Bias, reporting, and sharing: computational evaluations of docking methods

Ajay N. Jain

Received: 29 October 2007 / Accepted: 12 November 2007 / Published online: 13 December 2007

Abstract Computational methods for docking ligands to protein binding sites have become ubiquitous in drug discovery. Despite the age of the field, no standards have been established with respect to methodological evaluation of docking accuracy, virtual screening utility, or scoring accuracy. There are critical issues relating to data sharing, data set design and preparation, and statistical reporting that have an impact on the degree to which a report will translate into real-world performance. These issues also have an impact on whether there is a transparent relationship between methodological changes and reported performance improvements. This paper presents detailed examples of pitfalls in each area and makes recommendations as to best practices.

Keywords Docking · ROC · Evaluation · Bias · Enrichment · Virtual screening

Introduction

Discovery of novel lead compounds through virtual screening of chemical databases against protein structures is well established [1], but there is still much room for improvement in key aspects of algorithm performance, and this places a premium on the value of effective and accurate performance evaluations, a number of which have been published recently [1–5]. Many docking methods have been described, and they vary in their approaches to two components: scoring functions [6–11] and search methods [12–18]. Searching for optimal ligand poses is tightly coupled with scoring, since the optimal pose is *defined* by the extremum of the scoring function on a particular ligand and protein. In a formal sense, the docking problem is a search for a global optimum in an energy landscape that is defined by the scoring function, protein, ligand, and the degrees of freedom to be explored. Perturbations in the scoring function or search strategy clearly have an impact on the solutions that a method will report. Less obviously, even small changes in the protein or ligand can influence the *shape* of the energy landscape, not just the starting point of the search, which will also affect the solutions that will be reported. Results are strongly dependent on the precise atom locations and bond orders of both the protein and ligand, since not all degrees of freedom on either the protein or ligand are explored exhaustively.

There are three primary criteria for evaluating docking strategies: geometric docking accuracy, screening utility, and scoring accuracy. Geometric docking accuracy measures ability to generate and recognize the native conformation and alignment of a ligand bound to its cognate protein beginning from an arbitrary initial pose. The requirement is for the scoring function to have a global extremum at the location in parameter space that corresponds to the native bound pose and for the search algorithm to locate that extremum. Screening utility describes performance in ranking cognate ligands of a protein above non-ligands, as is desired in a virtual screening application. Here, the search algorithm must seek global extrema for each ligand, and the magnitudes of the extrema for different ligands must correctly rank true ligands above non-ligands. Scoring accuracy that is sufficient to produce correct rankings *within* a set of true ligands requires still more accuracy in the magnitude of extrema, but the search problem may be less complex,

A. N. Jain (✉)
University of California San Francisco, Box 0128,
San Francisco, CA 94143-0128, USA
e-mail: ajain@jainlab.org

since one may know experimental poses for examples within a chemical series that is being optimized, and these examples may be used to guide the search process.

In considering strategies for evaluating docking methods, it is important to understand that the operational application of the methods is to make predictions about ligands that are *not known*. If we accept this proposition, and we accept that method developers and evaluators desire to meaningfully communicate the likely real-world performance of the methods, there are implications for data sharing, design and preparation of data sets, and statistical reporting. In the case of binding mode prediction, we *do not know* the bound pose of the ligand being predicted. In the case of virtual screening or of affinity prediction, we *do not know* the bound pose or biological activity of any ligands being tested, though in the latter case we may have some information about binding geometry. It is critical in docking evaluations to prevent information that would not be known in practice from influencing the testing of a method. This sort of bias may often be obvious, but, as we will see, it can be quite subtle. Here, consideration of bias issues will lead to specific recommendations for data sharing and data set preparation.

With respect to statistical reporting, the goal should be to make correct conclusions in cases where correctness is objectively definable and to make use of metrics that support comparability between different reports that may involve some differences in methods or data sets. As a generality, both correctness and comparability are supported by use of metrics that are sensitive to *characteristics of distributions* of values (e.g., of scores or RMS deviations) as opposed to the specific samplings made to estimate the distributions. For docking accuracy and scoring accuracy, there has been relatively little variation in reporting metrics, but in the case of screening utility, a number of metrics yield counterintuitive (and sometimes clearly incorrect) conclusions. Here, specific recommendations will be made for metrics that avoid these problems.

Methods and data

This paper focuses on methods for evaluating docking approaches. All of the docking procedures themselves have been described previously and will be only briefly summarized here. The data and scripts for generating the results reported here are available by request from the author.

Data

The data to be used come from two sources: one illustrative screening case (PPAR-gamma) from the DUD data set [3] and a set of 100 docking accuracy protein/ligand pairs from the recent report of Surflex-Dock (called the Vertex100 set) [19], which was adapted from the work of Perola, Walters, and Charifson [4].

DUD

The DUD data set consists of protein structures, known active ligands, and a set of decoys specifically designed with knowledge of known active ligands that are intended to represent a difficult background. Within the DUD decoy set, a distinction is made between "self" decoys and "all" decoys. The former are target-specific, and the latter are the superset of all self-decoy sets for all 40 targets. For the PPAR-gamma case used here, there were 85 active ligands (ppar_gamma_ligands.mol2) and 3,127 target specific decoys (ppar_gamma_decoys.mol2). The superset of all DUD decoys included 124,413 molecules. These molecules were used unmodified. The protein structure for the PPAR-gamma case was PDB code 1FM9, which was provided as a PDB file with no protons. Protons were added with standard geometries and modified as described later. In addition to the DUD self and DUD all decoy sets, the 1,000 molecule ZINC-based [20] decoy set was also used (originally from [21] and also used in [3, 19]). The salient distinction between the DUD decoy sets and the ZINC set (apart from size) is that the latter was created with no knowledge of the active ligands of any particular target. Among the decoy sets used by Huang et al. in the DUD report that *did not* make use of knowledge of active ligands (termed the "Jain set" in that study), this set was the most challenging [3].

Vertex100

Perola et al. reported the docking accuracy of Glide, GOLD, and ICM on 200 complexes, 100 of which were available from the authors [4]. Complexes for that study were selected where a binding constant was available, where the ligand and protein interaction was non-covalent, and where crystallographic resolution was less than 3.0 Å. Ligands were selected to have molecular weight between 200 and 600, 1 to 12 rotatable bonds, be drug/lead-like, and be structurally diverse. Proteins were selected from multiple classes to be relevant for drug discovery. Critical to the observations discussed here, the protons of the proteins were *optimized* in the presence of the cognate ligand for each protein. In making use of this benchmark for this study, all structures were used beginning with the precise coordinates provided in order to illustrate the effects of different strategies for protein preparation.

