J Comput Aided Mol Des (2012) 26:185–197
DOI 10.1007/s10822-011-9539-5

Are predefined decoy sets of ligand poses able to quantify scoring function accuracy?

Oliver Korb · Tim ten Brink · Fredrick Robin Devadoss Victor Paul Raj · Matthias Keil · Thomas E. Exner

Received: 14 September 2011 / Accepted: 23 December 2011 / Published online: 10 January 2012

Abstract Due to the large number of different docking programs and scoring functions available, researchers are faced with the problem of selecting the most suitable one when starting a structure-based drug discovery project. To guide the decision process, several studies comparing different docking and scoring approaches have been published. In the context of comparing scoring function performance, it is common practice to use a predefined, computer-generated set of ligand poses (decoys) and to reevaluate their score using the set of scoring functions to be compared. But are predefined decoy sets able to unambiguously evaluate and rank different scoring functions with respect to pose prediction performance? This question arose when the pose prediction performance of our *piecewise linear potential* derived scoring functions (Korb et al. in J Chem Inf Model 49:84–96, 2009) was assessed on a standard decoy set (Cheng et al. in J Chem Inf Model 49:1079–1093, 2009). While they showed excellent pose identification performance when they were used for rescoring of the predefined decoy conformations, a pronounced degradation in performance could be observed when they were directly applied in docking calculations using the same test set. This implies that on a discrete set of ligand poses only the rescoring performance can be evaluated. For comparing the pose prediction performance in a more rigorous manner, the search space of each scoring function has to be sampled extensively as done in the docking calculations performed here. We were able to identify relative strengths and weaknesses of three scoring functions (ChemPLP, GoldScore, and Astex Statistical Potential) by analyzing the performance for subsets of the complexes grouped by different properties of the active site. However, reasons for the overall poor performance of all three functions on this test set compared to other test sets of similar size could not be identified.

Keywords Docking · Ranking · Conformational space · Sampling · Active-site properties

Introduction

Structure-based drug design and protein–ligand docking is applied in many drug discovery projects for hit identification [1]. The goal of a docking approach is to predict the relative orientation and the conformation of a ligand in the active site of a protein (*pose prediction*) and to estimate the binding affinity for this ligand (*scoring*). In a *virtual screening* campaign, a large database of ligands is docked and ranked according to their predicted binding affinity to identify highly active ligands. For solving the problems of pose and binding affinity prediction, a large number of different docking algorithms and scoring functions have been proposed. Although it has been over 25 years since the first docking tool DOCK was published in 1982 [2],

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9539-5) contains supplementary material, which is available to authorized users.

O. Korb · T. ten Brink · F. R. D. Victor Paul Raj · T. E. Exner (✉)
Department of Chemistry and Zukunftskolleg, University of Konstanz, 78457 Konstanz, Germany
e-mail: thomas.exner@uni-konstanz.de

O. Korb
The Cambridge Crystallographic Data Centre, Cambridge, UK

M. Keil
Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, QC H3A 2R7, Canada

there is still no universal docking tool available which outperforms all others on every system. Even worse, for some protein–ligand complexes no docking tool produces reliable results [3, 4]. The search for the best program has resulted in a large accumulation of comparative studies of docking programs that have been published in the literature during the past few years [3–15]. These studies report success rates, i.e. percentage of complexes for which the program identifies a pose reasonably close to the experimental structure on the highest rank, of 30–80% depending on the docking tools and the protein targets used [3–13]. In most studies, the scoring functions were identified as the main cause for failure. For speed reasons, scoring functions employed in docking calculations usually consist of simplified potential functions describing the protein–ligand interactions. Due to limitations, e.g. solvent and entropic effects as well as induced fit or conformational selection phenomena are only implicitly included in the parameterization, it is almost certain that no perfect agreement of the predicted and measured binding affinities can be obtained. Nevertheless, it is important to characterize the strengths and weaknesses of scoring functions, so that the most appropriate one can be selected for a specific problem at hand. In this respect, Cheng et al. [10] recently published a study, in which they compared 16 published scoring functions on a large test set consisting of 195 diverse protein–ligand complexes for which high-resolution crystal structures and reliable binding constants are available. The *docking power*, *ranking power*, and *scoring power* were analyzed, which are defined as follows [10]. *Docking power* describes the ability to identify a ligand pose close to the experimentally observed structure out of the set of computer-generated decoys. *Ranking power* and *scoring power* refer to the abilities to correctly rank different ligands bound to the same protein according to their binding affinities and to produce binding scores that are correlated with experimental binding affinities for a diverse set of targets, respectively.

In our opinion, docking tools and, thus, also docking scoring functions are designed to generate meaningful docking poses for a variety of targets, which will then be further investigated by other experimental or theoretical methods. Therefore, while other scoring functions were developed with the aim of reproducing binding affinities, our scoring functions were parameterized to accurately identify experimentally determined ligand poses. As described below in more detail, this strategy resulted in a good performance in redocking experiments of two well-known docking test sets. When assessing our scoring functions in terms of the *ranking* and *scoring power* directly on the crystallographic structures of the Cheng et al. [10] test set as done in the original study to avoid bias due to incorrect poses, a poor performance (Pearson correlation coefficient of 0.436 between docking score and log K_d) is observed. This is not unexpected since our scoring functions were solely trained for the purpose of pose prediction. Even if scoring functions trained on experimental binding activities perform slightly better on these tasks, other publications [11, 12] came to the conclusion that docking scoring functions are nowadays unable to reliably reproduce binding affinities and more sophisticated methods have to be applied to obtain reliable estimates. Therefore, we will concentrate here on the docking power while the *ranking* and *scoring power* will not be discussed further.

With this work we try to answer the question if predefined decoy sets can be used to reliably rank scoring functions with respect to their pose prediction performance. We first followed the work of Cheng et al. [10] and used the decoy set to compare our two new scoring functions ChemPLP [16] (implemented in our docking tool PLANTS [17, 18] and GOLD [19–21]) and a re-parameterized PLP [16] (*piecewise linear potential*) to the other scoring functions with respect to the *docking power*. Even though the results obtained for our scoring functions are among the best of all scoring functions assessed, we will show with additional studies that this protocol evaluates, in our opinion, only the rescoring power (for the chosen protocol) of a scoring function. As shown by O'Boyle et al. [22], the performance of a specific scoring function is almost always better if it is used to rescore poses generated using another scoring function than if the function is directly used to generate the poses. Many studies have shown that consensus docking schemes benefit from this observation [23–38]. In order to evaluate the pose prediction performance of the scoring functions more objectively, the conformational space should be sampled as accurately as possible. A discrete set of ligand poses generated using one scoring function will not necessarily cover the set of local or even global optima of a second scoring function (since in some scoring functions small values (more negative) and in others large values correspond to favorable poses and, thus, a minimization (e.g. PLANTS) and maximization (e.g. GOLD) problem has to be solved, respectively, we will use the term "optimum" in the following to avoid the confusion arising from the need to use "minimum" and "maximum" in dependency on the particular scoring function). However, when using the second scoring function directly in a docking calculation optimizing all translational, rotational and torsional degrees of freedom, different and better scoring optima might be sampled. In the context of pose prediction, these optima might correspond to correct or incorrect ligand poses when compared to the experimentally observed complex structure. To sample the conformational space more accurately, all decoys were individually used as the input for a full docking optimization run in additional experiments using PLANTS. This setup aimed at minimizing the

influence of sampling errors, especially when large ligands with many rotatable bonds are docked. Additionally, small differences in the bond distances and angles that resulted from the decoy set generation might have an influence on the docking results (the binding poses generated by LigandFit were optimized in situ with the "*Smart Minimizer*" in Discovery Studio during the decoy generation). Similar calculations were carried out with the GOLD software [19–21] (version 4.0) using the GoldScore [19, 20] and ASP scoring functions [39].

