


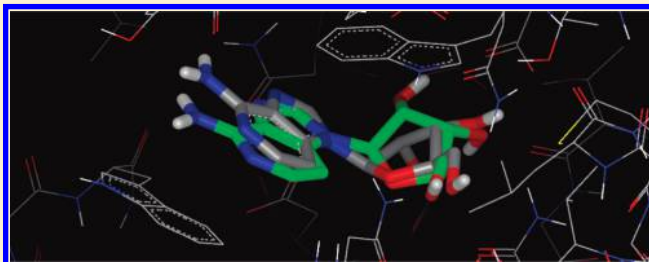
FRED Pose Prediction and Virtual Screening Accuracy

Mark McGann^{*,†}

OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, New Mexico 87508, United States

 Supporting Information

ABSTRACT: Results of a previous docking study are reanalyzed and extended to include results from the docking program FRED and a detailed statistical analysis of both structure reproduction and virtual screening results. FRED is run both in a traditional docking mode and in a hybrid mode that makes use of the structure of a bound ligand in addition to the protein structure to screen molecules. This analysis shows that most docking programs are effective overall but highly inconsistent, tending to do well on one system and poorly on the next. Comparing methods, the difference in mean performance on DUD is found to be statistically significant (95% confidence) 61% of the time when using a global enrichment metric (AUC). Early enrichment metrics are found to have relatively poor statistical power, with 0.5% early enrichment only able to distinguish methods to 95% confidence 14% of the time.



■ INTRODUCTION

Structure-based computer modeling methods, such as molecular docking programs, use the structure of a target protein for lead optimization and hit discovery in the drug development process. In hit discovery, no molecules active against the target protein are known, and docking is used to identify molecules in a large virtual database that are likely to be active by examining their shape and chemical complementarity to the active site. This process is known as virtual screening. The database can be a corporate collection, vendor database, or set of potentially synthesizable molecules. Because virtual screening is done *in-silico*, the only factor limiting the size of the database is the amount of CPU time available (a few million molecules is a realistic upper limit with a fast docking program and large cluster). Once the screen is complete, the top scoring molecules can be examined and a subset selected for further testing. The fraction of top scoring molecules that is active depends on the number of active molecules in the database overall and the effectiveness of the docking program. In lead optimization, one or more ligands active against a target protein are known, but they have insufficient binding affinity or other undesirable properties, such as cross reactivity or conflicts with existing patents and therefore must be modified. Docking is used to predict how potential modification will affect both the binding mode (pose) and binding strength of a ligand. To be effective at hit discovery and lead optimization, a docking program must be able to correctly place a ligand within the receptor site and estimate its binding affinity.¹

Ligand-based methods are a different class of virtual screening methods that use the information present in a known active ligand rather than the structure of a target protein for both lead optimization and hit discovery. Ligand-based methods often outperform structure-based methods (e.g., docking programs)

in retrospective virtual screening tests at a fraction of the CPU cost.² However, it has been shown that most virtual screening test data sets are unintentionally biased to favor ligand-based methods over structure-based methods.³ Also, structure-based methods can detect ligands that dock in a variety of binding modes. In contrast, ligand-based methods can only detect actives that bind in the same mode as the query ligand because the query ligand contains no information about binding modes other than its own. Two-dimensional ligand-based methods, such as Daylight Fingerprints⁴ or LINGOS,^{5,6} are further limited to detecting actives with chemical graphs similar to the query molecule (3-D ligand-based methods, such as ROCS^{7,8} and BROOD,⁹ are not limited in this way). Structure-based methods are thus in principle capable of finding actives with more diverse properties than ligand-based methods.¹⁰ Diversity is important because it is easier to find a molecule without undesirable properties, such as toxicity or cross-reactivity, in a set of diverse actives than it is in the same sized set of nondiverse actives (nondiverse actives will tend to all have the same undesirable properties).

While all docking programs use the structure of the target protein to detect active molecules, they differ greatly in the details of their implementation. Two of the basic parts of any docking program are the docking algorithm that creates trial poses within the active site and the scoring function that evaluates the fitness of each pose and the relative fitness of different ligands. Docking algorithms can be classified by how many degrees of freedom they have and how those degrees of freedom are searched. Most docking programs allow the ligand, but not the protein, to flex during docking. Methods of searching include (1) exhaustive searching, which systemically samples all possible poses (to a

Received: November 3, 2010

Published: February 16, 2011

given resolution); (2) stochastic searching, which randomly samples poses within the site; (3) anchor and grow, which orients an initial anchor fragment of the molecule and flexibly grows the remaining pieces of the molecule into the site, pruning any partially built solutions that are unlikely to score well; and (4) score optimization, which takes an initial ligand pose and drives it to an optimum of the scoring function (e.g., quasi-newton optimization). Scoring functions examine the shape and chemical complementarity of a pose with the active site and assign it a score. Ideally, ranking by score yields the identical ordering as ranking by binding energy, although in practice this is rarely the case. Common scoring function terms are shape (e.g., van der Waals interactions), hydrogen bonding, desolvation, electrostatics, and entropy. The implementation of scoring functions varies greatly from program to program, and even a publically available scoring function is often implemented in subtly different ways.

Given the potential usefulness of docking and the wide variety of methods available, standardized methods of evaluation are important.¹¹ Results of a docking study should be reproducible by the scientific community at large, and therefore, testing should be done using the officially released versions of software and publically available data sets. Recently published work by Cross et al.¹² uses publically available data sets to compare the performance of several molecular docking programs (Dock,^{13,14} FlexX,¹⁵ Glide,¹⁶ ICM,^{17,18} PhDock,^{19,20} and Surflex^{21–24}). This work extends the work of Cross et al.¹² to include the following:

- 1 The statistical confidence that the result on the retrospective test data sets will be observed on prospective data sets.
- 2 The consistency of each docking method (i.e., how much the results vary from target to target).
- 3 The probability that one method will outperform another on systems other than the ones in the test data set.
- 4 The results from the docking program FRED using the same publically available data sets.
- 5 Null hypothesis testing for virtual screening by randomizing the target protein to which the ligands in the test data set are docked (the hypothesis being that the structure of the protein is important to virtual screening). The null hypothesis test results are reported for FRED but not for other programs because the necessary data for the analysis were not available.

This work also examines the results of using the information present in the structure of a bound ligand to enhance docking performance. Most virtual screening programs make use of ligand-based or structure-based information exclusively. However it is common for both structure-based and ligand-based information to be available. For example, protein structures used for docking are often crystallized in the presence of a known binder, but the standard operating procedure for most docking programs is to remove the ligand from the site before docking (conversely, ligand-based approaches discard the protein structure). FRED is capable of using the bound ligand to guide the docking (although in its default mode it is a pure docking program). This approach blurs the line between ligand- and structure-based methods, and we refer to it as hybrid docking. (A related approach using ligand information to guide the selection of the receptor conformer to dock also was recently published by Lee et al.²⁵) Results of this hybrid docking approach are presented herein.

THEORY

FRED Docking. *Overview.* FRED docks molecules using an exhaustive search algorithm that systematically searches rotations and translations of each conformer of the ligand within the active site at a specified resolution. During the exhaustive search, unrealistic poses are filtered, and those that survive are scored. Following the exhaustive search, the 100 top scoring poses are subject to systematic solid body optimization (a local exhaustive search at a finer resolution than the global exhaustive search). The best scoring pose is then used to rank the ligand against other ligands in the screening database. The protein is held rigid during the docking process, as are the conformers of the ligand. Ligand flexibility, however, is implicitly included by docking a conformer ensemble of each molecule. A schematic of the FRED docking process is shown in Figure 1.

Conformer Generation. FRED treats each conformer of a molecule as rigid during the docking process, although the docking process is effectively flexible with respect to the ligand because multiple conformers of each ligand are docked into the site. Generating conformers prior to running FRED reduces run time (because conformation generation is done independent of the active site, it needs only to be done once for any given ligand database, rather than for every docking run). Any conformer generation program can be used to generate conformers for FRED. Within the OpenEye workflow, the program OMEGA^{26,27} is used to generate ligand conformers prior to docking. OMEGA uses torsion and ring libraries to identify and enumerate rotatable bonds and flexible rings. Conformers with internal clashes or high strain are then discarded. The remaining low strain conformations are then clustered on the basis of rmsd, and the cluster centers are retained, while all other conformers are discarded. Finally, if the number of remaining conformers exceeds a specified maximum number, then the conformers with the lowest strain energy are retained. By default, OMEGA uses a clustering threshold of 0.5 Å, and the maximum number of conformers generated is 200.

Exhaustive Search. Exhaustive search is the core of FRED's docking algorithm and is responsible for rapidly transforming ligand conformers into poses within the active site (a pose is a particular structure or arrangement of the ligand atoms within the active site), filtering bad poses, and scoring those that survive. The search is termed exhaustive because, to a specified resolution, all possible translations and rotations of each conformer in the active site are enumerated. A unique feature of the exhaustive search is that the ensemble of poses that is initially generated by rotating and translating each conformer is independent of the scoring function (i.e., two different scoring functions will score the exact same set of poses when using the exhaustive search). This makes the exhaustive search a useful tool for designing and comparing scoring functions because sampling and scoring issues can be examined separately.