Methods

Specific methodology is not the focus of this report. However, in order to illustrate the effects of data set preparation and different statistical metrics, Surflex-Dock was employed to provide real data. A very detailed exposition of Surflex-Dock has been published recently [19], and only a brief summary is presented here.

The scoring function used in Surflex-Dock (and in Hammerhead, which was Surflex's antecedent) has terms for hydrophobic contact, polar interactions, and entropic fixation costs for loss of torsional, translational, and rotational degrees of freedom. The terms, in rough order of significance, are: hydrophobic complementarity, polar complementarity, entropic terms, and solvation terms. By far the most dominant terms are the hydrophobic contact term and a polar contact term. Surflex-Dock also implements a molecular mechanics force field (based on DREIDING [22]) that supports all-atom optimization of ligand and/or protein while considering the non-bonded effects using the empirical function that guides docking.

A detailed account of the Surflex-Dock search algorithm can be found in the original paper [23]. Surflex employs an idealized active site ligand (called a protomol) as a target to generate putative poses of molecules or molecular fragments. Poses of the molecular fragments that tend to maximize similarity to a protomol are used as input to the scoring function and are subjected to thresholds on protein interpenetration and local optimization. The partially optimized poses of the fragments form the basis for further elaboration of the optimal pose of the full input ligand. The procedure identifies high scoring fragments that have compatible geometries to allow for merging in order to assemble a high scoring pose of the full input ligand. The whole molecules resulting from the merging procedure are pruned based on docking score, and are subjected to further gradient-based score optimization. The procedure returns a fixed number of top scoring poses.

Docking a ligand to assess docking accuracy was performed to be comparable to the report of Perola et al. (see [19] for details): `surflex-dock-v211.exe -pgeom dock_list ligand.mol2 protomol.mol2 protein.mol2 log`. To illustrate the effects of protein atom optimization, protein atoms were optimized using the following procedure: `surflex-exptl-v2204.exe popt xtal-lig.mol2 protein.mol2`. The procedure yielded optimized version of the protein and ligand ("`popt.mol2`" and "`opt.mol2`"). Subsequent docking made use of the modified protein for docking and the modified ligand from which to measure RMSD (precisely analogous to [15] and described in more detail in [19]). Screening a database against a protein was done as follows: `surflex-dock-v211.exe -pscreen dock_list test-db.mol2 protomol.mol2 protein.mol2 log`. This was done for the PPAR-gamma target for the known ligands as well as multiple decoy sets.

Results and discussion

Evaluating docking methodology is challenging for two reasons. First, the docking process itself is complex, since it is a multi-dimensional optimization problem in an energetic landscape that is highly non-linear. For a docking method that treats proteins as perfectly rigid, the combination of the protein coordinates and the scoring function define the landscape, with the ligand pose parameters (or possibly full ligand atomic coordinates) being subject to optimization. To the extent that a docking method to be tested *does not* vary certain parameters in its input (e.g., protein coordinates or ligand bond lengths/angles), those parameters affect either the shape of the landscape to be searched or the parts of the landscape that are accessible. This makes docking strongly dependent on choices of input preparation that vary between different practitioners. Second, the statistical treatment of some aspects of docking performance is difficult. In particular, screening utility assessment is made somewhat challenging since the typical application of screening involves a highly skewed population of actives (very few) versus inactives (very many) coupled with an operational cost function that varies from user to user. Some users will tolerate missing a substantial proportion of active molecules in return for effective elimination of all but a small proportion of inactive molecules; others will not find such a tradeoff acceptable. The following results will build a case for broad data sharing, minimal protein preparation, consistent ligand preparation, and for a small number of metrics to be reported consistently in studying the performance of docking methods.

Data set sharing: reproducibility and detection of bias

The first difficulty in docking application, particularly with respect to reproducing the results of another study, is that provision of a PDB code is not sufficient to define the actual input to most docking programs. The DUD data set is generally exemplary from the perspective of data sharing, with free and complete availability of structure files for ligands and decoys, in reasonable protonation states, and in a widely used file format [3]. However, the 40 proteins that form the targets for testing virtual screening are provided as unprotonated PDB files. Figure 1 shows the example of PPAR-gamma, which made use of the PDB structure 1FM9. Naïve addition of protons resulted in two strong

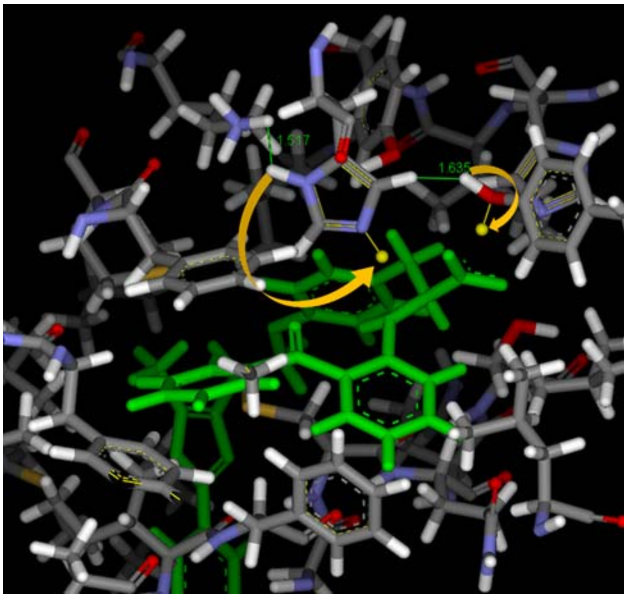
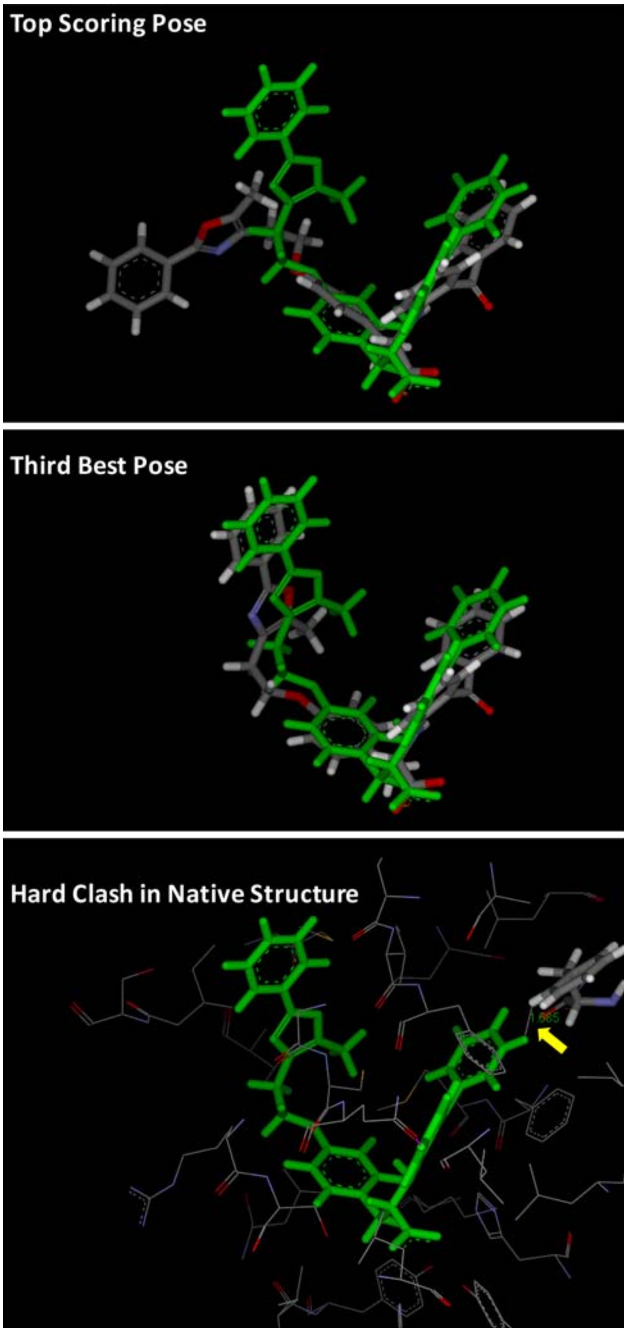