Materials and methods

As the PLANTS:ChemPLP and PLANTS:PLP scoring functions are described in detail in the literature [16], we will only highlight some of their features here. The functional forms of the scoring functions are based on parts of already published scoring functions and force fields. The *piecewise linear potential* (PLP) [40–43] scoring function is used in both cases to model steric complementarity of the protein and the ligand. In scoring function PLANTS:ChemPLP, terms of GOLD's ChemScore implementation [44] are added to introduce angle-dependent terms for hydrogen and metal bonding. The torsional potential from the Tripos force field [45] together with a heavy-atom clash term are employed to account for intra-ligand interactions. Starting from this functional form, a complete re-parameterization of the scoring function's weighting parameters was conducted. Using the best-performing parameter models regarding the reproduction of the experimentally observed ligand poses derived from the parameterization procedure for scoring function PLANTS:ChemPLP, 87% of the complexes of the Astex diverse set [46] and 77% of the CCDC/Astex clean list^{nc} (non-covalently bound complexes of the clean list) [16, 19] could be reproduced with root mean square deviations of <2 Å with respect to the experimentally determined structures. The success rates obtained for PLANTS:PLP are 84 and 71%, respectively.

As mentioned in the introduction, Cheng et al. [10] recently published a test set for the evaluation of scoring functions composed of 195 crystallographic complex structures (called primary test set in the original publication). For each of these complexes, they generated up to 100 alternative docking poses with four molecular docking programs, including LigandFit, Surflex, FlexX, and GOLD, which were then used for the evaluation of the *docking power* as described above. For a detailed description of the setup of the structures and decoy sets we refer to the original publication [10]. To evaluate the performance of PLANTS:ChemPLP and PLANTS:PLP on the decoy set, we took the complexes as given in this data set and rescored the crystal structures as well as all decoy poses with PLANTS (rescoring mode). For PLANTS:ChemPLP and PLANTS:PLP the published standard parameters [16] were used. Our two new scoring functions were then ranked with respect to the published data of the other state-of-the-art scoring functions presented in the original publication [10]. In addition to the results obtained using the standard parameters, we also present results without scoring of weak CH–O hydrogen bonds in PLANTS:ChemPLP. These hydrogen bonds have been found to be important in kinases, in which hydrogen bonds between a CH donor and an oxygen acceptor are frequently observed in the hinge region, but they are less important in other protein–ligand complexes [47, 48]. Recently, the ChemPLP scoring function was also implemented in GOLD version 4.0 [19–21]. To evaluate this implementation, the *docking power* was also determined using the rescoring option in GOLD.

For an objective comparison of scoring functions the decoy set needs to cover the conformational space such that the important optima (optima with the best scoring function values), especially the global optimum, are sufficiently close to one of these decoys. If this is not the case, the obtained success rates are only a measure of the rescoring power of the scoring function for this specific decoy set. Due to the roughness of the scoring functions, the relative ranking of the decoys could change with just small changes of the poses. Even more important, additional optima not represented by the conformations in the decoy set may exist, which may have better scores but may be less similar to the experimentally observed conformations. Due to the different functional forms of the scoring functions, the docking energy landscape is different for each scoring function, and therefore additional optima can exist not identified by the scoring functions used in the docking programs for the decoy generation. Such additional optima can have a negative effect on the pose prediction success rate and are therefore important for the evaluation of the *docking power* and the correct ranking of scoring functions with respect to this criterion.

That small conformational variations of the predefined decoy poses may strongly influence the ranking of scoring functions became first evident when the decoys were locally optimized using the Nelder-Mead Simplex (NMS) algorithm [49] implemented in PLANTS, which locally optimizes the rotational, translational and torsional degrees of freedom of the ligand as well as donor hydrogen positions in the protein binding site potentially involved in hydrogen bonding (see also the *supporting information*). To analyze these effects in more detail, exhaustive docking experiments were performed on all complexes with PLANTS. Standard *speed1* settings were employed [16] with spherical binding sites centered on the ligand and expanding 5 Å beyond each ligand atom. All decoy structures were used as the input for these docking

experiments to increase the sampling of the conformational space. Since the decoys are all the same molecule differing only in their conformation, the algorithm has been given *n* times the number of iterations of a single standard docking run, where *n* is the number of decoy structures for one complex. Due to its stochastic nature, the PLANTS algorithm can benefit from this additional sampling time so that there is a higher probability of finding the global optimum. Additionally, since in the final optimization step of some docking programs used for the decoy generation, bond lengths and angles are optimized in addition to the torsion angles, the structures deviate slightly from each other, which could also result in distinct docking poses. As will be shown in the "Results and discussion" section, the exhaustive docking experiments resulted in an extreme decrease in the success rates compared to the rescoring of the predefined set of poses. We performed the same exhaustive docking experiments with the GOLD program using the GOLD:GoldScore and GOLD:ASP scoring functions to see if the same observations regarding the decrease in the success rates can be made. For these experiments also standard settings and the binding site definition described above were used.

Results and discussion

Evaluating the *docking power* on the original decoy set

To allow for a direct comparison of the new results with the original ones, we will first follow the discussion of the original study. The term *docking power* refers to the percentage of top-ranked ligand poses selected by a specific scoring function resembling closely enough the one observed in the crystal structures [10]. The success rates for our scoring functions rescoring the decoy set are shown in Fig. 1 using three different rmsd cutoffs (rmsd < 1.0, <2.0, and 3.0 Å, respectively). Additionally, the 16 scoring functions from the original publication are reproduced for comparison. PLANTS implements two different versions of the ChemPLP function. The first one (standard version) takes weak CH–O hydrogen bonds into account (PLANTS:ChemPLP) while the second one does not (PLANTS:ChemPLP/noCH–O). The results for GOLD's implementation of ChemPLP (GOLD:ChemPLP) are also presented and reasons for the slightly different behavior of the PLANTS and GOLD implementations are discussed in the *supporting information*. The results seem very encouraging since the three ChemPLP versions are ranked second, third, and fifth if an rmsd cutoff of <2.0 Å is chosen (success rates: 81, 81, and 79%, respectively). Only GOLD:ASP performs slightly better (success rate: 82.5%). Noteworthy, PLANTS:PLP reaches a respectable fourth place (success rate: 80.5%) behind GOLD:ASP, PLANTS:ChemPLP and PLANTS:ChemPLP/noCH–O and shows a similar performance to another PLP derivative, DS:PLP1 (success rate: 75.4%). This is remarkable considering the simple functional form and the small number of atom types used in this scoring function. When looking at an rmsd cutoff of <1.0 Å PLANTS:PLP even outperforms PLANTS:ChemPLP (69.7% vs. 65.1%). An interesting observation is that the PLANTS:ChemPLP success rates with and without accounting for CH–O hydrogen bonds are exactly the same for this test set. As already mentioned, this type of hydrogen bonds has been shown to be important for kinases. For some kinase complexes in the test set the