The exhaustive search begins by enumerating every possible rotation and translation of a ligand's conformers within the active site to a default translational and rotational resolution of 1 and 1.5 Å, respectively. The rotational resolution is the maximum distance any atom will move in a single rotational step; because there is a lever arm effect in a rotational step, most atoms will move significantly less than the maximum distance. The number of poses enumerated is dependent on the shape of the active site, the number of conformers, and the shapes of the conformers; however, the number of poses per molecule is typically in the

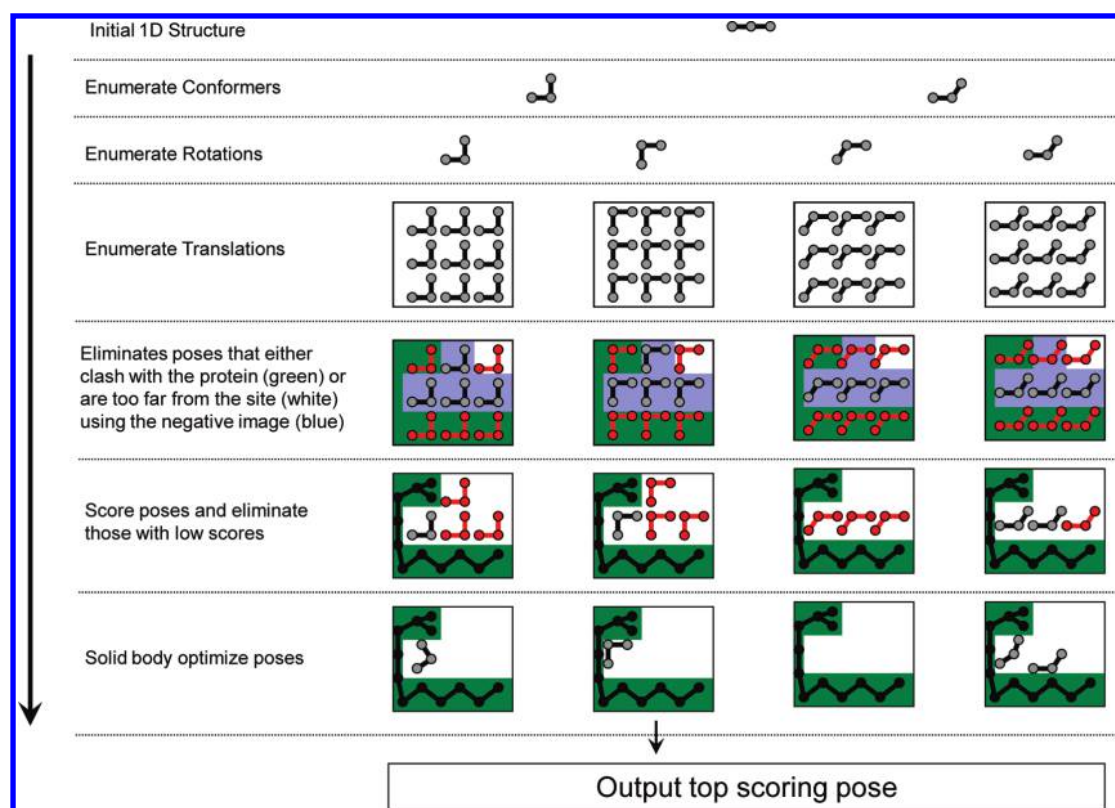


Figure 1. Schematic of the FRED docking process.

range of one hundred thousand to one hundred million. Once the initial pose ensemble is generated, a negative image of the receptor site (described below) is used to filter out poses that clash with the protein or extend too far from the active site. Typically, from ten thousand to ten million poses survive the filtering process (i.e., a 1 order of magnitude reduction). The final step of the exhaustive search is to score the remaining poses. The scored poses of all conformers are sorted by score, and by default, the top 100 are sent to the next step in the docking process: optimization.

FRED uses a shape that complements the receptor site, called the negative image, as a quick filter to determine which poses in the initial exhaustive search ensemble can be discarded because they clash with the protein or are too far from the active site. The negative image of the active site is created by contouring a shape potential field that complements the active site and is constructed as follows:

- 1 A set of molecular probes (molecules with a single conformer) representing common shapes of drug-like molecules are docked into the active site, using a modified FRED algorithm that does not require a negative image of the site.
- 2 Poses of each molecular probe are scored using the Gaussian Shape Scoring Function.²⁸
- 3 Top scoring poses of each probe are converted into density fields that are then averaged together to form the final shape potential field. To convert a pose into a density field, each atom is treated as having a spherical Gaussian density distribution, and the weight of each pose is scaled by the pose's overall rank and shape score.

The shape potential field that is created has high values at positions where ligand atoms make many contacts with atoms

of the receptor site without clashing and at positions some ligand atoms are likely to occupy when other atoms of the ligand make good contacts with the receptor (e.g., bridging positions that ligand atoms will likely occupy when a ligand is stretched between two pockets).

Optimization and Pose Selection. Following the exhaustive search, the top 100 scoring poses from the exhaustive search are optimized using a systematic solid body optimization, and the top scoring pose is selected as the final docking pose. This optimization is a local exhaustive search at half the resolution of the full exhaustive search (thus, by default the rotational and translational step size are 0.75 and 0.5 Å, respectively). Each pose to be optimized takes a positive and negative step for each rotational and translational degree of freedom, and all combinations are enumerated, for a total of 729 nearby poses (six degrees of freedom with three positions yields 3^6 or 729 poses), and the best scoring pose is retained in place of each original pose).

FRED Scoring. Overview. FRED uses scoring functions at three places during the docking process: the exhaustive search, optimization, and final scoring. The scoring functions available in FRED are Shapegauss,²⁸ PLP,²⁹ Chemgauss, Chemscore,³⁰ Screenscore,³¹ Chemical Gaussian Overlay (CGO), and Chemical Gaussian Tanimoto (CGT). In this work, however, only the results using default scoring (Chemgauss) and hybrid scoring (Chemgauss and CGO) are reported in the primary results. Both Chemgauss and CGO scoring are described below.

Chemgauss. The Chemgauss scoring function uses Gaussian-smoothed potentials to measure the complementarity of ligand poses within the active site, and it is the default scoring function of FRED v2.2.5. Chemgauss that recognizes shape interactions, hydrogen bonding interactions with the protein, hydrogen bonding interactions with implicit solvent, and metal–chelator

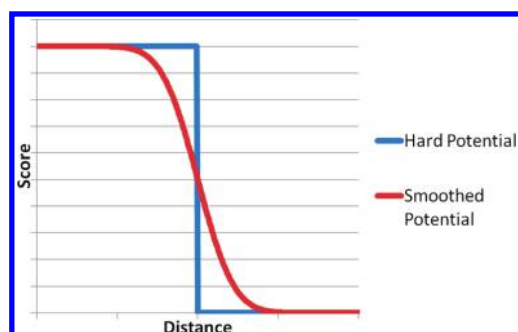


Figure 2. Example of a smooth potential (red) generated by convoluting a hard potential (blue) with a Gaussian.

interactions. All interaction potentials in Chemgauss are initially constructed using step functions to describe the interaction of atom pairs (or other chemical points) as a function of distance. These interactions are mapped onto a grid that is then convoluted with a spherical Gaussian function, which smooths the potential (Figure 2), making it less sensitive to small changes in the ligand position. Smoothing the score in this way serves two purposes: First, the exhaustive search can be run at a lower resolution than would be required if the score were not smooth because small changes in position do not cause large changes in score. Second, smoothing makes the function more tolerant of small deviations in the protein structure,²⁸ which should reduce the error associated with the rigid protein approximation by accounting for the ability of the protein to make small structural rearrangements to accommodate the ligand.

Shape interactions in Chemgauss are based on a united atom model (i.e., only the coordinates of the heavy atoms centers are relevant to the shape calculation). Each ligand heavy atom is assigned a fixed clash penalty score if the distance between it and a protein heavy atom is less than the sum of the van der Waals (VdW) radii; otherwise, it is assigned a score proportional to the count of the number of protein heavy atoms within 1.25 and 2.5 times the sum of the VdW radii (atoms within 2.5 count one tenth as much as those within 1.25). For each ligand-heavy atom, a penalty equal to two close protein atom contacts is subtracted to represent the VdW interactions with solvent water that are lost when the ligand docks. This score is precomputed at grid points throughout the active site, and the resulting grid is then smoothed as described above.

Hydrogen bonding groups are modeled with one or more lone-pair or polar-hydrogen positions that describe the directionality of potential hydrogen bonds (with respect to the hydrogen bonding group's heavy atom). Donor groups have lone pair positions that represent the possible location of the donor hydrogen atoms relative to the donating molecule, while acceptors have lone-pair positions representing the possible locations of the donated hydrogen relative to the acceptor. A hydrogen bond is detected and assigned a constant score when a hydrogen bonding position on the ligand is within 1.0 Å of a complementary hydrogen bonding position on the protein (i.e., when the polar-hydrogen position of a donor overlaps the lone-pair position of an acceptor). If the ligand hydrogen bonding group has multiple polar-hydrogen and/or lone-pair positions (groups can be both donors and acceptors), then this calculation is performed for each position and the result is summed. As with all Chemgauss terms, the hydrogen bond potential is precomputed at grid points throughout the site and then smoothed.

Hydrogen bonds with solvent that break when the ligand docks into the active site are penalized by the Chemgauss scoring function. Broken protein–solvent hydrogen bonds are accounted for by calculating how many hydrogen bonds water can make with the protein at the position of each heavy atom of the docked ligand, and the penalty score assigned is proportional to the number of hydrogen bonds. Broken ligand–solvent hydrogen bonds are accounted for by calculating desolvation positions around each hydrogen bonding group on the ligand that represent the positions water could occupy when making a hydrogen bonding interaction with the protein. A penalty is then assessed that is proportional to the number of desolvation positions that can no longer be occupied by water because the water in these positions would clash with the protein. As before, this potential is placed on a grid and smoothed.

Chelating interactions between protein metals and ligand chelating groups are accounted for by Chemgauss (protein–chelator and ligand–metal chelating interactions are not). For each chelator on the ligand, one or more chelating positions are calculated. If a protein metal is within 1.0 Å of any chelating position of a chelating group, then a fix score is assigned; otherwise a zero score is assigned. As before, this potential is placed on a grid and smoothed.

Chemical Gaussian Overlay (CGO). Chemical Gaussian Overlay (CGO) function scores by measuring how well a pose overlays the structure of a bound ligand within the active site. The overlay accounts for the overall shape of the molecules as well as the position of hydrogen bonding and metal chelating groups. This scoring function requires a bound ligand pose as well as the structure of the target protein (for reasons described below). Typically, the ligand structure is obtained from X-ray crystallography along with the structure of the target protein, although a docked ligand could also in principle be used.