Fig. 1 Proteins lacking hydrogens are insufficient for reproduction of docking results. Here, automatically added protons generated intra-protein clashes within PPAR-gamma (PDB structure 1FM9). The histidine tautomer was changed as well as the rotamer for the hydroxyl on the tyrosine as part of the protein preparation process

clashes within the protein near the active site. These were relieved by a change of tautomer for an imidazole and a change of rotamer for a hydroxyl. Both changes make a significant difference in the energetic landscape to be searched, since they result in hydrogen bond donors on the protein becoming accessible to the binding pocket. Given that there is no generally accepted, automatic, and freely available procedure for determining proton positions given a PDB file, it is incumbent on investigators to provide the structures that were used as input to their docking method, unless there is a scientifically defendable reason for making use of proprietary structures. Obviously, if a method only makes use of polar protons (or no protons at all), the structures as used in the study should be provided.

Figure 2 shows Surflex-Dock's performance in docking the cognate ligand of 1FM9 beginning from a minimized random input pose. The top scoring pose places the carboxylate correctly, but the phenyl-oxazole is misplaced, resulting in an RMSD of 4.4 Å. The third best scoring pose is much better, with an RMSD of 1.1 Å. Note that the phenyl-ketone in the upper right is shifted down from the experimental structure. The bottom panel shows a substantial hard clash between the ligand and protein in the native structure, amounting to an interpenetration between aromatic hydrogens of over 0.5 Å. The Surflex-Dock scoring function, while being less stiff in terms of penalization of inter-atomic overlap than, for example, a Lennard–Jones treatment, will not accept such an overlap when there are other solutions available. In this case, the result is a higher score for a clearly incorrect predicted pose than for one much closer to being correct. Nominally, this represents a hard docking failure.


Top Scoring Pose

Third Best Pose

Hard Clash in Native Structure

Fig. 2 The top picture shows the top scoring docked pose of the native ligand to the PPAR-gamma DUD target, corresponding to 4.4 Å RMSD. The middle picture shows the third-best scoring pose (1.1 Å RMSD). There is a fundamental difficulty in that the structure, as solved experimentally, has a hard clash between aromatic hydrogens corresponding to an overlap of over ½ Angstrom

In considering how to interpret this result, it is important to recognize that the problem of redocking a ligand into the structure of a protein for which one already knows the experimental solution is artificial. It is reasonable to construct and use benchmarks of this type to establish an upper bound on how well docking methods might work in the best possible circumstance: when one has a protein structure that does not change much on binding the ligand under consideration. Let us imagine that the structure in Fig. 2 was solved with the isopropyl analog instead of the phenyl. In that case, no hard clash would be evident, but the question of whether the phenyl would "fit" and how it might score would be relevant. This is the cross-docking question, and it is the most relevant one in the operational application of docking methods. If a method is able *in the docking process* to adjust the protein conformation such that the correct pose scores best, that is the behavior that will lead to desirable operational performance. However, adjustment of the protein conformation with knowledge of the correct bound pose of the ligand makes a somewhat artificial test explicitly biased.

Bias in data sets: from subtle to unsubtle

Some groups, though, still take this approach, as follows. Clashes such as those shown in Fig. 2 are ameliorated by optimizing the conformation of the protein and its cognate ligand simultaneously as part of the structure preparation process within a self-docking study where the optimization method is related to the scoring function to be tested in the docking procedure [15]. Further, RMSD is computed from the *optimized* cognate ligand coordinates. Docking begins from a randomized pose, but the optimization procedure has modified the energetic landscape such that the combination of the optimized protein coordinates along with the scoring function *guarantee* a local extremum at precisely the coordinates of the optimized cognate ligand. Figure 3 shows the result of applying optimization to the protein and ligand using Surflex's scoring function for the non-bonded terms and a DREIDING force field for the bonded terms (as detailed in [19]) and docking to PPAR-gamma beginning from the same randomized input pose as used in Fig. 2. The atomic motions of the protein are very small, with an overall RMSD of 0.06 Å, and for the ligand as well (0.4 Å). The bottom overlay in Fig. 3 shows the top-scoring pose that resulted from docking as in Fig. 2 but using the modified protein instead. The resulting pose has 0.7 Å RMSD from the native crystallographic ligand coordinates (compared with 4.4 Å using the unmodified protein), and lower still if measured from the optimized ligand coordinates.