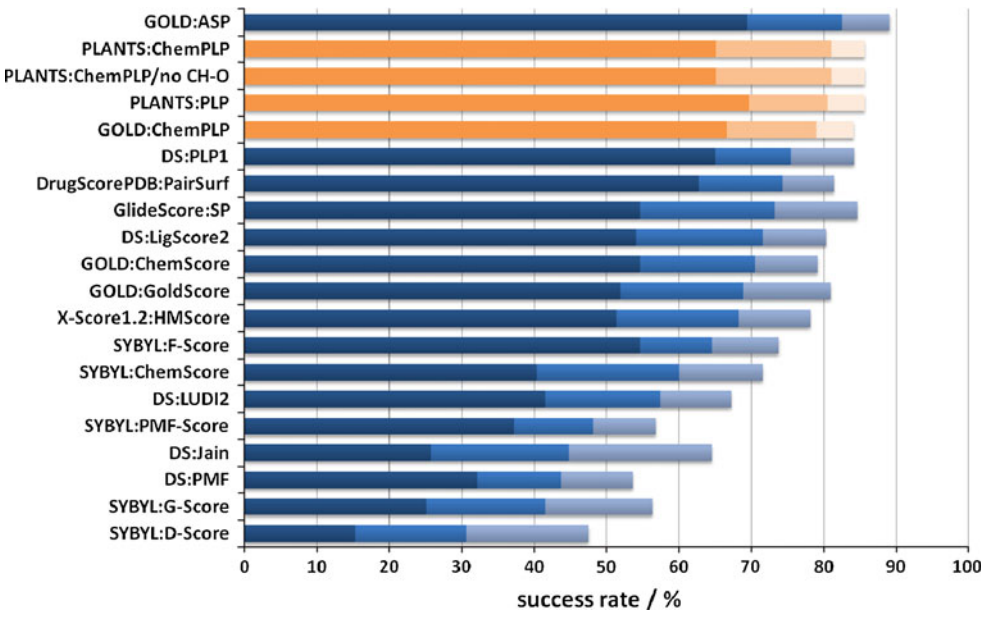

Fig. 1 Comparison of the success rates of 16 scoring functions taken from the original study (*blue bars*) and 4 from our study (*orange bars*). Scoring functions are ranked by the success rates when the acceptance rmsd cutoff is <1.0 Å (*dark bars*), 2.0 Å (*medium dark bars*) and 3.0 Å (*light bars*)

calculated scores are reduced by up to 3 units when considering CH–O interactions, corresponding to the contribution of exactly one weak hydrogen bond. As will be shown below, the reduced scores actually lead to different results with respect to the success rates if full docking experiments are performed on these complexes.

Figure 2 compares the success rates of all scoring functions when the experimentally determined ligand conformations are included or excluded from the decoy set. The success rates for PLANTS:ChemPLP decrease slightly if the correct pose is not included. For PLANTS:PLP, the success rate drops by almost 7% showing that, even if there are decoys very close to the experimentally observed structure, the results could potentially be improved for these scoring functions (but also for e.g. GOLD:ASP, DS:PLP1, DrugScore^{PBD}:PairSurf, and X-Score1.2:HM-Score; all showing decreases of approximately 5% and more) by generating decoys even closer to the experimental structure. This pronounced decrease is, in our opinion, the first indication that the decoy set is not adequate to probe all relevant regions of the respective scoring function landscapes.

Exhaustive docking of the test set

As mentioned before, our scoring functions have shown an excellent pose identification performance. However, the rescoring experiments carried out in the original publication [10], and followed here up to this point, evaluate, in our opinion, the performance of a consensus docking scheme rather than the performance of a scoring function when used within a docking procedure. This is in line with the work by O'Boyle et al. [22], in which the authors stated that in every case in their docking/rescoring scheme the mean rank obtained for a scoring function on its own is either the same as or poorer than when the scoring function is used to rescore poses generated using another scoring function. The scoring function landscape of a specific scoring function could exhibit optima different to the ones of other scoring functions and especially to those used in the docking programs that were employed for the generation of the decoy poses. Some of these optima, although corresponding to ligand poses with large rmsd values compared to the crystal structure, could show better scoring function values than the ones close to the crystal structure and, thus, the global optimum would represent an incorrect pose prediction result. However, these optima including the global one may not be sampled when performing rescoring on a *predefined* set of poses. Thus, the performance of one specific function obtained in studies using decoy sets will not solely be determined by its own scoring performance but also by the features of the scoring functions used for the decoy generation. Highly correlated scoring functions will describe the energy surface in a similar way and, thus, they will have optima in similar regions and the correction of a docking failure is less likely in a rescoring experiment. Scoring functions with high complementarity, i.e. which describe the binding event using different, non-redundant terms, will more likely be able to differentiate between true and false binding poses when used in a consensus scoring scheme. To reduce the influence of rescoring effects and obtain a truly objective comparison of scoring functions, all local optima, including the global optima, must be taken into account, which is of course practically impossible.

In order to achieve a better coverage of the conformational space (translational, rotational, and torsional degrees

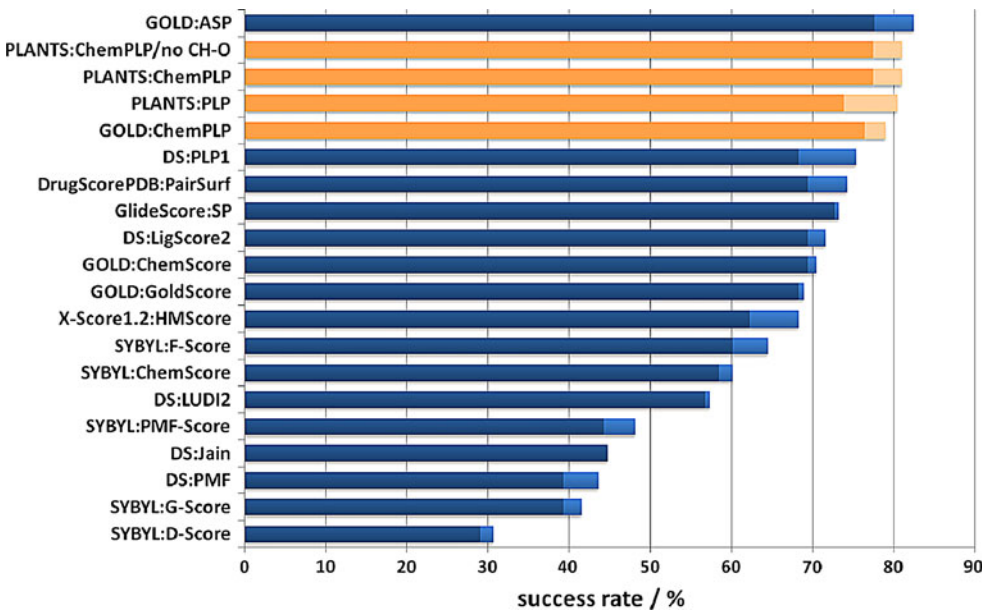

Fig. 2 Comparison of the success rates of 16 scoring functions taken from the original study (*blue bars*) and 4 from our study (*orange bars*) when the true binding pose is included in the decoy set (*light bars*) or not (*dark bars*). The acceptance rmsd cutoff is <2.0 Å. Scoring functions are ranked by the success rates when the true binding poses are considered (*light bars*)