CGO represents molecules as a set of spherical Gaussian functions describing their shape and chemistry (acceptors, donors, and chelators). The Gaussians representing the shape of the molecule are centered at the heavy atom positions: those for donors are centered on polar-hydrogen positions (i.e., positions where the donating hydrogen could be when it is involved in a hydrogen bond), those for acceptors are centered on lone-pair positions (i.e., positions where a donating hydrogen could be when a hydrogen bond is formed), and those for chelators are centered at chelating positions (i.e., locations where a metal could have a chelating interaction). The overlap of the Gaussians on the docked ligand to those on the bound ligand is computed for each type of Gaussian (e.g., shape, donor, acceptor and chelator) by summing the overlap of individual pairs of Gaussians. The overlap of each individual pair is calculated by integrating the product of the two. To prevent chemistry not relevant to binding from contributing to the overall score, when calculating the chemistry overlaps (i.e., acceptor, donor, and chelator), only groups that make the interaction with the protein are considered (e.g., a chelator that does not interact with a metal on the protein is ignored in the overlap calculation). Thus, while CGO is primarily a ligand-based scoring function, it indirectly uses some information from the protein structure. The sum of all four types of overlaps is the CGO score.

■ EXPERIMENTAL SECTION

Test Sets. This work is based on the results obtained by Cross et al.¹² with several docking programs and two retrospective data

sets. The first data set is used to determine the ability of the docking programs to ascertain the correct pose of a ligand within the active site, while the second measures the virtual screening performance of the docking program. In addition to a reanalysis of these results, this work also includes the result from the FRED docking program using the same test data sets.

The structure reproduction data set used by Cross et al.¹² contains 68 protein–ligand complexes, 64 of which are drawn from the PDB and 4 of which are listed as in-house. The structure reproduction results and analysis presented herein are based on the results of the 64 publicly available protein–ligand complexes, which we will refer to as the Cross2009 data set. To test this data set, each ligand is removed from the protein active site and given to a docking program that redocks the ligand into the site, and the docked structure is then compared to the known experimental structure. Assessing the pose prediction performance of a docking program using a self-docking test set is of limited scientific value because the protein structure is fixed in the perfect conformation for the ligand to dock. In real world research settings, the structure the protein adopts in the presence of a ligand is not known *a priori*, and docking programs using the rigid protein approximations (e.g., all docking methods examined herein) do not even attempt to find the correct protein structure. Generally, the protein structure used in prospective research settings is from an X-ray crystallography experiment in which the protein was crystallized in the presence of a single ligand. Thus, self-docking represents a highly idealized experimental condition that is never seen in prospective research settings. Nevertheless, self-docking results are commonly reported, and we do not have access to cross docking results on publicly available data sets for the docking programs tested by Cross et al.¹² Therefore, we report and analyze self-docking results in this work.

The DUD³² data set is a publicly available virtual screening test set and is the second data set used by Cross et al.¹² and this work. It consists of 40 protein targets each with a set of molecules known to be active against the protein target and 36 decoys for each active ligand. The decoys were selected to have physical properties similar to the active such as molecular weight, cLogP, and number of hydrogen bonding groups, but different topologies. A potential deficiency of the DUD data set is that many actives are close analogs with very similar physical properties and topologies. However, DUD remains the largest publicly available virtual screening data set, and it is difficult to obtain statistically significant results in smaller data sets.

In addition to the standard virtual screening tests on the DUD data set, this work also includes results for FRED in which the protein targets of the DUD data set have been randomized. This testing is done to detect any bias that may be present in the DUD data set that allows actives to be distinguished from decoys independent of the receptor site. The pairings of actives and decoys to protein target are listed in the Supporting Information. Pairings are not strictly random. Some attempt was made to ensure that the ligands were docked into a receptor site of roughly the same size as the correct receptor and to ensure that the new receptor was not of the same class as the original.

FRED Setup. Receptor and Ligand Preparation. A receptor file is a specialized file used by FRED that describes the active site. It contains the structure of the target protein, the location of the active site (described by a box), the negative image of the active site used by the exhaustive search (see the Theory Section for a description of the negative image), and optionally the structure of a bound ligand (a bound ligand is required to use the CGO

scoring functions). The receptor file can be set up either interactively using a GUI supplied with the FRED program or automatically by the FRED command line using the structure of a target protein, together with a box enclosing the active site or a bound ligand (or both). Receptors set up for the Cross2009 data set used the automatic method with a box and protein. The box enclosing the active site was determined by creating an initial box around the bound ligand and then extending each side of the box by 4 Å (this technique produces boxes that are reasonable on visual inspection without any subjective element to the box creation). Receptors for the DUD virtual screening targets were set up by hand using the interactive GUI.

Ligand conformers were generated with OMEGA version 2.3.2 prior to running FRED. Ligands from the Cross2009 structure reproduction set were extracted from their PDB complex using an interactive GUI and converted to SMILES before being passed to OMEGA; thus, the ligands have no memory of their correctly docked conformation. DUD ligands were downloaded directly from the DUD Web site³³ and converted to SMILES before being passed to OMEGA. Many of the DUD ligands downloaded from the DUD site have enumerated tautomer states. In our virtual screening analysis, only the highest scoring tautomer is included in the final ranked list used to calculate AUC and early enrichment.

Search Resolution and Scoring Functions. The docking algorithm and scoring function are the primary scientific characteristics of any docking program. In this work, we present the results from FRED version 2.2.5 using the default and a hybrid scoring method and using the default and a high resolution docking algorithm. Thus, there are four modes of operation for FRED (Table 1) because the docking algorithm and scoring function can be modified independently (a list of the OMEGA and FRED options used for each mode are listed in the Supporting Information). The “FRED CG” mode listed in Table 1 is the default “out of the box” method for FRED. Docking times vary by system, but on average, docking times per ligand on a single CPU (2.4 GHz Xeon/Linux) are shown in Table 1 (also see Figure II in the Supporting Information).

The FRED scoring methods examined herein are the default and a hybrid scoring method. The default scoring method uses the Chemgauss scoring function for the exhaustive search, optimization, and final scoring (CG modes in Table 1). The hybrid scoring method modifies the default scoring method by using the Chemical Gaussian Overlap (CGO) scoring function for the exhaustive search, while still using Chemgauss for optimization and final scoring (the hybrid modes in Table 1). The CGO scoring function is primarily a ligand-based scoring function that favors poses that match the shape of a known bound ligand and make the same hydrogen bonding and metal chelating interactions with the protein. The term hybrid is chosen because this scoring method uses both ligand and protein information to screen molecules, and thus, it is a hybrid of ligand-based and structure-based virtual screening methods. Hybrid method results are not included in the structure reproduction in the figures and tables because the hybrid scoring function makes direct use of the structure of the bound ligand being searched, which would not be possible in a prospective study. Hybrid results are reported in the virtual screening because different ligands are being docked by the bound ligand used by CGO, and thus is predicative of prospective virtual screening work.

FRED’s docking algorithm is tested in default and high resolution modes. The docking algorithm is an exhaustive search that

Table 1. FRED Methods^a

name	docking function	scoring function	search resolution	CPU time/compound
FRED CG	Chemgauss	Chemgauss	default	~5 s
FRED HYBRID	Chemical Gaussian Overlay	Chemgauss	default	~1 s
FRED-HR CG	Chemgauss	Chemgauss	high	~50 s
FRED-HR HYBRID	Chemical Gaussian Overlay	Chemgauss	high	~20 s

^a “Docking function” is the scoring function that is used during the exhaustive search. “Scoring function” is the scoring function used to optimize the top poses from the exhaustive search and to compare one ligand to another.

Table 2. Methods Examined by Cross et al.¹²

method	program version	notes
DOCK ^{13,14}	6.1	
FlexX ¹⁵	2.0.3	
Glide HTVS	4.5	Glide in high throughput virtual screening mode
Glide SP	4.5	Glide in standard precision mode
Glide XP ¹⁶	4.5	Glide in extra precision mode
ICM ^{17,18}	3.5	
PhDock ^{19,20}	N/A	
Surflex ^{21–24}	2.1	Surflex-Dock without ring-flexing
Surflex-Ring	2.1	Surflex-Dock with ring-flexing

Table 3. Metrics Used To Evaluate Docking Performance

metric	type	description
rmsd	structure reproduction	Root mean squared displacement of docked poses heavy atoms compared to the crystallographically determined structure.
AUC	virtualScreening	The area under the Receiver Operator Characteristic (ROC) curve. Also the probability that an active will score better than a decoy.
ROC(0.5%)	virtual screening	The value of the ROC curve (i.e., the fraction of actives recovered) at 0.5% decoys recovered.
ROC(1%)	virtual screening	The value of the ROC curve (i.e., the fraction of actives recovered) at 1.0% decoys recovered.
ROC(2%)	virtual screening	The value of the ROC curve (i.e., the fraction of actives recovered) at 2.0% decoys recovered.
ROC(5%)	virtual screening	The value of the ROC curve (i.e., the fraction of actives recovered) at 5.0% decoys recovered.
ROC(10%)	virtual screening	The value of the ROC curve (i.e., the fraction of actives recovered) at 10% decoys recovered.

tests all possible placements of a ligand within the active site to a given resolution. The resolution of the docking algorithm is defined by the rotational and translation step size of the exhaustive search, the number of poses passed from the exhaustive search to the optimization step, and the resolution and number of the conformers generated by OMEGA. The default resolution setting for OMEGA is a maximum of 200 conformers, and the rmsd duplicate removal threshold is 0.5 Å rmsd; those for FRED are a translational step size of 1.0 Å, a rotational step size of 1.5 Å, and 100 poses optimized. In high resolution mode, OMEGA retains a maximum of 1000 conformers and a duplicate removal threshold of 1/3 Å, and FRED uses a translational step size of 1.0 Å, a rotational step size of 1.0 Å, and optimizes 1000 poses.