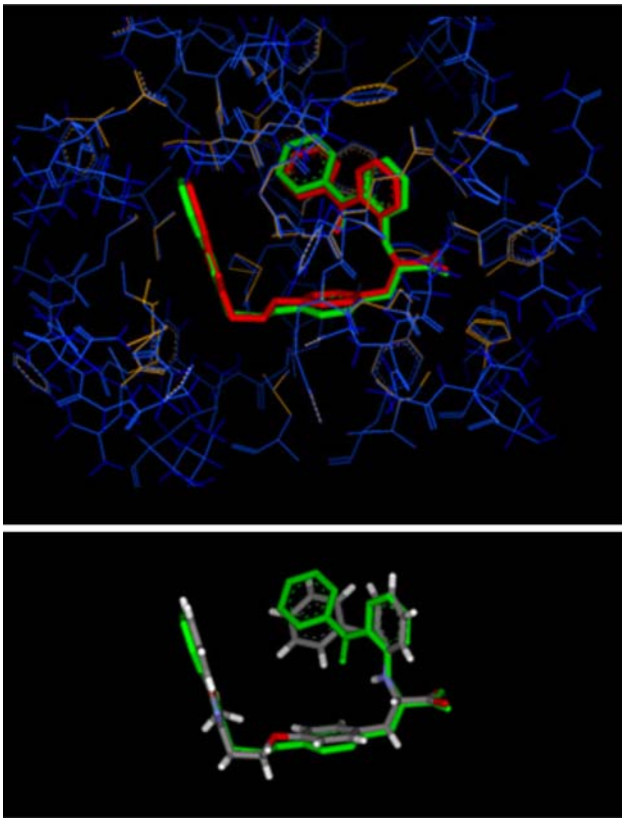


Fig. 3 Top picture shows the original protein structure (blue) with the original ligand structure (green) along with the optimized protein structure (orange) and the optimized ligand structure (red). The overall change to the protein was 0.06 Å RMSD for the heavy atoms, and all individual atomic motions were small. The bottom picture shows the ligand redocked from the same coordinates as from Fig. 2 but using the modified protein structure and corresponding protomol. Measured from the original crystallographic coordinates, the top scoring pose (green) now yield 0.7 Å RMSD (compared with 4.4 Å using the unmodified coordinates)

Figure 4 shows that this effect is not anecdotal (data taken from [19]). On the Vertex100 docking accuracy set, Surflex-Dock exhibits a nominal improvement at the 1 Å success threshold from 30% success using unmodified proteins to over 50% using "optimized" proteins and measuring RMSD from optimized cognate ligands. The source of the improvement is subtle. Docking proceeds from randomized input ligand poses (in this case, CORINA conformations were used), which clearly have *no memory* of their bound state. However, the docking search procedure can reliably generate poses that are fairly close to correct among the top twenty solutions for most proteins. The challenge is in ranking the top twenty returned poses such that the top scoring one is close to correct. By perturbing the energetic landscape defined by the protein coordinates *using a function related to the docking scoring function*, a memory of the coordinates of the optimized cognate ligand pose is created in the form of a local extremum in this perturbed energetic landscape. The docking procedure exploits this extremum by making use of local optimization to refine the final poses returned by the search procedure, thus ensuring that some of the final poses have moved very close to what is known *a priori* to

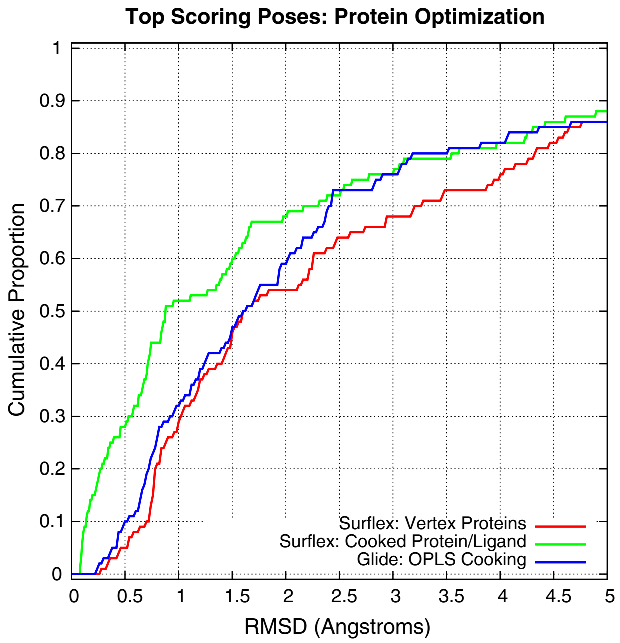


Fig. 4 Cumulative proportion of top scoring RMSD over 100 complexes are shown, with results for Surflex using the original Vertex data (red), Surflex using optimized proteins and reference ligands (green), and Glide results using the Vertex prepared proteins (whose hydrogen atoms were optimized using Glide's preferred forcefield). This is not a fair comparison, but it is analogous to multiple published reports

be a local optimum. By measuring "correctness" from the optimized cognate ligand, a systematic improvement in RMSD is observed *on top of* the increased ease of ranking the poses. It is not common to employ this procedure, but some investigators have [15]. Optimization of protein protons only and without biased RMSD computations is more common [4].

It is possible to magnify this effect by allowing larger displacements of protein atoms, and it is *impossible* to detect this effect if protein coordinates are not made available to accompany reports of the performance of docking methods. Given the data presented in Fig. 4, it would be easy to claim the following:

> "On a respected benchmark of pharmaceutical relevance, using analogous preparation procedures to those established by other investigators, Surflex's docking accuracy has been shown to be superior to other methods, especially with respect to the proportion of extremely well-docked top-scoring poses (<1.0 Å RMSD)."

However, the more helpful conclusion is as follows:

> "On a respected benchmark of pharmaceutical relevance, the protein preparation procedures established by other investigators have been shown to yield systematic memory effects of a ligand's cognate pose within the protein structure itself. The memory arises from the relationship between the energy function used in protein/ligand preparation to the one used for scoring in docking. This is true whether moving all atoms or moving just protons. An additional bias arises from measuring RMSD based upon optimized cognate ligand coordinates."

Additional details on this point are available in the recent Surflex-Dock report [19]. However, there are two observations to be emphasized here. First, preparation procedures that make use of knowledge of what is to be predicted (here the cognate ligand pose) must be *avoided* in constructing test cases for molecular docking. In practice, this means that protons should be placed on proteins according to default geometries, with modifications of tautomers and rotamers only where clearly required. Further, optimization of any protein atomic positions *outside of the docking process* must be avoided unless such optimization makes no use of a ligand to be docked or a close analog thereof. Second, since bias can hide in small coordinate changes, the message from the previous section on data sharing must be taken seriously.

The generous approach to data sharing from Perola, Walters, and Charifson [4] of the Vertex100 data led to the foregoing analysis and observations about hidden bias in protein conformations. A data set shared by Cummings et al. [24] led to a related observation about hidden bias in ligand conformations [19]. As detailed in previous work, common methods for generating 3D structures of ligands can lead to scaffold-specific differences in strain energy [19]. These differences can influence the outcome of screening utility studies, either increasing or decreasing the nominal performance of a docking method. Figure 5 shows cumulative histograms of the strain energies for the 85 DUD PPAR-gamma ligands and the 3,127 PPAR-gamma decoys. If such strain is unresolvable by a docking method through torsional variation, and if the method is sensitive to, for example self-clashing within the ligand, the final scores of the docked ligands may be systematically higher or lower than the decoys. In the PPAR-gamma case, there is a slight advantage for the true ligands. From a theoretical perspective, strain that is correlated with prediction class (active or decoy) is a systematic bias in a screening benchmark. From a practical perspective, it can dominate the results of a test of screening performance. In the recent Surflex-Dock report, for three of four screening examples, ligand minimization vastly improved screening enrichment due to high strain present in the cognate ligands that was in excess of the strain present in the decoys [19].