of freedom), we assessed the PLANTS:ChemPLP scoring function using full docking experiments on the same test set. As described in the "Materials and methods" section, we used all decoy structures as the input for the docking program PLANTS. In this way, each ligand is docked between 21 and 101 times to its corresponding protein structure. These multiple docking experiments aimed at minimizing sampling errors and, thus, the important optima and especially the global optimum are more likely to be identified. When looking at the structure with the lowest scoring function value for each complex identified over all runs, these structures have an rmsd to the experimental ones lower than 2 Å in 106 of the 195 complexes, which corresponds to a success rate of 54%. In comparison with the over 80% success rate obtained in the rescoring of the decoy set, this result is rather sobering. To identify the reason for this, we looked at the scoring function values of the best-ranked poses. For all complexes the scoring function values are lower compared to the poses of the original decoys. The average improvements in the scores are 20 and 10 units with respect to the original and locally minimized decoys, respectively. In only one case (protein-tyrosine phosphatase, PDB code 1nny) the scoring function value of the locally optimized decoy (see *supporting information*) is lower than the one obtained in the docking experiment, which can be explained by sampling problems. Using even longer sampling times, the optimum identified by the local optimization should also be found with the docking approach resulting in a successful docking (rmsd local optimization: 0.66 Å, rmsd docking: 6.09 Å). For all other complexes, sampling is not the main issue, it is the existence of additional highly favorable scoring function optima located far away from the experimentally observed pose, which are not represented by structures of the decoy set.

Although some decrease in the docking success rates was expected, the very low success rate is astonishing since much higher success rates were obtained with other similar-sized test sets [16, 19]. To test if this is a specific problem of PLANTS and the ChemPLP scoring function, we performed corresponding docking runs with the other docking program available to us: GOLD using GoldScore and the Astex Statistical Potential (ASP). GOLD:GoldScore predicts exactly the same number of complexes correctly as PLANTS:ChemPLP (106 complexes, 54% success rate) while GOLD:ASP is slightly better with 110 correctly predicted complexes, corresponding to a success rate of 56%. This shows that for the test set studied here it is much harder to reproduce the correct binding pose than with other published data sets. As already mentioned in the introduction, the two docking programs can achieve success rates of around 80% for the CCDC/Astex [19] and the Astex diverse data set [46] using the same standard settings for the scoring functions and search settings as used in the study presented here [16, 19, 46]. Looking at single complexes, some correlations between the scoring functions can be identified (see Fig. 3). There is a large cluster (67 complexes) in the lower left corner in which the best-scored pose of each scoring function has an rmsd of <2 Å representing a successful docking. The correct binding pose of 51 additional complexes cannot be reproduced by any of the scoring functions. But there are also complexes that can be docked correctly with one (44 complexes) or two (33 complexes) scoring functions but not with the other ones. Thus, the ranking of the scoring functions is highly dependent on the complex at hand. As we will show below, it is very hard to identify features of the binding site or the ligand that can be used to spot problematic cases or even to choose the optimal scoring function. It seems that at the moment the only possibility is to test different docking approaches and scoring functions on known complexes of the same protein and use the one giving the best results.

It is also surprising that GOLD:GoldScore shows such a fall-off in the success rate since it was one of the scoring functions used for the decoy set generation [10] and should, therefore, not profit from the rescoring effects described above. Even if the success rate for the decoy set (69%) was lower than for PLANTS:ChemPLP, there is still a 15 % decrease observed in our new docking experiment. Due to the multiple docking runs performed in this work, the search space is explored more thoroughly, resulting in the identification of additional optima. The sampling issue is also obvious when looking at the success rates for each individual complex over the 21–101 docking runs, which used all decoy structures as input (see Fig. 4). Only for around 130 complexes the success rate is either above 90% or below 10%, i.e. the docking approach is able to identify the correct pose in almost all docking runs or is never able

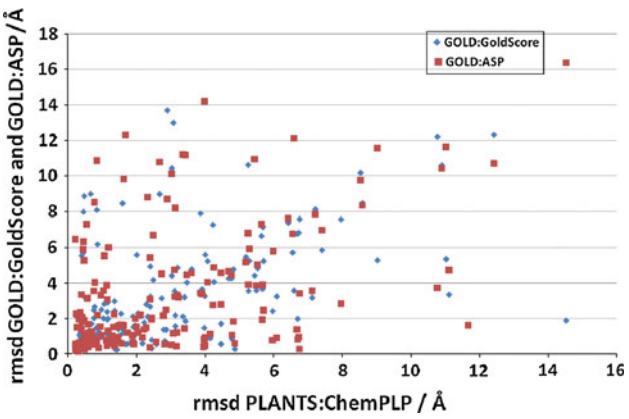

Fig. 3 Docking rmsd values obtained for the 195 complexes of the test set using the PLANTS:ChemPLP scoring function in comparison to GOLD:GoldScore and GOLD:ASP

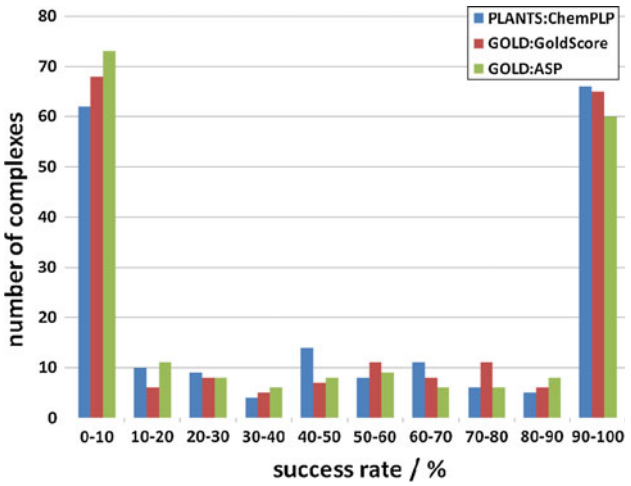

Fig. 4 Success rates for the 195 complexes averaged over the 21–101 docking runs performed with the decoy structures as input. For many complexes, success rates are between 10 and 90% indicating sampling problems

to do so. For the remaining complexes a variety of optima are produced in the different runs, which shows that at least the amount of sampling carried out in this work, or even more, is necessary to identify the global optimum. The small structural differences with respect to bond angles and lengths might also contribute to the different results per docking run, but this is very hard to quantify.