Docking Programs Tested by Cross and Co-Workers. Cross et al.¹² tested the structure reproduction and virtual screening performance of six docking programs, two of which were tested in more than one run mode (Table 2). The reanalysis of these results presented herein requires results from each individual target of the Cross2009 and DUD data sets. These data were not included in the published article by Cross et al.;¹² however, the authors have graciously supplied these values, which we have included in the Supporting Information. Cross' published results included structure reproduction for Glide XP. However, they did not include virtual screening results for Glide XP, and hence, Glide XP's virtual screening results are not reported here.

Ideally all structures in the Cross2009 data set should be new to the programs tested. However, we note that the this data set shares roughly half of its structures with the Astex-Gold structure reproduction data set.³⁴ FRED was not optimized to give the best possible results on the Astex-Gold data set, but consideration was given to the structure reproduction success rate on the Astex-Gold data set when the parameters for FRED were chosen. We would prefer to use a completely new data set to test structure reproduction. However, when comparing to other docking programs, we are constrained to the test data sets in which there is data for the other docking programs. We would also not be surprised, given the high profile of the Astex-Gold data set, that some other docking programs besides FRED have been similarly indirectly trained on the Cross2009 data set.

Metrics. Overview. This work uses the same set of metrics used by Cross et al.¹² to measure virtual screening and structure reproduction performance. These metrics were chosen in order to directly compare FRED's performance to the programs analyzed by Cross. We note that this set of metrics also includes the metrics recommended by a recent paper covering the pros and cons of different metrics.¹¹ The set of metrics is shown in Table 3.

These metrics were computed for FRED for each target in the respective structure reproduction or virtual screening data set. This set of FRED docking results was then combined with the

original Cross results, and the combined set of data was then reanalyzed as described in the Analysis of Results section.

Note that because of the sheer volume of data, the primary figures and tables describing virtual screening results only include AUC and ROC(1%) results. Results for the other early enrichment values are discussed herein, and the complete set of figures and tables for the early enrichment results are included in the Supporting Information.

Issues with rmsd. This work uses rmsd as a measure of the quality of a docked pose. A docked pose close to the reference structure will have a low rmsd value, while high values indicate a pose that is far from the reference structure and hence incorrectly docked. Unfortunately, while the formal range of rmsd is $[0, \infty]$, the range of meaningful values is much smaller, typically $[0.5 \text{ \AA}, 2.0 \text{ \AA}]$. Large rmsd values lack significance because they represent failed dockings, but the degree of failure is not particularly relevant to an evaluation of docking accuracy. For example, one might argue that 12 Å is a more incorrect docking than 10 Å, but what matters is only that both dockings are incorrect. In practice, this makes the reporting of mean rmsd of dubious value. For example, take two docking methods A and B that obtain rmsds of (0.5, 10, 10, and 11.5 Å) and (6, 6, 6, and 6 Å), respectively, on a set of four trials. Method A is clearly the better method because it docked one structure reasonably correctly (0.5 Å), while method B docked none correctly. However, if we were to naively compare the mean rmsd of methods A (8 Å) and B (6 Å), we would incorrectly conclude that method B is the best method.

Very low rmsd values are also of limited value because little scientific inference can be drawn from rmsd values lower than the experimental error of the structure. Coordinate precision is the most relevant measure of experimental error in the X-ray structure of the ligand (both the original Cross et al.¹² study and this work use the raw crystallographic coordinates for the protein structure). The mean coordinate error of the Cross2009 data set is approximately 0.3 Å; thus, rmsd values lower than 0.3 Å are not necessarily better than rmsd 0.3 Å. The ability to draw scientific inference does not abruptly stop at the experimental error; rather, it slowly decreases as the experimental error is approached and crossed. However, in practice, it simplifies analysis significantly to simply choose a hard cutoff and consider any rmsds lower than the cutoff value to be equivalent to the cutoff value. This approximation is reasonable given that in prospective docking experiments the ligand will not be docked into the cognate structure of the protein, and the rigid protein approximation will likely make it impossible to achieve docking accuracy on the order of the experimental error.

Because of these issues with both large and very small rmsd values, we do not report average rmsd values but rather the fraction of systems that docked with less than a given rmsd cutoff value. The cutoff values used are 2.0, 1.5, 1.0, and 0.5 Å.

Analysis of Results. Overview. The metrics described in the previous section are measurements of performance on individual systems of a test data set (either structure reproduction or virtual screening). This section describes how these sets of metric data are analyzed. The goal of this analysis is to answer the following questions with reasonable statistical rigor.

- 1 What is the expected mean performance of a given method?
- 2 What is the consistency of a method?
- 3 What is the probability that one method will outperform another on average?

- 4 What is the probability that one method will outperform another on any given system?
- 5 Do different methods tend to perform well or poorly on the same system?

Questions 1 and 2 are questions of absolute performance, i.e., how well does this program perform independent of the performance of other programs. Questions 3, 4, and 5 are questions of relative performance, i.e., how does one program compare to another.

Absolute Performance. Absolute performance is measured both in terms of expected mean performance on new systems (Question 1) and how well a docking method is expected to do on an individual new system or single trial performance (Question 2). True mean performance is the average performance that would be obtained if all possible systems were tested, while consistency is a measure of how much performance varies from system to system and is described as a distribution of likely values.

The estimate of true mean performance (Question 1) is reported as a 95% confidence interval for both structure reproduction and virtual screening metrics (Table 3). The measured mean performance on the retrospective data set, known as the sample mean in statistics, always falls within this confidence interval and is the most likely value of the true mean (i.e., the mean performance that would be obtained if the set of all possible systems were tested). The size of the confidence interval is a function of the consistency of the method and the number of test systems in the data set. Highly consistent methods tend to get the same result (with respect to a given metric) on all systems, and thus, it is easier to estimate the true mean performance for these methods (i.e., the confidence interval will be smaller for highly consistent methods if all other factors are equal). The size of the confidence interval is also inversely proportional to the square root of the size of the data set. Thus true mean performance will be more accurately estimated with larger data sets than smaller ones.

The consistency (Question 2) of virtual screening is also reported as a 95% confidence interval. The outcome (as measured by a given metric) of a single virtual screen is expected to fall within this range 95% of the time. Because drug discovery focuses on one target at a time, this confidence interval is highly relevant to computational chemists in drug discovery settings. The consistency of a method is an intrinsic property of the method and is not a direct function of the size of the data set. While consistency will be more accurately estimated with large data sets, the size of the confidence interval does not approach zero as the size of the data set increases (in contrast, the confidence interval for the true mean performance does). Also the confidence interval for the outcome of a single trial will always be larger than the confidence interval estimating the true mean performance ($\sigma_{\text{mean}}^2 \cong \sigma_{\text{trial}}^2/N$).

The consistency structure reproduction performance (Question 2) is not reported because structure reproduction is evaluated by measuring if a docked pose is less than a given rmsd. Thus, there are only two possible outcomes for a single trial, success or failure, rather than a continuous range of outcomes; and hence, a 95% confidence interval is always success *and* failure (unless the fraction of successes is greater than 95% or less than 5%, in which case the range is success *or* failure, respectively). These results do still have a variance, which is underlying measure of consistency, that can be calculated directly from the fraction of successful dockings (variance = $p(1 - p)$, where p is the fraction of successful dockings).

Relative Single Trial Performance. Absolute performance measures are useful for evaluating how effective a method is. However what is often of more interest is the relative performance of two or more methods (e.g., how does Method A compare to Method B). Like absolute performance, relative performance can be measured in terms of both how consistent relative performance is and what the mean relative performance is. Herein, relative performance is measured as the probability that one method will do better than all other competing methods on average (Question 3) and on a single system (Question 4).

The relative performance is measured as a probability because the outcome of a method on a new system is not known exactly but rather is described by a distribution of likely results. A probability distribution function (PDF) describes the likely outcomes (where the outcome is the value of the metric used to measure performance) when the method is used, and by testing on retrospective data sets this distribution can be estimated. If the results are discrete PDF(x) is the probability that a result of x will be obtained, where x is the metric used to measure performance. If the results are continuous PDF(x) is the probability density at x , and $\int_{x_1}^{x_2} \text{PDF}(x) dx$ is the probability of getting a result between x_1 and x_2 . Given two methods, A and B, each with a given PDF and a metric spanning the range $[x_{\min}, x_{\max}]$, the probability P that A will be better than B on a new system is

$$P = \int_{x_{\min}}^{x_{\max}} \text{PDF}_A(x) \left[\int_{x_{\min}}^x \text{PDF}_B(x') dx' \right] dx \quad (1)$$

This is simply the probability that A will get a result of x , $\text{PDF}_A(x)$, times the probability that the result of B will be worse than x , $\int_{x_{\min}}^x \text{PDF}_B(x') dx'$, integrated across all possible values of the metric x .

Equation 1 assumes that outcomes of A and B are uncorrelated; however, we can account for the target correlation by calculating P for each individual target in the test data set and averaging the result across all targets in the data set. We also wish to compare Method A to several methods at once. To do this, we replace the inner integral in eq 1 with a product of integrals: one integral for each method being compared (the probability that several methods will all do worse than x is the product of the individual probabilities for each method). Thus, given the result of M methods on a retrospective data set with N targets, the probability that method m will be the best in a (single) new system is

$$P_m^{\text{trial}} = \frac{1}{N} \sum_n \int_{x_{\min}}^{x_{\max}} \left[\text{PDF}_{m,n}(x) \prod_{i, i \neq m} \left(\int_{x_{\min}}^x \text{PDF}_{i,n}(x') dx' \right) \right] dx \quad (2)$$

where x is the metric used to evaluate performance, x_{\max} is the maximum possible value of the metric, x_{\min} is the minimum possible value of the metric, N is the number of systems in the retrospective data set, n is a given target in the retrospective data set, and $\text{PDF}_{m,n}$ is the probability distribution function describing the likely outcome (i.e., metric value) of method m on target n . $\text{PDF}_{m,n}$ is derived from the outcome of the docking program on the retrospective data.