Both of these illustrations highlight subtle issues of bias, where benchmark preparation imparts a positive or

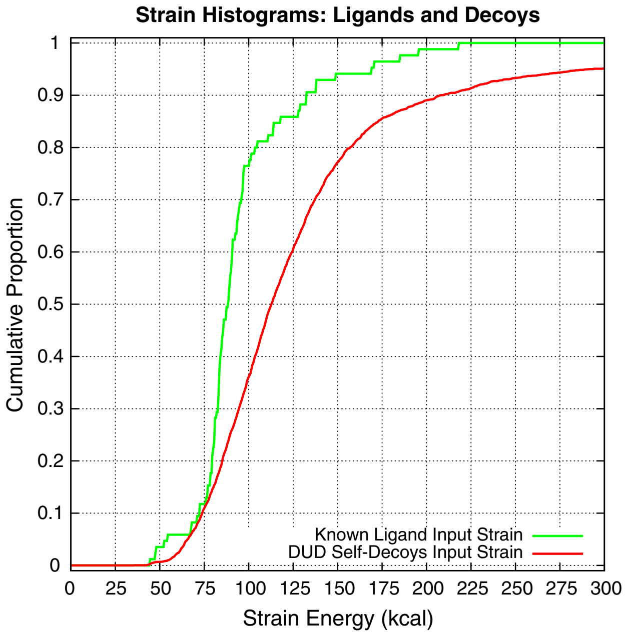


Fig. 5 The distributions of strain energy for PPAR-gamma ligands and their designed decoy set are systematically biased in favor of the ligands

negative systematic influence on results stemming from the conformations of proteins or ligands. There are much less subtle source of bias to avoid. For screening utility studies, privileged treatment of actives (multiple or hand-selected protonation states, tautomers, ring conformations) over decoys is somewhat common but should be avoided. *Post facto* results selection is less common (e.g., reporting screening enrichment for a subset of ligands that are known to dock correctly [25]) but should also be avoided. For docking accuracy, similar observations apply. The general principle is simple: use of any knowledge of the correct answer in a manner that is incongruent with operational application of a method is inappropriate in assessing the performance of a method.

Statistical reporting: distributions and standards

Whether reporting docking accuracy, screening utility, or scoring accuracy, the process is essentially the same. A collection of examples is used to *estimate* the performance of a method, with the hope that the specific collection of examples is representative of the *population* of examples that will be encountered in a real-world application. Some aspects of benchmarks vary widely from study to study; in particular the number of cases used and the numbers of ligands and/or decoys employed. Performance metrics will generally be sensitive to sample size changes in the sense that smaller sample sizes yield larger variance in the metrics. However, sensible metrics will, on average, yield the same value independent of the sample size. This is a desirable property, since it enhances the comparability of performance data from different reports.

In the case of docking accuracy, the most widely used metric is proportion of correctly docked top scoring poses with 2.0 Å RMSD or less. If one makes use of a sample size of 100, 200, or 300, from a parent population of 1,000 complexes, assuming that the specific choice of which complexes to use was random, the proportion of correct dockings will be very similar. Except for a few outlier reports (which may use metrics like average RMSD), docking accuracy is measured by this intuitive and sensible population statistic. While there are reasonable arguments to be made that RMSD has flaws, until there is wide consensus that a different metric is preferable, the current practice should continue. In the case of quantitative assessment of scoring accuracy, owing to the long history of QSAR, the metrics used are generally well-grounded in statistics and have the correct behavior with respect to changes in sample size.

However, for screening enrichment, we have little agreement on which metrics to employ, and a number of the most widely used have poor behavior with respect to sample size issues. In a screening enrichment experiment for a particular target, some set of actives and some set of decoys are required. Recall from above, the particular actives are being used to derive an *estimate* for the behavior of the docking method on the population of ligands that will bind the protein site in question. The particular decoys are being used to derive an *estimate* for the behavior on molecules that do not bind the site. Typically, there are tens of ligands available as known actives, and the general practice is to make use of as many as possible. Decoy sets have a large range, from about 1,000 to more than 100,000. There are two critical points to be made here. First, since the accuracy in estimating the separation of the populations is strongly limited by the size of the sample of actives, the size of the sample of decoys is not important as long as it is substantially larger (e.g., 10-fold) than the size of the sample of actives. Second, most of the metrics that have been proposed and used to report screening enrichment are inappropriately and strongly affected by the ratio of active to decoy population size, and this sensitivity can lead to clearly incorrect interpretation of results.

Figure 6 shows smoothed histograms (left) and cumulative histograms (right) for the docking scores of PPAR-gamma ligands (green), DUD self decoys (red), and ZINC decoys (blue). It is the relationship between these curves that gives rise to all metrics for reporting screening enrichment. It is clear from visual inspection that the DUD self-decoys yield sharply higher scores than the ZINC

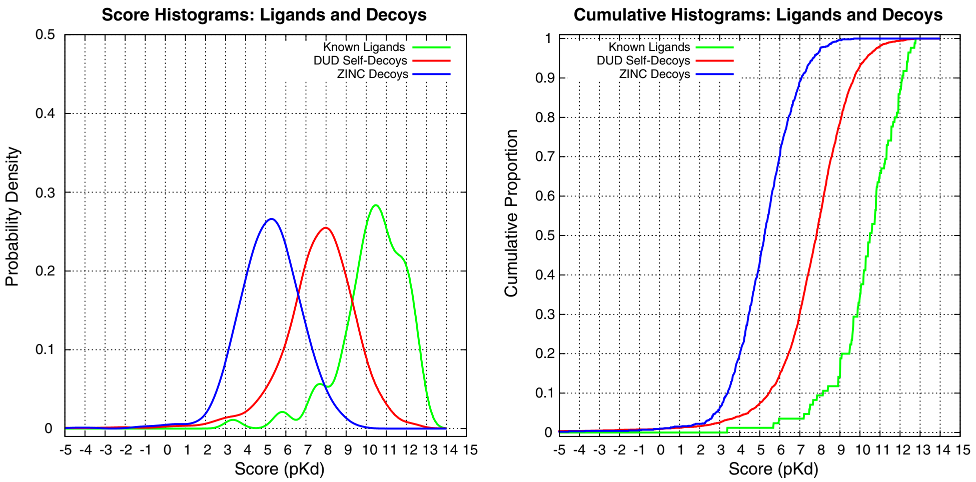


Fig. 6 Histograms (left) show the relationship of the scores of the true ligands of PPAR-gamma (green), the DUD self decoys (red), and the ZINC decoys (blue). Cumulative histograms (right) offer the advantage of allowing for direct reading of true and false positive rates. For example, at the score threshold of 8, 90% of the true ligands are found, less than 5% of the ZINC decoys, and about 50% of the DUD self decoys. Clearly, the DUD self-decoy set is more difficult to differentiate from the true ligands than the ZINC set. This is true for early enrichment and for overall enrichment since the DUD self distribution is cleanly shifted to the right of the ZINC distribution