Correlation of pose prediction success rates with molecular properties of the binding site and the ligand

Recently, studies have appeared in the literature [4, 12, 50] attempting to explain the successes and failures of docking approaches and scoring functions by analyzing the physicochemical properties of the complexes. Following the same idea we used physicochemical properties mapped on the molecular surface [51] to describe features of the binding site and the ligand. Since not all ligands are located in a deep cavity, cavity detection algorithms were not able to identify all binding sites correctly. Therefore, the molecular surface of the protein up to 3 Å away from the ligand was defined as the binding site. The properties used to characterize the binding site were the total surface area of the site, the local lipophilicity, the electrostatic potential, the cavity depth quantifying the concavity of the binding site, and the hydrogen-bond density [51]. All these properties were calculated at each point of the surface and then averaged over the binding site using the Molcad molecular modeling software [52–54].

In Fig. 5 the root mean square deviation of the best-scoring poses using the three scoring functions is plotted versus the lipophilicity, cavity depth, and hydrogen-bond density, respectively. One trend apparent from these plots is that it seems harder to dock ligands into shallow binding sites. For the rest of the properties it is hard to find any explanation for a docking failure as only a small number of complexes exhibits extreme values for any property, so any conclusions drawn from these results would probably be statistically insignificant. This is also the case for the

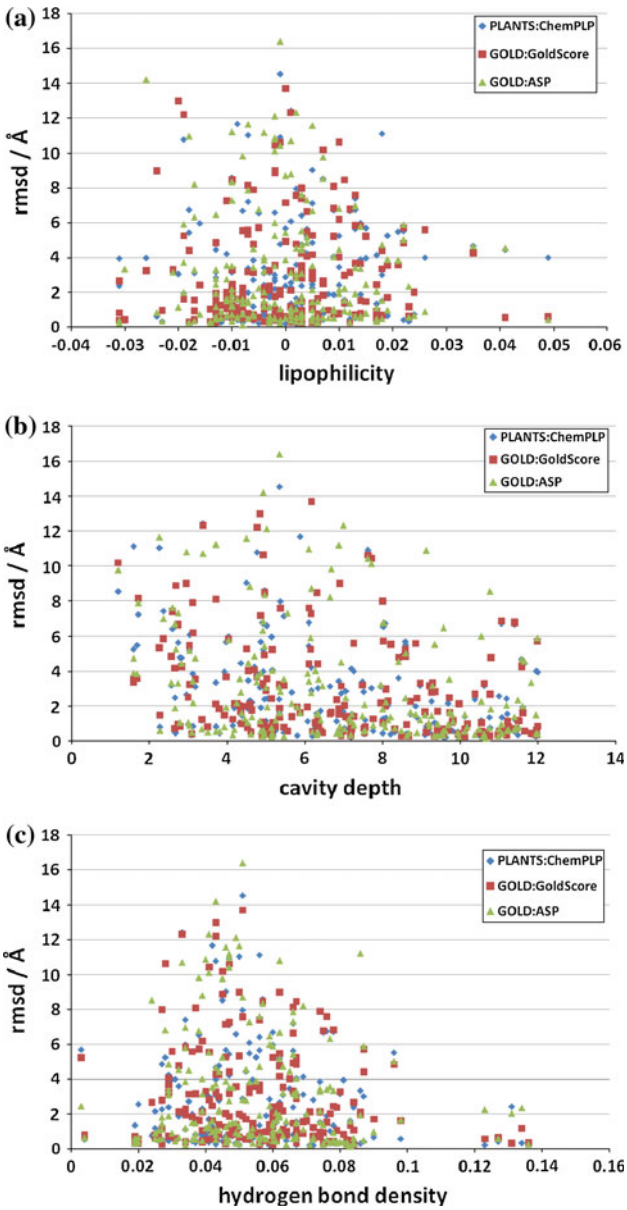

Fig. 5 Root mean square deviation of the best scoring docking pose versus **a** the local lipophilicity, **b** the cavity depth and **c** the hydrogen-bond density averaged over the binding site. **a** Positive and negative values correspond to lipophilic and hydrophilic binding sites, respectively. **b** Small and large values correspond to very shallow binding sites and deep cavities, respectively. On average lower rmsd values can be observed for the deep cavities located in the part of the plot. **c** Small and large values correspond to almost no and many hydrogen bonding sites, respectively

properties not shown here such as the binding-site size, where one would expect that larger ligands and, thus, larger binding sites would result in more sampling and scoring function problems, but no such correlation could be observed. Additional evidence that success or failure in docking cannot solely be explained by the properties of the binding site alone is given in Fig. 6. The test set was constructed such that for each protein target three ligands with different binding affinities were included. The figure shows the number of targets for which zero, one, two, or all three ligands were predicted correctly. For the majority of targets, successful and unsuccessful docking runs are observed.

Thus, the properties of the ligands were expected to have a strong influence on docking performance. However, performing an analysis of ligand properties similar to the one presented for the protein binding sites above revealed hardly any correlation of ligand features with docking failures. Only the ligand surface area, which essentially is a measure for the ligand size, showed some correlation. As expected, with increasing ligand size, the success rates decreased (see *supporting information*). A property related to the ligand size and frequently used to characterize ligands is the number of ligand rotatable bonds (see Fig. 7). Reasonable success rates of over 60% are only obtained for ligands with less than 5 rotatable bonds. Noteworthy, for this range of rotatable bonds GOLD:ASP even reaches over 70%. With an increasing number of rotatable bonds the success rates drop rapidly until they reach <10% for 20 rotatable bonds and more using PLANTS:ChemPLP. Overall, the distribution of rotatable bonds is similar to other test sets [19], so that the worse results cannot be attributed to a larger number of large and flexible ligands. On the contrary, complexes in all ranges of rotatable bonds show a similar decrease in the success rate compared to the corresponding ranges in the other test sets [16]. A last possibility to explain the poor performance is that the test set was specially designed to also include ligands with low binding affinity. There seems to be a slight tendency that low affinity complexes are harder to dock than high affinity ones. Individual results are presented in the *supporting information.*

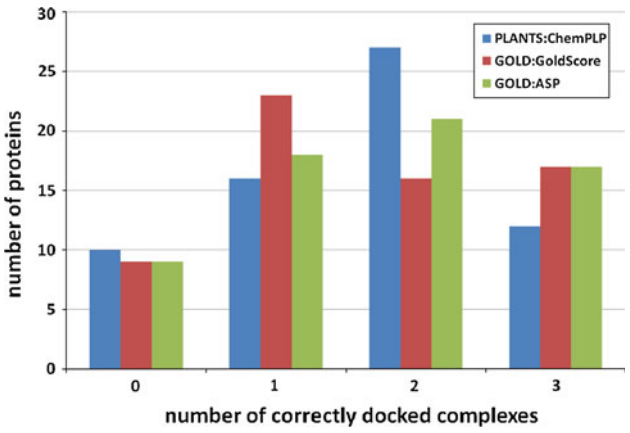

Fig. 6 Number of protein targets into which none, one, two, and all three corresponding ligands can be docked successfully (top-ranked solution rmsd <2 Å)