For virtual screening, the expected outcome of method m on target n is represented by a distribution ($\text{PDF}_{m,n}$) because the actives and decoys in the test data set are only a subset of all

possible actives and decoys. $\text{PDF}_{m,n}$ is thus the expected outcome if method m were retested on target n with a different set of actives and decoys. In this work we assume that $\text{PDF}_{m,n}$ for the virtual screening results follows a normal distribution. While the true distribution cannot be normal for metrics with bounded ranges, such as the metrics used in this work, we have found that in practice this approximation is reasonable and gives very similar results (the calculated probabilities are mostly within 1–2%) to numerically bootstrapping $\text{PDF}_{m,n}$ or using a binomial distribution. Given a normal distribution the formula for the $\text{PDF}_{m,n}$ for virtual screening (VS) becomes

$$\text{PDF}_{m,n}^{\text{VS}}(x) = \sqrt{\frac{1}{2\pi\sigma_{m,n}^2}} \exp\left(-\frac{(x - X_{m,n})^2}{2\sigma_{m,n}^2}\right) \quad (3)$$

where $X_{m,n}$ is the virtual screening metric [either AUC or ROC($x\%$), Table 3] method m obtained on target n of the retrospective data set, and $\sigma_{m,n}$ is the variance of the distribution. The variance can be calculated by numerical bootstrapping, but the data required to perform a numeric bootstrap are not available for the Cross results. However, analytic estimates for $\sigma_{m,n}$ are available that depend only on $X_{m,n}$ and the number of actives and decoys in the test data set (which are known for the Cross results). Therefore, these analytic formulas (described below) are used to estimate $\sigma_{m,n}$ for the virtual screening result examined herein.

The analytic formulas to estimate $\sigma_{m,n}$ are different for AUC and the ROC($x\%$) metrics. For AUC, the variance, $\sigma_{m,n}$, is estimated analytically³⁵ as follows

$$\sigma^2 = \frac{\text{AUC}(1 - \text{AUC}) + (n_a - 1)(Q_1 - \text{AUC}^2) + (n_d - 1)(Q_2 - \text{AUC}^2)}{n_a n_d} \quad (4)$$

$$Q_1 = \frac{\text{AUC}}{2 - \text{AUC}}$$

$$Q_2 = \frac{2\text{AUC}^2}{1 + \text{AUC}}$$

where n_a is the number of actives, n_d is the number of decoys for target n , and AUC is the measured area under the Receiver Operator Characteristic (ROC) curve for method m on target n . For the ROC($x\%$) metrics, the variance, $\sigma_{m,n}$, is estimated analytically³⁶ using the following formula

$$\sigma^2 = \frac{f_a(1 - f_a)}{n_a} + \left(\frac{f_a}{f_d}\right)^2 \frac{f_d(1 - f_d)}{n_d} \quad (5)$$

As before, n_a is the number of actives, and n_d is the number of decoys for target n . f_d is the fraction of decoys for the given ROC($x\%$) metric, e.g., for ROC(1%) f_d is 1%. f_a is the fraction of actives recovered once f_d decoys have been recovered (i.e., the measured ROC($x\%$) metric value) by method m on target n .

Estimating relative performance of virtual screening using eq 2 (and eq 4 or 5 as appropriate with eq 3) accounts for the correlation in performance the methods may have on targets in the retrospective data set. For example, if two docking methods, A and B, have a sample mean AUC of 0.67 and 0.68, respectively, on the retrospective data set and the 95% confidence interval for the true mean value is ± 0.1 AUC from the measured (sample) mean, one would initially assume that the measured difference in mean AUC is not statistically significant because the estimate of the true mean varies much more than the measured difference of

sample means. However, if for each individual target in the data set the AUC of Method B was exactly 0.01 greater than that of Method A, then the difference in performance would be statistically significant because even though the AUC varies significantly from target to target for any given target, method B always does better than method A. This is an example of methods that are highly correlated. While methods are rarely as perfectly correlated as in this example, by accounting for correlation, the relative performance of docking methods can typically be distinguished with more statistical certainty than if correlation is ignored (they can also be distinguished with less certainty if the methods are anticorrelated; however, in practice this is quite rare).³⁷ Equation 2 does not account for the correlation of methods on individual compounds in the virtual screening database; however, recent work indicates that the correlation due to individual compounds is small compared to the target correlation.³⁷ We also do not have the required data from the methods listed in Table 2 to account for compound correlation.

The Cross2009 structure reproduction data set has only one ligand per target, and thus, it is not possible to estimate the likely variance in outcomes if different ligands were docked to the target protein. Therefore $\text{PDF}_{m,n}$ is approximated as a Dirac delta function for the structure reproduction results (eq 6).

$$\text{PDF}_{m,n}^{\text{SR}}(x) = \delta(x - \text{rmsd}_{m,n}) \quad (6)$$

where $\text{rmsd}_{m,n}$ is the rmsd obtained by method m on target n of the retrospective data set limited to the range [0.5 Å, 2.0 Å] (values less than 0.5 Å are set to 0.5 Å, and those greater than 2.0 Å are set to 2.0 Å). The rmsd range is capped in this calculation because, as discussed earlier, values outside of this range have limited meaning (very low rmsd values are below the experimental error crystal structure coordinates, and high rmsd values are simply wrong regardless of the value, e.g., 8 Å vs 10 Å rmsd). We would also like to note that this rmsd range was selected for the reasons discussed prior to performing any calculations, and no attempt was made to optimize FRED's performance by adjusting the range. Using this form of $\text{PDF}_{m,n}$ in eq 2, the probability that a method will be the best becomes simply the fraction of systems in the retrospective data set in which the method was the best (in the case of ties, the probability is split equally between the methods that tied).

Relative Mean Performance. Relative mean performance is measured as the probability, P_m^{mean} , that one method will do better than another on average. If the true trial probability is greater than 0.5, then method m is the best on average because it is the best on more than half the systems. Equation 2 calculates a trial probability, so this calculation would seem straightforward. However, the value calculated by eq 2 is the probability that method m will be the best on a single system *in* the retrospective data set, not the true probability that method m will be the best on systems *outside* the retrospective data set. Equation 2 does estimate the true probability, and the variance of that estimate is $P_m^{\text{trial}}(1 - P_m^{\text{trial}}/N)$ (thus, when N becomes large the variance approaches zero and P_m^{trial} is a perfect estimate of the true trial probability). If N is large enough for the Central Limit Theorem to apply then the probability distribution function describing the likely values of true trial probability P_m^{true} is normal (eq 7)

$$P_m^{\text{true}}(x) = \sqrt{\frac{N}{2\pi P_m^{\text{trial}}(1 - P_m^{\text{trial}})}} \exp\left(-\frac{N(x - P_m^{\text{trial}})^2}{2P_m^{\text{trial}}(1 - P_m^{\text{trial}})}\right) \quad (7)$$

If $m = 2$ (i.e., two systems are being compared), then the chance that the mean performance of one method is better than another (P_m^{mean}) is the same as the probability that the true trial probability is greater than 0.5, which is the integral of eq 7 from 0.5 to 1 or

$$P_m^{\text{mean}} = \frac{1}{2} \left[1 + \text{erf} \left((P_m^{\text{trial}} - 0.5) \sqrt{\frac{N}{2P_m^{\text{trial}}(1 - P_m^{\text{trial}})}} \right) \right] \quad (8)$$

Performance Correlation. Correlation is a measurement of the independence of two outcomes or, in the present case, the independence of the outcomes of two docking methods with respect to a given metric. Correlated docking methods will tend to perform well on the same systems (target correlation) or find similar active molecules for a given target (ligand correlation). As mentioned previously, the present work calculates and accounts for target correlation but not ligand correlation (the relative ranking of individual ligands is assumed to be uncorrelated) as the required underlying data for the methods listed in Table 2 was unavailable, and ligand correlation tends to be a small effect relative to target correlation.³⁷

Correlation is measured herein using the Pearson Correlation Coefficient, which has a range of $[-1, 1]$. A positive value indicates the methods are correlated (i.e., tend to do well on the same systems), while a negative value indicates they are anticorrelated (i.e., tend to do well on different systems), and zero indicates that they are independent. The Pearson Correlation Coefficient of two methods $m1$ and $m2$ ($\rho_{m1,m2}$) can be expressed as follows

$$\rho_{m1,m2} = \frac{\sigma_{m1} + \sigma_{m2} - \sigma_{\text{diff}}}{2\sqrt{\sigma_{m1}\sigma_{m2}}} \quad (9)$$

where σ_{m1} and σ_{m2} are the variances of the expected outcomes (with respect to a given metric) of the two methods, and σ_{diff} is the variance of the expected difference in outcomes. Variance can be calculated from the underlying probability distribution functions which for any given method m (PDF_m) is the average probability distribution across all targets or

$$\text{PDF}_m(x) = \frac{1}{N} \sum_n \text{PDF}_{m,n}(x) \quad (10)$$

where N is the number of targets systems, x is the metric being used, n is the number of targets in the data set, and $\text{PDF}_{m,n}$ comes from eq 3 for virtual screening results and eq 6 for structure reproduction. The probability distribution function for the expected difference in outcomes (PDF_{diff}) is again obtained by averaging the probability distribution functions for the difference in expected outcomes for each individual target as follows

$$\text{PDF}_{\text{diff}}(x) = \frac{1}{N} \sum_n \text{PDF}_{\text{diff},n}(x) \quad (11)$$

where $\text{PDF}_{\text{diff},n}$ is the expected difference in outcomes for target n , and is calculated with the following equation

$$\text{PDF}_{\text{diff},n}(x_{\text{diff}}) = \int_{\max(x_{\min}, x_{\min} - x_{\text{diff}})}^{\min(x_{\max}, x_{\max} - x_{\text{diff}})} \text{PDF}_{m1,n}(x) \text{PDF}_{m2,n}(x + x_{\text{diff}}) dx \quad (12)$$

Given the probability distribution functions the variances need by eq 9 can then be computed as follows