decoys and that the overlap of the DUD scores with the scores of the true ligands is substantially larger. Any sensible metric for measuring enrichment should report that the DUD self-decoy set in this case is more challenging than the ZINC set. It *should not* matter that the ZINC decoy set contained 1,000 molecules and that the DUD set contained 3,127 molecules. The uncertainty in the overlap of the score distributions comes from the limitation of 85 active ligands. Figure 7 shows equations for three methods of characterizing enrichment: the receiver-operator-characteristic (ROC) curve, the enrichment (or recovery) curve, and the enrichment factor curve. All are computed in a similar fashion. Given a set of scores for actives and decoys, a threshold is varied to divide the molecules into putatively active and inactive. At every threshold, the number of actual actives and falsely predicted actives are tallied, and these values give rise to the curves. The ROC curve plots the true positive rate on the *Y*-axis and the corresponding false positive rate on the *X*-axis. The enrichment plot is quite similar, but it plots the percentage of actives recovered on the Y axis and the proportion of database screened on the *X*-axis. The enrichment factor plot shares its *X*-axis with the enrichment plot, but the *Y*-axis reflects the observed hit rate to the expected one.

ROC Plot

$$y=\frac{L_t}{N_L}$$

$$x=\frac{D_t}{N_D}$$

Enrichment Plot

$$y=\frac{L_t}{N_L}\cdot 100$$

$$x=\frac{L_t+D_t}{N_L+N_D}\cdot 100$$

Enrichment Factor Plot

$$y=\frac{\dfrac{L_t}{L_t+D_t}}{\dfrac{N_L}{N_L+N_D}}$$

$$x=\frac{L_t+D_t}{N_L+N_D}\cdot 100$$

Fig. 7 The three sets of equations define three approaches to analyzing screening utility. We consider different thresholds t to select a subset of compounds. The total number of true ligands is N_L and total number of decoys N_D. L_t is the number of true ligands in the subset, D_t is the number of decoys in the subset, and $(L_t + D_t)$ is the total number selected. Common metrics include the area under the ROC curve, maximum enrichment factor, and enrichment factor at a specific percentage of database screened (e.g., EF1% is the value of y when x is 1 in the enrichment factor plot)

ROC analysis is very well established, having been developed in the first half of the 20th century. The behavior of ROC curves is well understood from a statistical perspective: the ROC AUC for a class predictor is equivalent to the probability that the predictor will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to the Wilcoxon test of ranks and is also closely related to the Gini coefficient. Among the desirable statistical properties of ROC AUC include insensitivity to skew (the relative sizes of the population of actives and inactives). This, and other, formal aspects of ROC analysis are discussed at length by Fawcett [26]. Their practical advantages for quantifying screening enrichment will be discussed here. Figure 8 shows the corresponding ROC plots for the DUD self-decoy set (red, 3,127 molecules), the ZINC decoy set (blue, 1,000 molecules), and the DUD all decoy set (green, 124,413 molecules). As expected, the ZINC decoy set yielded higher early and overall enrichment, and the DUD self-decoy set is clearly more challenging. The DUD all decoy set was only marginally more difficult than the ZINC set. Table 1 shows a number of metrics corresponding to

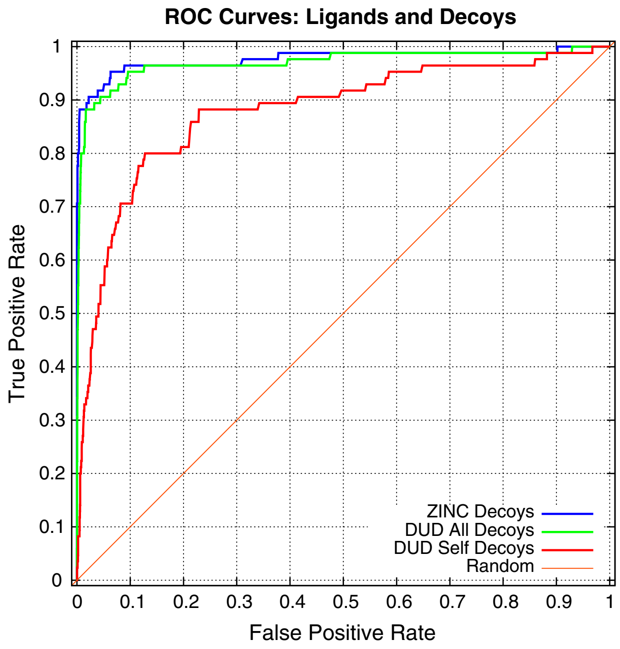


Fig. 8 The ROC plots of Surflex docking performance on PPAR-gamma using three decoys sets make it clear that the DUD self-decoy set (red) is more challenging than the ZINC decoy set (blue) or the DUD all decoy set (green). The latter two are essentially equally challenging

this example. The ROC-based metrics include: area under the ROC curve, its corresponding 95% confidence interval, and true positive rates at three "early" false positive rate values (1%, 5%, and 10%). From these data, it is clear that the DUD self-decoy set is statistically significantly more challenging than either the ZINC or DUD all decoy sets. Further, while we see excellent early enrichment in two cases (80–90% at a false positive rate of 1%), the early enrichment is lower for the DUD self-decoy set. Table 1 also shows the values for other metrics, including EF1% (the enrichment factor at 1% of database screened), EF′(70) [27], RIE (robust initial enhancement [28]), and BEDROC [29]. EF′(70) measures the enrichment for recovering 70% of the known actives, where enrichment is computed based on the average percentile rank of the actives within the entire database [27]. Both EF1% and RIE lead to the *clearly incorrect* conclusion that early enrichment is *better* with the DUD self-decoy set than with the ZINC set, even though measuring early enrichment was the specific motivation for use of these metrics.

Why is this? All of the non-ROC metrics in Table 1 are affected by the ratio of the number of actives to inactives in a screening experiment instead of being invariant to the ratio. EF, EF′, and RIE are all biased to report high values with large decoy sets, and BEDROC is biased to report higher values with smaller decoy sets. In this example, there were 85 actives, which was clearly the limiting factor in estimating the separation of score populations between actives and decoys. Randomly selected sub-populations from the DUD all decoy score population of 10% and 1% were made to empirically test the different metrics' sensitivity to population ratio variation. Table 1 shows all metrics computed for five different 10% sub-populations of the DUD all decoy scores and for five different 1% sub-populations. In each case, the ROC area was 0.97, with a 95% confidence interval of 0.94–0.99. This is a result of the ROC metrics being unaffected by skew. However, EF1%, EF′(70), and RIE 1% all decreased as the population of decoys shrank (RIE by nearly 100-fold), and BEDROC moved in the opposite direction.