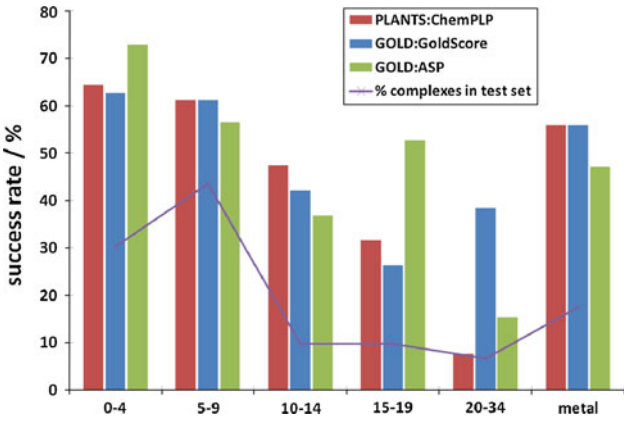

Fig. 7 Dependence of the success rate on the number of ligand rotatable bonds and the presence of metal ions in the binding site. Results are presented for scoring functions PLANTS:ChemPLP, GOLD:GoldScore and GOLD:ASP. Additionally, the respective percentage of complexes part of the data set is reported

Clustering of complexes according to molecular properties of the binding site and the ligand

Since the performance of the scoring functions cannot be explained by a single property as shown in the last paragraph, we finally tried to look at the performance problem from a multidimensional perspective by clustering the complexes based on six properties. The four active-site properties surface size, lipophilicity, cavity depth and hydrogen bond density were used and additionally the number of rotatable bonds of the ligand and the binding affinity were considered. The KNIME workflow management system [55] was used to perform the clustering. First, the six properties were normalized by a linear mapping to the range between 0 and 1 (min–max normalization: the minimum value is mapped to 0 and the maximum value to 1) to remove the bias due to the different scales. These were then taken as input for the divisive hierarchical cluster algorithm of KNIME with average linkage and the Euclidian distance of the normalized properties as the similarity criterion. This clustering produces a hierarchy of clusters by a top-down approach, which starts with one large cluster, in which all data points are included, and performs splits of the most dissimilar cluster recursively until each data point is assigned to its individual cluster generating a tree-like structure. By looking at the

assignment of the data points to the clusters after each split visualized by a dendrogram, one can get an idea about the structure of the dataset and determine optimal numbers for the cluster count. In this way, the hierarchical clustering has the advantage that the number of clusters has not to be selected a priori as different cluster counts can be chosen according to the data at hand without rerunning the clustering.

The dendrogram for the clustering of the 195 complex structures can be found in the *supporting information*. In the first three splits, three small clusters are separated from the rest of the complex structures. In the fourth split, the remaining very large cluster is divided up into two additional clusters at a ratio of 3:2. These resulting five clusters will be used for the first analysis. The number of complexes and the average values of the normalized properties are presented in Table 1, while the success rates and average rmsds for the three scoring functions PLANTS:ChemPLP, GOLD:GoldScore, and GOLD:ASP are reported in Table 2. Please note that due to the small number of complexes in some of the clusters, the statistical relevance may be questionable. The complex structures of the first and the third cluster can be characterized by small and deeply buried binding sites. In the case of very polar binding sites (cluster A1) PLANTS:ChemPLP and GOLD:GoldScore show a good to excellent performance. For lipophilic binding sites (cluster A3), GOLD:GoldScore is still performing well while for GOLD:ASP and PLANTS:ChemPLP inferior success rates are observed. Complexes with extremely large ligands are grouped in the third small cluster (cluster A2). As expected, all scoring functions show a bad pose prediction performance for these complexes. PLANTS:ChemPLP predicts none of the complexes in this cluster correctly, which may be attributed to sampling issues. Due to the large number of complexes in the remaining two clusters, it is harder to characterize these since the spread in the properties leads to almost identical average values. However, in direct comparison it seems that the larger and shallower binding sites are assigned to cluster A4 and the deeper cavities to cluster A5. Again, complexes with deeper binding sites are easier to predict and thus better success rates are obtained for cluster A5 in comparison to cluster A4 for all three scoring functions.

To analyze the large clusters in more detail, we also looked at the distribution of the complex structures when considering 10 clusters (results are presented in Tables 3 and 4). The first five clusters (B1 to B5) are either identical to the small three clusters described above (A1 to A3) or are too small to be statistically significant. Cluster A4 is divided into two clusters, where the high-affinity complexes and the low- to medium-affinity complexes are assigned to cluster B7 and B8, respectively. The complexes of cluster B7 are on average also slightly larger, deeper, and more hydrophilic than the ones of cluster B8. All these properties are favorable for predictions using empirical scoring functions, which is confirmed by the above-average success rates of PLANTS:ChemPLP and GOLD:GoldScore. The remaining three clusters originate from cluster A5. Cluster B6 can again be characterized by small, hydrophilic and deeply buried ligands but with less hydrogen-bonding capabilities. Since the properties are in between the ones of cluster A1 and A3 it makes sense that also the success rate of B6 falls in between the ones observed for these clusters. The remaining complexes of cluster A5 are separated according to higher (cluster B9) and lower (cluster B10) affinity. The success rates for these two clusters are comparable with the only exception of

Table 1 Number of complexes and average values of the normalized properties, and success rates for the scoring functions PLANTS:ChemPLP, GOLD:GoldScore, and GOLD:ASP after dividing the dataset into 5 clusters

Cluster	No. of complexes	log K_d	Lipophilicity	H-bonding	Cavity index	Surface area	No. of rot. bonds
A1	5	0.35	0.47	0.96	0.92	0.09	0.14
A2	9	0.47	0.30	0.35	0.32	0.878	0.82
A3	6	0.21	0.74	0.15	0.91	0.14	0.13
A4	72	0.44	0.37	0.38	0.27	0.39	0.29
A5	103	0.38	0.37	0.39	0.68	0.23	0.17

Table 2 Success rates and average rmsd for the scoring functions PLANTS:ChemPLP, GOLD:GoldScore, and GOLD:ASP after dividing the dataset into 5 clusters

Cluster	PLANTS:ChemPLP		GOLD:GoldScore		GOLD:ASP	
	Success rate	rmsd	Success rate	rmsd	Success rate	rmsd
A1	0.80	0.77	1.00	0.62	0.40	1.51
A2	0.00	6.42	0.33	5.50	0.22	8.88
A3	0.33	3.33	0.67	2.03	0.50	2.24
A4	0.44	3.48	0.40	3.70	0.43	3.56
A5	0.66	1.91	0.63	2.23	0.70	2.30

Table 3 Number of complexes and average values of the normalized properties, and success rates for the scoring functions PLANTS:ChemPLP, GOLD:GoldScore, and GOLD:ASP after dividing the dataset into 10 clusters

Cluster	No. of complexes	log K_d	Lipophilicity	H-bonding	Cavity index	Surface area	No. of rot. bonds
B1	3	0.14	0.56	0.01	0.85	0.23	0.20
B2	3	0.53	0.32	0.45	0.92	0.45	0.50
B3	3	0.28	0.91	0.29	0.98	0.05	0.07
B4	5	0.36	0.47	0.96	0.92	0.09	0.14
B5	9	0.47	0.30	0.35	0.32	0.88	0.82
B6	7	0.48	0.09	0.48	0.90	0.16	0.16
B7	9	0.73	0.28	0.33	0.40	0.60	0.39
B8	63	0.39	0.38	0.39	0.25	0.36	0.27
B9	49	0.49	0.39	0.27	0.64	0.31	0.19
B10	44	0.22	0.40	0.50	0.67	0.14	0.12