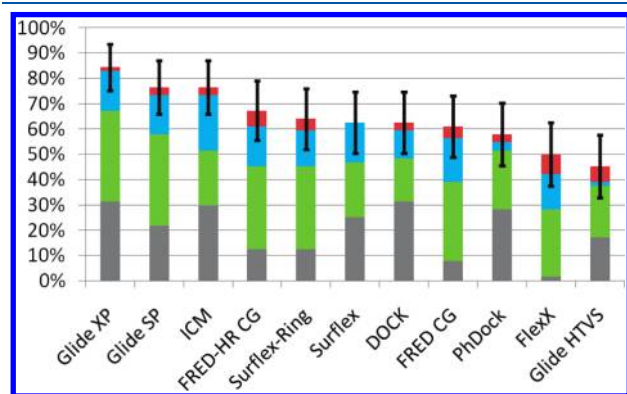
$$\sigma = \int_{x_{\min}}^{x_{\max}} (x - u)^2 \text{PDF}(x) dx \quad (13)$$

Table 4. Probability That the Row Method Will Do Better than the Column Method in a Single Structure Reproduction Docking Experiment

	Glide XP	Glide SP	ICM	FRED-HR CG	Surflex-Ring	Surflex	DOCK	FRED CG	PhDock	FlexX	Glide HTVS
Glide XP	50%	55%	61%	75%	69%	64%	57%	77%	66%	83%	73%
Glide SP	45%	50%	52%	69%	62%	60%	56%	70%	61%	74%	69%
ICM	39%	48%	50%	59%	68%	59%	54%	69%	55%	74%	62%
FRED-HR CG	25%	31%	41%	50%	49%	42%	40%	56%	51%	67%	55%
Surflex-Ring	31%	38%	32%	51%	50%	42%	39%	59%	42%	68%	55%
Surflex	36%	40%	41%	58%	58%	50%	53%	60%	49%	66%	60%
DOCK	43%	44%	46%	60%	61%	47%	50%	64%	55%	73%	60%
FRED CG	23%	30%	31%	44%	41%	40%	36%	50%	43%	55%	50%
PhDock	34%	39%	45%	49%	58%	51%	45%	57%	50%	61%	54%
FlexX	17%	26%	26%	33%	32%	34%	27%	45%	39%	50%	46%
Glide HTVS	27%	31%	38%	45%	45%	40%	40%	50%	46%	54%	50%

Table 5. Probability That the Row Method Will Do Better than the Column Method on Average in Structure Reproduction

	Glide XP	Glide SP	ICM	FRED-HR CG	Surflex-Ring	Surflex	DOCK	FRED CG	PhDock	FlexX	Glide HTVS
Glide XP	50%	81%	96%	100%	100%	99%	87%	100%	100%	100%	100%
Glide SP	19%	50%	65%	100%	97%	95%	84%	100%	96%	100%	100%
ICM	4%	35%	50%	94%	100%	92%	73%	100%	77%	100%	97%
FRED-HR CG	0%	0%	6%	50%	45%	10%	5%	84%	55%	100%	77%
Surflex-Ring	0%	3%	0%	55%	50%	10%	4%	92%	10%	100%	81%
Surflex	1%	5%	8%	90%	90%	50%	69%	95%	45%	100%	95%
DOCK	13%	16%	27%	95%	96%	31%	50%	99%	77%	100%	95%
FRED CG	0%	0%	0%	16%	8%	5%	1%	50%	13%	77%	50%
PhDock	0%	4%	23%	45%	90%	55%	23%	87%	50%	96%	73%
FlexX	0%	0%	0%	0%	0%	0%	0%	23%	4%	50%	27%
Glide HTVS	0%	0%	3%	23%	19%	5%	5%	50%	27%	73%	50%

**Figure 3.** Bar chart with success rates for docking within 0.5 Å (gray), 1.0 Å (gray + green), 1.5 Å (gray + green + blue), and 2.0 Å (gray + green + blue + red). Error bars are the 95% confidence interval for the true mean performance (i.e., the mean if all possible systems were tested) at 2.0 Å. Methods are sorted by success rate at 2.0 Å, and rmsds are for the top scoring pose.

where $\text{PDF}(x)$ is the probability distribution function from either eq 10 (when calculating σ_{m1} and σ_{m2}) or eq 12 (when calculating σ_{diff}), and u is the mean value of $\text{PDF}(x)$.

RESULTS AND DISCUSSION

Structure Reproduction. Absolute Performance of Best Pose. The probabilities that a given method will have a docking accuracy less than 0.5, 1.0, 1.5, and 2.0 Å rmsd are shown in

Figure 3. These methods tend to dock ligands very well (less than 1.0 Å) or very poorly (greater than 2.0 Å). The average success rates at 0.5, 1.0, 1.5, and 2.0 Å rmsd are 19%, 47%, 60%, and 64%, respectively. Very few ligands are docked between 1.5 and 2.0 Å (4% on average) and only a small fraction between 1.0 and 1.5 Å (13% on average). This indicates that when a docking program recognizes the correct binding mode it can generally reproduce it with high accuracy.

Relative Performance. The results in Figure 3 describe the absolute structure reproduction performance of individual methods. However, the question of most interest to researchers using molecular docking programs is given two different methods, which one is likely to give the best result when run in a prospective research setting. To evaluate relative performance, the probability that a given method will be the best for a single run is calculated (using eqs 2 and 6), and the results are shown in Table 4. The most differentiable pair of programs is Glide XP and FlexX with a 17% chance that FlexX will do better in structure reproduction than Glide XP. No probabilities in Table 4 are greater than 95% or lower than 5%. Therefore, on the basis of these results *we cannot say with 95% statistical certainty that any docking method examined here will outperform another on any given system.*

The mean performance of methods can be differentiated with much more statistical certainty than the trial performance. The mean relative performance is the probability that one method will outperform another on average across a large number of new systems. Table 5 shows the pairwise relative mean performance of methods calculated from the data in Table 4 using eq 8. Many

Table 6. Cross2009 Systems Groups by Number of Programs That Correctly Docked (rmsd < 2.0 Å)^a

number of programs that docked correctly to the systems	count of systems	systems
0	3	1qpc, 1xid, 1xie
1	2	1lna, 1tmn
2	7	1atl, 1qbr, 1fm9, 1o61, 1ew8, 1hfc, 1jap
3	8	1flr, 1tyl, 1bma, 1mld, 1mrk, 1nco, 2src, 4dfr, 1hyt
4	3	1glq, 1hvr, 1nq7
5	13	2cmd, 1hsl, 1jpa, 1mts, 1slt, 2aa2, 2tmn, 1mmq, 1wap, 2cli, 1qbu, 1qpd, 1srj
6	13	1aqw, 1coy, 1txi, 1a28, 1a6w, 1d3h, 1f3d, 1mrg, 1ydr, 2ctc, 1abf, 1aoe, 3ert
7	14	1abe, 1mvc, 1fcz, 1l2i, 1nhz, 1z95, 1a4q, 1c83, 1ke5, 1l1t, 2gbp, 2qwk, 3tpi, 5abp

^a The total count of programs is 7 (Glide, ICM, FRED, Surflex, DOCK, PhDock, and FlexX). One method was chosen for each program that had multiple methods (Glide SP for Glide, FRED CG for FRED, and Surflex for the Surflex program).

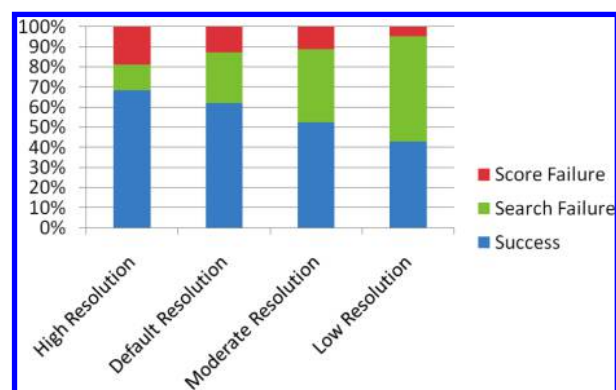
but not all methods are distinguishable to 95% certainty by this measure. Methods with similar absolute performance tend to be more difficult to differentiate but could be distinguished with by using a larger test data set (unlike the trial performance, the mean performance of any two methods can always be distinguished with a large enough data set).

Relative Difficulty of Targets. Some targets are easier to dock into than others. Easy targets may be ligands with few rotatable bonds or well-defined active sites. Easy targets are more common than hard targets in the Cross2009 data set (Table 6). There were 14 systems that every single docking method successfully docked a ligand within 2.0 Å and only 3 that every program got wrong.

FRED Structure Reproduction Analysis (Score vs Search Failures). FRED performs comparably to other docking programs in this test. Using Chemgauss, the default scoring function, the success rate at 2.0 Å is slightly above average using the high resolution search algorithm (67%) and slightly below average using the default search resolution (61%). Using the hybrid search method, the success rates at 2.0 Å are 95% and 92% using the high and default resolution docking algorithm, respectively. Hybrid FRED (FRED HYBRID and FRED-HR HYBRID) uses the correctly docked structure of the ligand to score poses during the exhaustive search, and thus, hybrid results are not predictive of FRED's structure reproduction performance in prospective research settings. However, the hybrid results do indicate the effectiveness of FRED's search algorithm and suggest that score is currently the factor most limiting FRED's structure reproduction success rate.

Determining why docking failures occur is a critical part of improving docking success rates. While the above results indicate that scoring rather than the search algorithm is the limiting factor in FRED's structure reproduction, FRED has a strong division between scoring and the docking algorithm used to create trial placements of ligands within the active site that allows for a more quantitative analysis of search failures (i.e., FRED did not test the correct pose) vs scoring failures (i.e., FRED did not recognize the correct pose).