The non-ROC metrics are not measuring the separation of populations; they are measuring a confabulation of

Table 1 Performance of Surflex-Dock on PPAR-gamma, comparing multiple metrics for measuring virtual screening utility

Decoy Set	N Decoys	ROC Area	95% CI	TP % at 1% FP	TP % at 5% FP	TP % at 10% FP	EF1%	EF′(70)	RIE1%	BEDROC (alpha 20)
ZINC	1,000	0.977	0.949–0.995	88	92	96	13	12	13	0.959
DUD Self	3,127	0.879	0.830–0.920	26	55	71	25	10	25	0.525
DUD All	124,414	0.970	0.941–0.991	80	91	95	80	236	1,021	0.853
DUD 10%Sample 1	12,432	0.969	0.940–0.991	80	91	95	70	100	112	0.861
DUD 10%Sample 2	12,464	0.969	0.940–0.990	80	91	95	70	92	109	0.856
DUD 10%Sample 3	12,346	0.970	0.936–0.990	80	91	95	70	101	111	0.862
DUD 10%Sample 4	12,299	0.970	0.942–0.99	80	91	95	71	90	107	0.857
DUD 10%Sample 5	12,417	0.970	0.940–0.991	80	91	95	70	94	111	0.864
DUD 1%Sample 1	1,248	0.970	0.941–0.991	80	91	95	16	15	16	0.920
DUD 1%Sample 2	1,232	0.971	0.941–0.992	81	91	95	15	15	15	0.934
DUD 1%Sample 3	1,228	0.970	0.940–0.991	80	91	95	15	15	15	0.924
DUD 1%Sample 4	1,178	0.966	0.939–0.988	71	89	92	15	13	14	0.894
DUD 1%Sample 5	1,256	0.969	0.939–0.990	80	91	95	15	14	14	0.902

EF and RIE both incorrectly suggest that the DUD self-decoy set is *less* challenging than the ZINC set for this test case, which is clearly false population separation with active to inactive sample size ratio. Table 2 shows the average summary results for the three different sample sizes of the DUD all decoy set. Metrics such as RIE report *orders of magnitude* different values when *no distributional effect* exists. This makes comparing results from different studies very difficult since the ratio of actives to inactives in a screening test is highly variable, arbitrary, and unrelated to the screening performance of a method. Figure 9 shows the cumulative histograms for the scores from the PPAR-gamma ligands (green), the DUD self decoys (red), and for the DUD all decoys including all ten sub-populations (cyan, red, and blue). The population characteristics of the 10% sub-population are so similar to the parent population that all five curves *fall within* the thick line of the parent population. Even the 1% populations (of about 1,200 scores) make only small excursions from the parent distribution. Very large decoy sets do not enhance the precision with which we can measure screening enrichment: they are simply a waste of processing power. They add no information about the population characteristics of non-ligands that cannot be estimated from much smaller decoy sets.

The source of the sensitivity to skew for enrichment factor analysis is easily traced back to the equations in Fig. 7. Apart from the constant factor of 100, the ROC plot and enrichment plot are the same for the *Y*-axis and differ only in the *X*-axis. However, as the total number of decoys (N_D) becomes very large relative to the number of actives (N_L), the equation for the abscissa becomes exactly that defined for the ROC plot (since N_L can be ignored and L_t/N_D is very small for any threshold t). The ROC plot is insensitive to skew since it avoids sums that involve *both* the numbers of actives and inactives. The enrichment plot achieves this only asymptotically with very large decoy set sizes. The same analysis applies for the enrichment factor plot, but in this case, the simplification applies to both the *X*-and *Y*-axes. Here, the equation for the ordinate simplifies to the ratio of true positive rate to false positive rate (*y* and *x*, respectively, from the ROC plot). Referring back to Table 2 using all 124,414 DUD decoys, EF1% was 80, exactly matching the ratio of TP rate to FP rate at 1%. As the number of decoys was reduced, the TP/FP ratio at 1% did not change significantly (see also Table 1), but the EF1% value shifted downward dramatically. The downward shift with smaller decoy sets is easily explained. Considering again the equations from Fig. 7, the enrichment factor tends toward one as the number of decoys N_D (and consequently also D_t) tends toward zero. ROC analysis is isomorphic to enrichment factor analysis minus the skew.

The larger goal of methodological evaluation of screening utility is, of course, to communicate relevant information to users that will reflect real-world performance. The ROC-based metrics support quantitative statistical conclusions through the use of confidence intervals. So, a claim that, for example, ligand preminimization improves screening performance can be quantitatively supported [19]. Further, provision of true positive recovery rates at fixed (low) false positive rates addresses two issues. First, it disambiguates the case of an ROC area of, say, 0.8 that can arise with excellent early enrichment *or* with poor early enrichment, which has been a common theoretical objection to ROC area [29]. Also, these values can be directly extrapolated to practical screening. Given, for example, a true positive rate of 80% at a false positive rate of 1% (as with the PPAR-gamma example above using the ZINC or DUD all decoy sets), a user screening a database of 100,000 ligands could conclude that assaying the top 1,000 ranked ligands from a screen would identify the bulk of the actual actives present in the database. There are caveats to this though. This will be true only if the characteristics of actives and decoys in a study reflect the characteristics of the populations of actives and inactives within the database to be screened in practice. All metrics are limited in this fashion though, and the issue of optimal design of active and decoy test sets is complex since their relevance is defined by properties intrinsic to *future* application of a method.