Table 4 Success rates and average rmsd for the scoring functions PLANTS:ChemPLP, GOLD:GoldScore, and GOLD:ASP after dividing the dataset into 10 clusters

Cluster	PLANTS:ChemPLP		GOLD:GoldScore		GOLD:ASP	
	Success rate	rmsd	Success rate	rmsd	Success rate	rmsd
B1	0.67	2.29	0.67	2.25	0.67	1.26
B2	1.00	1.17	1.00	0.97	0.67	2.52
B3	0.00	4.38	0.67	1.81	0.33	3.22
B4	0.80	0.77	1.00	0.62	0.40	1.51
B5	0.00	6.42	0.33	5.50	0.22	8.88
B6	0.57	1.80	0.71	1.40	0.57	1.46
B7	0.78	2.31	0.78	2.25	0.56	3.73
B8	0.40	3.65	0.35	3.91	0.41	3.53
B9	0.65	1.80	0.63	2.28	0.67	2.58
B10	0.66	2.09	0.59	2.40	0.75	2.09

GOLD:ASP showing higher success rates for the low-affinity complexes.

In summary, the clustering was able to separate the dataset into groups with specific active-site features and some differences in the success rates of these groups could be explained by these features. For example PLANTS:ChemPLP and GOLD:ASP seem to be suited for deep, polar binding sites while GOLD:GoldScore performs better for larger, lipophilic complexes. The better performance of PLANTS:ChemPLP on polar targets can probably be explained by the fact that lipophilic interactions are only described by the simple PLP potential while for hydrogen bonds a more sophisticated distance- and angle-dependent potential is used. For GOLD:ASP the preference for polar binding sites is not that easily explainable, since lipophilic and polar potentials are both derived from known complex structures. However, since lipophilic interactions are less well defined in general, deriving discriminatory distance-dependent knowledge-based potentials seems to be harder. The better performance of GoldScore on nonpolar targets may be attributed to the explicit use of a slightly softened van der Waals potential which has a high weight relative to hydrogen-bonding interactions. Overall, the differences in the performance observed for the different groups are, however, not highly pronounced and therefore the overall low number of correctly predicted complexes cannot be attributed to a single or a small number of properties.

Conclusion

In this study we compared the rescoring of a predefined discrete decoy set [10] to performing full docking calculations in order to assess the pose prediction performance of scoring functions. If the predefined decoy set is used to rank our new scoring functions with respect to the other published ones [10], they belong to the best-performing class with respect to pose generation accuracy. If an acceptance rmsd cutoff of <2.0 Å and the original decoy set are used success rates of 81.0 and 80.5% are obtained for PLANTS:ChemPLP and PLANTS:PLP, respectively.

Unfortunately, this picture changes drastically if the decoy set is locally optimized (see *supporting information*) or if full docking experiments are performed. For

PLANTS:ChemPLP, GOLD:GoldScore, and GOLD:ASP, which we used as examples here, but probably all other scoring functions, more poses different to the decoys of the original data set with very favorable scores can be identified when the scoring function is used directly in the docking process. One reason for the success of rescoring campaigns is that by combining the docking and rescoring function wrong optima, identified by only one of these two functions are avoided [22]. Thus, the rank of all functions only used for rescoring are probably better than if the complete conformational space (translational, rotational, and torsional degrees of freedom) had been searched, so that the setup of the original study to some extent disfavors the scoring functions used for pose generation.

Overall, these results show that scoring functions can be compared objectively with respect to their docking performance only if the ligand conformation generation and scoring steps are not separated. A predefined set of decoys will never be able to represent all local optima for all potentially existing scoring functions. Scoring functions with a high complementarity to the ones used in decoy generation will artificially be favored due to the different description of the binding events. The combination of such complementary functions in a consensus or rescoring scheme will lead to a better discrimination of true and false binding poses and an elimination of docking failures. Only an accurate sampling of the fitness landscape of a specific scoring function is able to identify most of the local optima and, critically, the global optima. For the test set studied here a pronounced decrease in the success rates can be observed when comparing the results of the decoy set with the full docking. The success rates decrease by approximately 25–30% for PLANTS:ChemPLP and GOLD:ASP and by 15% for GOLD:GoldScore. If the evaluation is based on the original decoy set, PLANTS:ChemPLP and GOLD:GoldScore are ranked 2nd and 11th (see Fig. 1), respectively, but they achieved the same success rates in the full docking experiments. While the study of Cheng et al. [10] describes the rescoring performance of a set of scoring functions applied to predefined poses, our full docking experiments present a more detailed picture of the fitness landscape underlying the specific scoring functions. This is also of interest in the context of virtual screening applications, where in the first instance the scoring function values obtained for each compound are used for the generation of a ranked database.

The overall low number of correctly predicted complexes in our full docking experiments is unexpected. For similar-sized data sets much higher success rates have been reported [16]. We attempted to analyze why this is the case by correlating physicochemical properties of the binding sites with the success rates. While for most properties the results were inconclusive, a dependence on the ligand size with decreasing success rates for larger ligands was apparent. Moreover, the multidimensional analysis revealed a preference of PLANTS:ChemPLP for polar active sites and, as expected, a preference of all scoring functions for deep cavities. Apart from the properties studied here, additional factors like the protonation states of the ligands and the protein structures or water molecules in the binding site may have an impact on the docking results [56–59]. While for the experiments carried out here the dataset was used as supplied in order to be comparable to previous studies, further experiments taking these factors into account may be of interest.

The results presented here as well as many studies published in the past demonstrate that today's scoring functions remain far from being perfect. Identifying the best one for a target at hand is a nontrivial task as the outcome of scoring function comparisons is highly dependent on the dataset used as well as the performance evaluation protocol itself. If no information on the scoring function performance for a particular target is available, a worthwhile option can be to look at the scoring function performance on highly similar targets provided that any exist. In case no related targets exist, the use of consensus scoring schemes may be the protocol of choice due to their relatively high robustness to outliers [23–38]. Given the lucky situation that structural and/or affinity data for the target at hand is available, assessing the performance of these scoring functions within the same docking package is obviously the most effective way for making an informed decision about which one to choose in the discovery process. Also the option of training a target-specific scoring function should be taken into account if enough training and test instances are available [60–62]. Although we in principle agree that a decoupling of sampling and scoring would be highly desirable, we believe that, for the reasons given in this study, the consideration of scoring function rankings derived from rescoring discrete decoy sets should be the last option in the decision making process.

Researchers tend to independently repeat scoring function benchmark exercises for the same target, sometimes coming to the same conclusion, sometimes identifying an alternative scoring protocol. A step forward with respect to avoiding redundant evaluations or to report alternative protocols could be a community database aiming at storing the outcome of specific docking and scoring protocols applied to a particular target. This could facilitate the decision making process of researchers working on the same or related targets and at the same time inform researchers involved in the development of scoring

functions where the shortcomings are. We would really like to see such an initiative in the near future.