Classifying a docking failure as a search or score failure requires that we know the best possible score a correctly docked pose can have. This score can be approximated by scoring the experimental structure. However, while the experimental structure is in the space of correctly docked poses, the score we would like to know is the best score of any pose in the space of correctly docked poses (i.e., the global optimum in score may be very near but not exactly at the experimentally determined structure of the ligand). To better approximate the best possible score a correct pose could have, we optimize the experimental pose with a

**Figure 4.** Docking success rate (rmsd < 2.0 Å) vs search resolution for FRED with Chemgauss scoring.

systematic solid body optimization that cannot move the pose outside the rmsd cutoff (2 Å) for successful docking. The resulting optimized score is then used as an estimate of the best possible score a correctly docked structure could have (correct is defined as within 2.0 Å rmsd in this context). It is important to emphasize that this optimization is performed solely to determine the best possible score of a correctly docked structure, and we do not use the resulting structure for rmsd calculations (this would introduce a favorable bias into the results). All rmsd calculations are performed with respect to the experimentally determined structure, not the optimized structure.

As part of this analysis we introduce two new search resolutions for the FRED docking algorithm, in addition to the high and default resolutions described in the Experimental section of this work. The moderate resolution settings for OMEGA are a maximum of 200 conformers and a rmsd duplicate removal threshold of 0.75 Å rmsd, and those for FRED are a translational step size of 1.5 Å, a rotational step size of 2.0 Å, and 100 poses optimized. In low resolution mode, OMEGA retains a maximum of 200 conformers and a duplicate removal threshold of 1.0 Å, and FRED uses a translational step size of 2.0 Å, a rotational step size of 2.5 Å, and optimizes 200 poses. The rates of docking successes (rmsd < 2.0 Å), search failures, and score failures are shown as a function of resolution in Figure 4.

These results show that simply increasing search resolution will not necessarily improve docking results. At higher resolutions, docking begins to fail not because the correct structure is never examined but rather because the scoring functions are unable to recognize the structure as correct. This suggests why methods such as flexible protein docking have not been widely

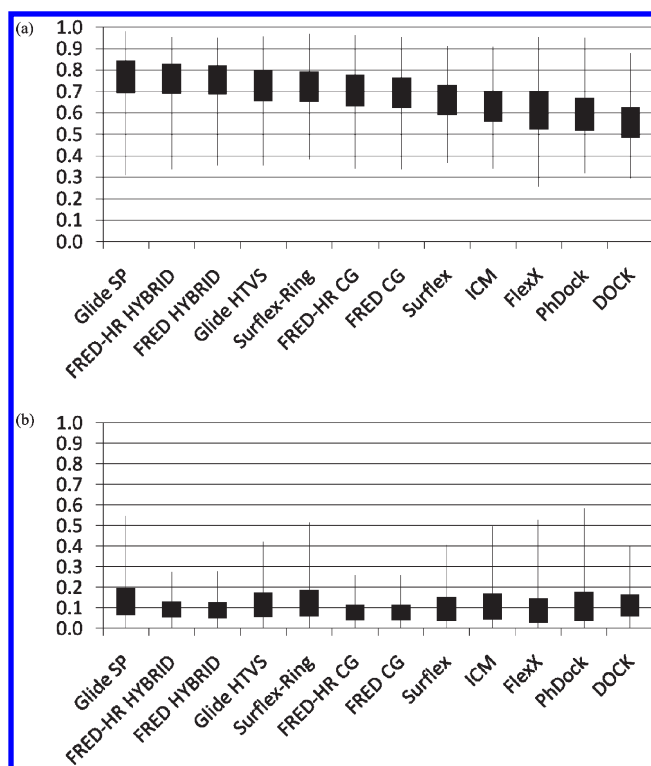


Figure 5. Estimated true virtual screening performance based on results from the DUD data set using metric (a) AUC and (b) ROC(1%). Expected average performance for a method that orders molecules randomly is 0.5 for AUC and 0.01 for ROC(1%). The 95% confidence interval for the true mean performance of a given method is shown by the thick black bar. The 95% confidence interval for a given methods performance on a single target is shown by a black line. Methods are ordered by mean AUC on the DUD data set. (See Figure I in the Supporting Information for this same set of results, including all early enrichment metrics listed in Table 3.).

adopted. Flexible protein docking searches a much larger space of structures but has not been proven to be generally more effective than rigid protein docking despite the vastly increased amount of CPU time required (often one or more orders of magnitude). The likely reason is that most scoring functions are not designed to account for the energetics of changes in the protein structure, and docking success is limited by the scoring function's inability to select the correct structure from incorrect structures in the flexible docking search space.

Virtual Screening. Absolute Performance. Predicted virtual screening performance is shown in Figure 5 (see Appendix A for an explanation of how the confidence intervals were obtained) both in terms of how well each method is expected to dock to a single new system (i.e., consistency) and how well each method is expected to do on average docking to many new systems (i.e., mean performance). On average, all methods, with one exception, detect active molecules at a rate better than random (to 95% confidence); however, the consistency of these methods is very poor (i.e., docking tends to do very well on one system and very poorly on the next). The single trial 95% confidence interval of every method has an upper limit above 0.9 AUC and a lower limit below 0.4 AUC range (DOCK's confidence interval 0.88 to 0.39 is an exception). Therefore, we can say little about the expected outcome of any method of when docking to a new system and likely results range from nearly perfect enrichment to worse than random!

This same result is observed using any of the early enrichment metrics (Figure 5 only includes ROC(1%) early enrichment, see Figure I of the Supporting Information for the complete set of early enrichment results). This lack of consistency is a property of the docking methods themselves and not a direct function of the size of the data set. Using a larger retrospective data set than DUD would reduce the confidence intervals of the true mean performance but not the size of the confidence interval for trial performance.

Relative Performance. The results in Figure 5 describe the absolute virtual screening performance of individual methods; however, the question of most interest to researchers using molecular docking programs is given two different methods, which one is likely to give the best result when run in a prospective research setting? We measure relative performance as the probability that a given method will do better than the method it is being compared to in a single virtual screen as described in the Experimental section. The probability that a method will be the best (i.e., have the best metric) is calculated using eqs 2, 3, and 4 or 5 as appropriate [eq 4 for AUC and eq 5 for ROC(1%)], and the results are shown in Table 7.

In a pair-wise comparison of methods, no methods performance on a single virtual screen is distinguishable from another to 95% confidence using any of the 6 virtual screening metrics (Table 3). Averaging across all virtual screening metrics (Table 3) and all pairs of methods, the probability that the method that has a better average metric value on DUD data set will also do better on a new target system is 61%. Broken down by metric type, the probability is 66%, 63%, 62%, 60%, 58%, and 57% for AUC, ROC(10%), ROC(5%), ROC(2%), ROC(1%), and ROC(0.5%), respectively. The high individual probability between any pair of methods, using any metric, is 86% (FRED-HR HYBRID has an 89% probability of getting a better AUC than DOCK when tested on a new system). Thus, we cannot say with any statistical certainty (generally accepted as 95% confidence) that any docking method will do better than another on a single new target system using any of the virtual screening metrics.

Methods are significantly more distinguishable in terms of their expected average performance on many new systems as opposed to the expected result on a single new system. The probability that one method will outperform another on average (calculated using eq 8 and the trial probabilities shown in Table 7) is shown in Table 8. Across all metrics and all method pairs, the average statistical distinguishability (where we define statistical distinguishability as the nondiagonal elements of Table 8 or 1 minus the element value, whichever is greater) is 85%. Broken down by metric, the average statistical distinguishability is 90%, 88%, 87%, 84%, 81%, and 78% for AUC, ROC(10%), ROC(5%), ROC(2%), ROC(1%), and ROC(0.5%), respectively. In general, methods that perform similarly on the DUD data set tend to be less distinguishable than those that performed differently. Using a larger data set than DUD would allow the mean performance of all methods to be distinguishable to 95% certainty, unlike trial performance, which is not directly a function of the size of the data set.

The correlation of any pair of docking methods is shown in Table 9. The values are computed (as described in the Analysis of Results section) using eq 9. Docking methods are overwhelmingly positively correlated (i.e., have a correlation coefficient greater than 0). The most positively correlated result is 0.91 (FRED-HR HYBRID and FRED HYBRID using AUC), while the least correlated method is -0.14 [Surflex-Ring and DOCK using ROC(0.5%)]. The average correlation coefficient across all

Table 7. Comparison of Virtual Screening Methods^a

(a)	Glide SP	FRED-HR Hybrid	FRED Hybrid	Glide HTVS	Surflex Ring	FRED-HR CG	FRED CG	Surflex	ICM	FlexX	Ph-dock	Dock
Glide SP	50%	53%	54%	67%	63%	67%	70%	77%	77%	77%	80%	86%
FRED-HR Hybrid	47%	50%	60%	59%	61%	72%	74%	75%	78%	79%	81%	89%
FRED Hybrid	46%	40%	50%	57%	60%	69%	72%	74%	77%	77%	81%	88%
Glide HTVS	33%	41%	43%	50%	51%	55%	58%	67%	72%	71%	77%	80%
Surflex Ring	37%	39%	40%	49%	50%	55%	57%	75%	71%	75%	76%	87%
FRED-HR CG	33%	28%	31%	45%	45%	50%	61%	62%	68%	68%	72%	80%
FRED CG	30%	26%	28%	42%	43%	39%	50%	60%	66%	66%	70%	77%
Surflex	23%	25%	26%	33%	25%	38%	40%	50%	58%	61%	66%	75%
ICM	23%	22%	23%	28%	29%	32%	34%	42%	50%	54%	58%	65%
FlexX	23%	21%	23%	29%	25%	32%	34%	39%	46%	50%	53%	62%
Ph-dock	20%	19%	19%	23%	24%	28%	30%	34%	42%	47%	50%	57%
Dock	14%	11%	12%	20%	13%	20%	23%	25%	35%	38%	43%	50%