However, the issues around performance metrics for measuring screening utility are less complex. Many of the metrics that have been proposed over the past several years lead to clearly incorrect conclusions in comparing the relative degree of challenge posed by the different decoy sets shown here. Most of these effects are due to skew: the

Table 2 Performance of Surflex-Dock on PPAR-gamma, comparing the effect of randomly subsampling the DUD all decoy set by 10% and 1%

Decoy Set	N Decoys	ROC Area	TP % at 1% FP	TP % at 5% FP	TP % at 10% FP	EF1%	EF′(70)	RIE1%	BEDROC (alpha 20)
DUD All	124,414	0.970	80	91	95	80	236	1,021	0.853
10%Sample Means	12,392	0.970	80	91	95	70	95	110	0.860
1%Sample Means	1,228	0.969	78	90	95	15	14	15	0.915

The ROC-based metrics are essentially invariant, since the population of as few as 1,200 decoys is sufficient to estimate the characteristics of the parent distribution. The other metrics are all susceptible to skew, with EF, EF′, and RIE all favoring large decoy sets and BEDROC favoring small decoy sets

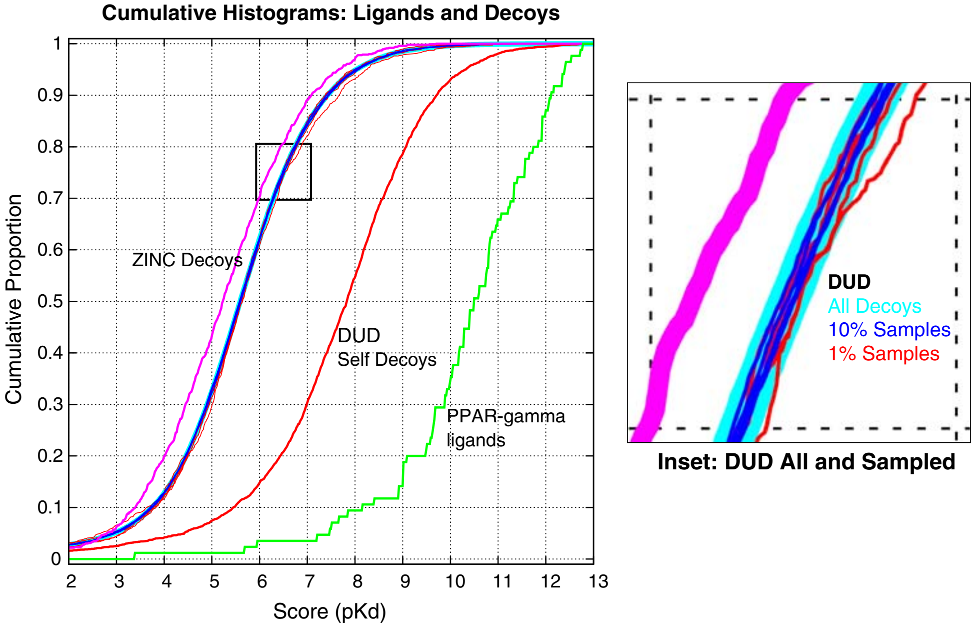


Fig. 9 The plot shows the distributions that underlie all of the performance metrics computed for Table 2. The ZINC decoy score distribution (purple) and DUD self-decoy distribution and PPAR-gamma ligand (green) distribution are all as in Fig. 6, shown in thick lines. The DUD all decoy distribution (cyan) is shown in cyan, and is very slightly higher-scoring than for the ZINC decoys. The inset shows how little the distributions change by making use of much smaller decoy sets. The 10% samplings (thin blue lines: about 12,000 decoys each) are contained *within* the line plotted for all decoys (about 120,000 decoys), and all but one of the 1% samplings (thin red lines: about 1,200 decoys each) are contained within the all decoy line as well. In no case does the subsampling yield significant differences in ROC-based metrics, with the DUD full decoy set and all subsamplings generating ROC 95% confidence intervals of 0.94–0.99

inappropriate dependence on the relative sample sizes used to estimate performance on actives and on inactives. It is likely that methods can be developed that address the early enrichment issue that are sound in this sense. However, straightforward use of ROC curves, with quantified areas and confidence intervals, along with true positive rates at specific false positive rates corresponding to early enrichment avoid the skew problem, are statistically meaningful, and have very intuitive meanings. Since these metrics lead to transparent comparability across studies, they are recommended as the baseline that should accompany any report of virtual screening performance.

Conclusions

Docking is an important and venerable approach for drug design, but this has not translated into standards in the field, either formal or community-based, for the preparation of data sets, sharing of data sets, or statistical reporting in studies of the performance of docking methods. The combination of no requirements for data set sharing with the potential for subtle systematic bias to be present in the structures of proteins and ligands leads to serious problems. In the best case, we may not detect unwittingly embedded systematic bias in a report that makes a conclusion about the relationship between a technique and a performance advance. In the worst case, we create a moral hazard that provides an incentive to exploit such bias effects to create the *impression* of a performance advance. With respect to performance metrics, we see a similar problem. Idiosyncratic metrics substituting for a standardized set of well-vetted metrics leads, in the best case, to reports that are not comparable to other reports. In the worst case, there is an incentive to construct metrics that create the impression of a performance advance where there may be none.

Acknowledgements The author gratefully acknowledges NIH for partial funding of the work (grant GM070481). Dr. Jain has a financial interest in BioPharmics LLC, a biotechnology company whose main focus is in the development of methods for computational modeling in drug discovery. Tripos Inc. has exclusive commercial distribution rights for Surflex-Dock, licensed from BioPharmics LLC.

References

1. 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 (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49(20):912–931
2. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43(25):4759–4767
3. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801

4. Perola E, Walters WP, Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56(2):235–249
5. Miteva MA, Lee WH, Montes MO, Villoutreix BO (2005) Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. J Med Chem 48(19):6012–6022
6. Bohm HJ (1994) 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 8(3):243–256
7. Jain AN (1996) Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des 10(5):427–440
8. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) 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 11(5):425–445
9. Jain AN (2006) Scoring functions for protein-ligand docking. Curr Protein Pept Sci 7(5):407–420
10. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 295(2):337–356
11. Muegge I, Martin YC (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem 42(5):791–804
12. Welch W, Ruppert J, Jain AN (1996) Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chem Biol 3(6):449–462
13. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261(3):470–489
14. Jain AN (2004) Virtual screening in lead discovery and optimization. Curr Opin Drug Discov Devel 7 (4):396–403
15. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749
16. Goodsell DS, Morris GM, Olson AJ (1996) Automated docking of flexible ligands: applications of AutoDock. J Mol Recognit 9 (1):1–5
17. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267(3):727–748
18. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161(2):269–288
19. Jain AN (2007) Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. J Comput Aided Mol Des 21 (5):281–306
20. Irwin JJ, Shoichet BK (2005) ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model 45(1):177–182
21. Pham TA, Jain AN (2006) Parameter estimation for scoring protein-ligand interactions using negative training data. J Med Chem 49(20):5856–5868
22. Mayo SL (1990) DREIDING: A Generic Force Field for Molecular Simulations. J Phys Chem 94 (26):8897–8909
23. Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 46(4):499–511
24. Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP (2005) Comparison of automated docking programs as virtual screening tools. J Med Chem 48(4):962–976
25. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J Med Chem 49 (21): 6177–6196
26. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874
27. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47(7):1750–1759
28. Sheridan RP, Singh SB, Fluder EM, Kearsley SK (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. J Chem Inf Comput Sci 41(5):1395–1406
29. Truchon J-F, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. J Chem Inf Model 47(2):488–508