included complexes whose ligands are druglike (less than 13 rotatable bonds) and whose active site B-factor, as an indicator of structure quality, is lower than 20. The resulting benchmark contains 2161 complexes and is docked by GlamDock with a success rate of 60% (31% scoring and 9% sampling errors), despite its fully automatic preparation.

Several docking evaluations (for example refs 5 and 14) have employed rather stringent definitions of the binding site. In order to investigate the effect of such a definition on docking quality, we repeated the scPDB docking experiment on the basis of an active site consisting of only those atoms within 6 Å of any ligand heavy atom. This definition results in an improved success rate of 54% on the complete scPDB and 64% on the druglike high-quality subset defined above. These improvements of approximately 7% and 5% can be attributed to a lower number of both sampling and scoring errors, the latter caused by a reduction in the number of good-scoring false minima. However, we chose not to present these improved success rates as the main results of this work, as more compact binding sites, especially when they are defined so as to closely follow the reference ligand conformation, will always improve the results of redocking studies. In the most common applications of protein−ligand docking, that

is, the modeling of binding modes of new ligands without X-ray structures or in virtual screening of compound libraries, limitations of the binding site may or may not improve the results. In these applications, the merit of a binding site modeled closely on the known complex structure will depend on the homogeneity of ligand size and that of the binding modes of the novel ligands docked.

## CONCLUSION

In this paper, GlamDock 1.0, our current docking tool, has been described and validated. The energy function used in GlamDock (ChillScore) is a differentiable approximation to piecewise linear potentials such as Ludi,[33] ChemScore,[19,34] and others. It has been parametrized as an empirical potential, by fitting to the binding free energies in PDBbind.[20] The results obtained in the validation show that the combination of ChillScore and the Monte Carlo with minimization approach implemented in GlamDock is a simple and effective method for obtaining high docking accuracy at high computational efficiency. The results on the first benchmark set allow the comparison of GlamDock to state-of-the-art docking software and place it well among the best available methods. Given the good performance of GlamDock in the Kellenberger data set, we assume that the results obtained with it on the scPDB mirror approximately the expected accuracy of state-of-the-art docking tools on automatically prepared proteins and ligands.

Our analysis of those results allowed an extensive characterization of the strengths and weaknesses of GlamDock in particular. Docking accuracy depends mainly on the flexibility of the ligand and the resolution of the binding site structure. A number of secondary effects can be reduced to those two main factors; the poor accuracy in docking peptides is mainly due to their flexibility, while the drug-likeness, which is often used to limit benchmark sets, appears to have an effect only for very nondruglike compounds, showing three or more Lipinski rule violations. In that group, however, the ligands are also highly flexible. This dependence on docking accuracy is clearly due to the sampling problem as is evidenced by the increase in sampling errors with the number of rotatable torsions in the ligand. Sampling appears to also be relatively poor for hydrophobic ligands, and this may be an indication that our rigid body mapping strategy needs to improve the placement of nonpolar compounds. Finally, scoring errors are more probable for highly hydrophilic compounds, which again may suggest the limitations of the protonation strategy we used for the ligands and the protonation and general preparation of the binding sites. However, it may also be due to the crude approximation of solvation effects in empirical potentials such as ChillScore. The use of better solvent models could help improve the prediction in those cases.

The overall result of this study, namely, that current state-of-the-art docking manages to predict the correct structure in almost half of the cases in redocking on automatically prepared data, is of importance for forward virtual screening, where a given protein is tested against a database of ligands. This is even more important for inverse virtual screening, where for a given ligand a large database of receptors is screened to identify a possible target,[6,7] the large number of receptor structures further extends this problem.

In addition, we strongly propose the use of large and comprehensive benchmarks for the evaluation of docking tools as an alternative to limiting the data upfront to complexes deemed suitable for docking. Figures 7 and 10 show the effect of several features on the success rates and types of error: this information would have been lost if thresholds on these features had been used to limit the benchmark. Furthermore, the second-order analysis of descriptor dependencies can only be performed with statistical confidence if the number of complexes is large enough.

## AVAILABILITY

The .mol2 files of all complexes used in this study are available upon request from the authors. The authors plan to make GlamDock, which is implemented in Java, available within the next year.

## ACKNOWLEDGMENT

The authors acknowledge partial funding by the DFG project AP 101/1-2 and Dr. Esther Kellenberger and Dr. Didier Rognan for providing the 100-complex benchmark set.

**Supporting Information Available:** Tables with detailed results for all complexes from the Kellenberger benchmark and the scPDB data set, including all descriptor values used in this study. This information is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule−Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(2) Abagyan, R.; Totrov, M.; Kuznetsov, D. Icm: A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(3) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

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

(5) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499−511.

(6) Chen, Y. Z.; Zhi, D. G. Ligand−Protein Inverse Docking and Its Potential Use in the Computer Search of Protein Targets of a Small Molecule. *Proteins* **2001**, *43*, 217−226.

(7) Paul, N.; Kellenberger, E.; Bret, G.; Mueller, P.; Rognan, D. Recovering the True Targets of Specific Ligands by Virtual Screening of the Protein Data Bank. *Proteins* **2004**, *54*, 671−680.

(8) Rockey, W. M.; Elcock, A. H. Progress toward Virtual Screening for Drug Side Effects. *Proteins* **2002**, *48*, 664−671.

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

(10) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins* **2004**, *57*, 225−242.

(11) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein−Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(12) Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. sc-PDB: An Annotated Database of Druggable Binding Sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717−727.

(13) Apostolakis, J.; Plueckthun, A.; Caflisch, A. Docking Small Ligands in Flexible Binding Sites. *J. Comput. Chem.* **1998**, *19*, 21−37.

(14) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein−Ligand Docking Using GOLD. *Proteins* **2003**, *52*, 609−623.
(15) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew R. K.; Olson, A. J. A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(16) Marialke, J.; Koerner, R.; Tietze, S.; Apostolakis, J. Graph-Based Molecular Alignment (GMA). *J. Chem. Inf. Model.* **2007**, *47*, 591−601.
(17) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631−644.
(18) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible Docking Using Tabu Search and an Empirical Estimate of Binding Affinity. *Proteins* **1998**, *33*, 367−382.
(19) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(20) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(21) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein−Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.
(22) Choi, V. On Updating Torsion Angles of Molecular Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 438−444.
(23) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(24) Ruppert, J.; Welch, W.; Jain, A. N. Automatic Identification and Representation of Protein Binding Sites for Molecular Docking. *Protein Sci.* **1997**, *6*, 524−533.
(25) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A Novel Method for the Shape-Directed Rapid Docking of Ligands to Protein Active Sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289−307.
(26) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A* **1976**, *32*, 922−923.
(27) Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A* **1978**, *34*, 827−828.
(28) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220* (4598), 671−680.
(29) Bystroff, C. An Alternative Derivation of the Equations of Motion in Torsion Space for a Branched Linear Chain. *Protein Eng.* **2001**, *14*, 825−828.
(30) Deo, A. S.; Walker, I. D. Overview of Damped Least-Squares Methods for Inverse Kinematics of Robot Manipulators. *J. Intell. Robotic Syst.* **1995**, *14*, 43−68.
(31) Kullback, S.; Leibler, R. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(32) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein−Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2005**, *49*, 5856−5868.
(33) Boehm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein−Ligand Complex of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(34) 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*, 503−519.

CI7001236