Acknowledgment The authors thank Renxiao Wang for providing the diverse test set of 195 protein–ligand complexes as well as Colin Groom and John Liebeschuetz for helpful discussions. The work was supported by the Konstanz Research School Chemical Biology (KoRS-CB), the Zukunftskolleg and the Young Scholar Fund of the Universität Konstanz. O.K. acknowledges support of the Landesgraduiertenförderung Baden-Württemberg and the Postdoc-Programme of the German Academic Exchange Service (DAAD). Additionally, we thank the Common Ulm Stuttgart Server (CUSS) and the Baden-Württemberg grid (bwGRiD), which is part of the D-Grid system, for providing the computer resources making the computations possible.

References

1. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Nat Drug Dis 3:935–949
2. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin ThE (1982) J Mol Biol 161:269–288
3. von Korff M, Freyss J, Sander T (2009) J Chem Inf Model 49:209–231
4. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) J Chem Inf Model 49:1455–1474
5. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Proteins 57:225–242
6. Krovat EM, Steindl T, Langer T (2005) Curr Comput Aided Drug Des 1:93–102
7. Perola E, Walters WP, Charifson PS (2004) Proteins 56:235–249
8. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49:6789–6891
9. Hevener KE, Zhao W, Ball DM, Babaoglu K, Qi J, White SW, Lee RE (2009) J Chem Inf Model 49:444–460
10. Cheng T, Li X, Liu Z, Wang R (2009) J Chem Inf Model 49:1079–1093
11. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2005) J Med Chem 49(20):5912–5931
12. Englebienne P, Moitessier N (2009) J Chem Inf Model 49:1568–1580
13. Corbeil CR, Moitessier N (2009) J Chem Inf Model 49:997–1009
14. Chikji A, Bensegueni A (2008) J Proteomics Bioinform 1:161–165
15. Li X, Li Y, Cheng T, Liu Z, Wang R (2010) J Comput Chem 31:2109–2125
16. Korb O, Stützle T, Exner TE (2009) J Chem Inf Model 49:84–96
17. Korb O, Stützle T, Exner TE (2006) Lect Notes Comput Sci 4150:247–258
18. Korb O, Stützle T, Exner TE (2007) Swarm Intell 1:115–134
19. Nissink JWM, Murray CW, Hartshorn MJ, Verdonk ML, Cole JC, Taylor R (2002) Proteins 49:457–471
20. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727–748
21. Jones G, Willett P, Glen RC (1995) J Mol Biol 245:43–53
22. O'Boyle NM, Liebeschuetz JW, Cole JC (2009) J Chem Inf Model 49:1871–1878
23. Okamoto M, Masuda Y, Muroya A, Yasuno K, Takahashi O, Furuya T (2010) Chem Pharm Bull 58(12):1655–1657
24. Huang SY, Grinter SZ, Zou X (2010) Phys Chem Chem Phys 12(40):12899–12908
25. Zhong S, Zhang Y, Xiu Z (2010) Curr Opin Drug Discov Devel 13(3):326–334
26. Bar-Haim S, Aharon A, Ben Moshe T, Marantz Y, Senderowitz H (2009) J Chem Inf Model 49(3):623–633
27. Fukunishi H, Teramoto R, Takada T, Shimada J (2008) J Chem Inf Model 48(5):988–996
28. Teramoto R, Fukunishi H (2008) J Chem Inf Model 48(4):747–754
29. Teramoto R, Fukunishi H (2008) J Chem Inf Model 48(2):288–295
30. Renner S, Derksen S, Radestock S, Moerchen F (2008) J Chem Inf Model 48(2):319–332
31. Wolf A, Zimmermann M, Hofmann-Apitius M (2007) J Chem Inf Model 47(3):1036–1044
32. Teramoto R, Fukunishi H (2007) J Chem Inf Model 47(2):526–534
33. Betzi S, Suhre K, Chetrit B, Guerlesquin F, Morelli X (2006) J Chem Inf Model 46(4):1704–1712
34. Oda A, Tsuchida K, Takakura T, Yamaotsu N, Hirono S (2006) J Chem Inf Model 46(1):380–391
35. Miteva MA, Lee WH, Montes MO, Villoutreix BO (2005) J Med Chem 48(19):6012–6022
36. Xing L, Hodgkin E, Liu Q, Sedlock D (2004) J Comput Aided Mol Des 18(5):333–344
37. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB (2002) J Mol Graph Model 20(4):281–295
38. Charifson PS, Corkery JJ, Murcko MA, Walters WP (1999) J Med Chem 42(25):5100–5109
39. Mooij WT, Verdonk ML (2005) Proteins 61:272–287
40. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST (1995) Chem Biol 2:317–324
41. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PW (2002) Proteins 48:539–557
42. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PW (2003) Proteins 53:201–219
43. Verkhivker GM (2004) J Mol Graph Model 22:335–348
44. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Proteins 52:609–623
45. Clark M, Cramer RD III, Van Opdenbosch N (1989) J Comput Chem 10:982–1012
46. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW (2007) J Med Chem 50:726–741
47. Panigrahi SK (2008) Amino Acids 34:617–633
48. Panigrahi SK, Desiraju GR (2007) Proteins 67:128–141
49. Nelder JA, Mead R (1965) Comput J 7:308–313
50. Pencheva T, Soumana OS, Pajeva I, Miteva MA (2010) Eur J Med Chem 45:2622–2628
51. Keil M, Exner TE, Brickmann J (2003) J Comput Chem 25(6):779–789
52. Waldherr-Teschner M, Goetze T, Heiden W, Knoblauch M, Vollhardt H, Brickmann J (1992) MOLCAD—computer aided visualization and manipulation of models in molecular science. In: Post FH, Hin AJS (eds) Advances in scientific visualization. Springer Verlag, Heidelberg, pp 58–67
53. Brickmann J, Goetze T, Heiden W, Moeckel G, Reiling S, Vollhardt H, Zachmann C-D (1995) Interactive Visualization of Molecular Scenarios with MOLCAD/SYBYL. In: Bowie JE (ed) Data visualisation in molecular science: tools for insight and innovation. Addison-Wesley Publishing Company Inc., Reading, Mass, pp 83–97
54. Brickmann J, Keil M, Exner TE, Marhöfer R (2000) J Mol Model 6:328–340

55. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2007) KNIME: the Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds) Studies in classification, data analysis, and knowledge organization (GfKL 2007). Springer, pp 319–326
56. ten Brink T, Exner TE (2009) J Chem Inf Model 49:1535–1546
57. ten Brink T, Exner TE (2010) J Comput Aided Mol Des 24:935–942
58. Thilagavathi R, Mancera RL (2010) J Chem Inf Model 50:415–421
59. Verdonk ML, Chessari G, Cole JC, Hartshorn MJ, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) J Med Chem 38:6504–6515
60. Ravitz O, Zsoldos Z, Simon A (2011) J Comput Aided Mol Des 25:1033–1051
61. Seifert MHJ (2009) J Comput Aided Mol Des 23:633–644
62. Pham TA, Jain AN (2008) J Comput Aided Mol Des 22:269–286