(b)	Glide SP	FRED-HR Hybrid	FRED Hybrid	Glide HTVS	Surflex Ring	FRED-HR CG	FRED CG	Surflex	ICM	FlexX	Ph-dock	Dock
Glide SP	50%	57%	58%	60%	50%	63%	62%	66%	63%	70%	64%	56%
FRED-HR Hybrid	43%	50%	52%	46%	41%	61%	63%	57%	54%	65%	62%	53%
FRED Hybrid	42%	48%	50%	46%	43%	58%	62%	57%	52%	62%	60%	53%
Glide HTVS	40%	54%	54%	50%	44%	63%	62%	61%	56%	68%	62%	56%
Surflex Ring	50%	59%	57%	56%	50%	65%	63%	73%	61%	69%	66%	56%
FRED-HR CG	37%	39%	42%	37%	35%	50%	51%	51%	51%	56%	56%	45%
FRED CG	38%	37%	38%	38%	37%	49%	50%	52%	51%	55%	57%	44%
Surflex	34%	43%	43%	39%	27%	49%	48%	50%	48%	56%	51%	44%
ICM	37%	46%	48%	44%	39%	49%	49%	52%	50%	56%	54%	43%
FlexX	30%	35%	38%	32%	31%	44%	45%	44%	44%	50%	45%	39%
Ph-dock	36%	38%	40%	38%	34%	44%	43%	49%	46%	55%	50%	41%
Dock	44%	47%	47%	44%	44%	55%	56%	56%	57%	61%	59%	50%

^a Elements are the probabilities that the row method will have a better result (with respect to a given metric) than the column method for a **single virtual screen**. Metrics for each table are (a) AUC and (b) ROC(1%). See Table V of the Supporting Information for these same results, including all early enrichment metrics listed in Table 3.

differing method pairs and metrics is 0.27. Broken down by metric the average correlation coefficients are 0.40, 0.34, 0.33, 0.25, 0.17, and 0.14 for AUC, ROC(10%), ROC(5%), ROC(2%), ROC(1%), and ROC(0.5%), respectively.

The correlation of a docking method with itself must be 1.0 (i.e., if you run the same method twice you will get the same results). However, the self-correlation coefficients in Table 9 (the diagonal elements) have been calculated in the same way as correlation coefficients of differing docking methods, and this calculation assumes that the ligand ranking of any two methods is uncorrelated (specifically this occurs when using eq 12). The amount that the self-correlation coefficients in Table 9 deviate from 1.0 is therefore an estimate of the maximum error in these calculations because of the assumption that the ligand rankings are uncorrelated. The average self-correlation coefficient (calculated using eq 9 under the assumption that ligand ranking is uncorrelated) varies by metric and is 0.93, 0.90, 0.87, 0.81, 0.77, and 0.72 for AUC, ROC(10%), ROC(5%), ROC(2%), ROC(1%), and ROC(0.5%), respectively. Thus, AUC appears to be the least significantly affected by the ligand correlation, while early enrichment metrics are more affected (the more “early” the enrichment the more the effect). At most, ligand correlation appears to account for about 28%, using the ROC(0.5%) metric, of the overall correlation. This is also a worst case scenario as it is unlikely that two different methods will have perfect ligand correlation (except in the self-correlation case), and thus, assuming uncorrelated ligands will be less of an effect.

FRED Hybrid Docking. FRED’s Hybrid scoring function outperforms structure-based scoring with Chemgauss on the DUD data set at both default and high resolution. (Hybrid AUC is 0.76 and 0.75 at high and default resolution, respectively, while Chemgauss is 0.70 and 0.69, respectively.) This difference in performance is statistically significant in terms of predicted true mean performance. The difference between hybrid and Chemgauss scores is not statistically significant for a single trial, which is somewhat surprising given that hybrid and Chemgauss scoring used the same scoring function to optimize and score ligands (the hybrid method uses a ligand-based scoring function during the exhaustive search). This would lead us to expect a higher degree of correlation between these methods than most. Presumably the hybrid method does better in systems where most actives dock in the same binding mode as the known bound ligand, thus reducing the false positive rate for selecting the correct pose, and worse on systems where the actives dock in multiple modes by increasing the false negative rate. Overall hybrid docking appears to be a promising method of improving virtual screening results, although the same lack of consistency observed in traditional docking programs also occurs in hybrid docking.

Early versus Global Enrichment Metrics. While virtual screening methods will shift actives toward the top of the hit list, as a practical matter, only the first few hundred molecules at beginning of the hit list are ever examined and given serious consideration for further testing. Therefore, most metrics for measuring virtual screening performance are early enrichment

Table 8. Comparison of Virtual Screening Methods^a

(a)	Glide SP	FRED-HR Hybrid	FRED Hybrid	Glide HTVS	Surflex Ring	FRED-HR CG	FRED CG	Surflex	ICM	FlexX	Ph-dock	Dock
Glide SP	50%	60%	74%	99%	90%	94%	99%	100%	100%	100%	100%	100%
FRED-HR	40%	50%	68%	97%	87%	99%	100%	100%	100%	100%	100%	100%
Hybrid												
FRED Hybrid	26%	32%	50%	94%	80%	98%	100%	100%	100%	100%	100%	100%
Glide HTVS	1%	3%	6%	50%	47%	51%	75%	98%	99%	91%	100%	100%
Surflex Ring	10%	13%	20%	53%	50%	45%	52%	99%	98%	100%	100%	100%
FRED-HR CG	6%	1%	2%	49%	55%	50%	78%	97%	98%	98%	99%	100%
FRED CG	1%	0%	0%	25%	48%	22%	50%	91%	95%	95%	99%	100%
Surflex	0%	0%	0%	2%	1%	3%	9%	50%	76%	88%	91%	100%
ICM	0%	0%	0%	1%	2%	2%	5%	24%	50%	57%	83%	93%
FlexX	0%	0%	0%	9%	0%	2%	5%	12%	43%	50%	58%	96%
Ph-dock	0%	0%	0%	0%	0%	1%	1%	9%	17%	42%	50%	86%
Dock	0%	0%	0%	0%	0%	0%	0%	0%	7%	4%	14%	50%

(b)	Glide SP	FRED-HR Hybrid	FRED Hybrid	Glide HTVS	Surflex Ring	FRED-HR CG	FRED CG	Surflex	ICM	FlexX	Ph-dock	Dock
Glide SP	50%	82%	85%	89%	48%	95%	94%	98%	95%	99%	96%	79%
FRED-HR	18%	50%	58%	30%	14%	92%	95%	83%	68%	97%	93%	66%
Hybrid												
FRED Hybrid	15%	42%	50%	31%	20%	83%	93%	82%	59%	93%	89%	66%
Glide HTVS	11%	70%	69%	50%	24%	95%	94%	93%	76%	99%	93%	76%
Surflex Ring	52%	86%	80%	76%	50%	97%	95%	100%	93%	99%	98%	78%
FRED-HR CG	5%	8%	17%	5%	3%	50%	56%	54%	53%	79%	76%	25%
FRED CG	6%	5%	7%	6%	5%	44%	50%	59%	57%	75%	82%	21%
Surflex	2%	17%	18%	7%	0%	46%	41%	50%	41%	79%	57%	22%
ICM	5%	32%	41%	24%	7%	47%	43%	59%	50%	78%	71%	18%
FlexX	1%	3%	7%	1%	1%	21%	25%	21%	22%	50%	28%	8%
Ph-dock	4%	7%	11%	7%	2%	24%	18%	43%	29%	72%	50%	12%
Dock	21%	34%	34%	24%	22%	75%	79%	78%	82%	92%	88%	50%

^a Elements are the probabilities that the row method has a better **true mean** AUC than the column method. Metrics for each table are (a) AUC and (b) ROC(1%). See Table VI of the Supporting Information for these same results, including all early enrichment metrics listed in Table 3.

metrics, which only account for actives that appear in the beginning of the hit list. Early enrichment metrics, however, have some drawbacks. Either the range or the expected value of random enrichment is a function of the number of actives and decoys of most early enrichment metrics. (In the case of the ROC($x\%$) metrics used herein, the range is constant, but the expected average random value is a function of the number of actives and decoys, although this effect decreases as the number of actives and decoys becomes large. The effect is relatively small for most systems in the DUD data set, approximately 10%.) Early enrichment values are also calculated using less information than global enrichment values (i.e., they use only the information in the beginning of the hit list, while global enrichment metrics use information from the entire hitlist). Given this, it seems likely that early enrichment results would have less statistical significance versus global enrichment results.

The relative “statistical power” of global versus early enrichment metrics is illustrated in Figure 6. Using the AUC global enrichment metric, the average performance of two docking methods can be distinguished to 95% confidence 61% of the time. This percentage steadily decreases as we switch to ROC-(10%) and then progressively “earlier” early enrichment metrics [the probability is 14% for ROC(0.5%)]. In principle, by using a

large enough data set, the relative average performance of two methods could be distinguished to 95% confidence 100% of the time, even using a relatively information poor early enrichment. However, constructing a data set the size of DUD is already a large undertaking and constructing a large enough data set is no easy task. In addition, the ability to distinguish the single trial performance of two methods cannot be directly addressed by increasing the size of the data set.

The ability of global versus early enrichment to distinguish single trial performance is illustrated in Figure 7. The same trend observed for mean performance (Figure 6) occurs here with global enrichment (AUC) doing the best (albeit very poorly with 18% even at 75% confidence) and early enrichment metrics being progressively less able to distinguish methods the “earlier” they are.

Structure Reproduction versus Virtual Screening. Virtual screening and structure reproduction performance (shown in Figure 8) are not highly correlated. Several programs have relatively poor virtual screening performance compared to their structure reproduction performance (DOCK, PhDock, and ICM). This is a plausible result because many elements of scoring that vary from ligand to ligand are constant when comparing poses. For virtual screening to be accurate, all elements of the scoring

Table 9. Pearson Correlation for Virtual Screening Using (a) AUC and (b) ROC(1%)^a

(a)	Glide SP	FRED-HR Hybrid	FRED Hybrid	Glide HTVS	Surflex Ring	FRED-HR CG	FRED CG	Surflex	ICM	FlexX	Ph-dock	Dock
Glide SP	0.95	0.57	0.56	0.76	0.55	0.50	0.55	0.47	0.26	0.28	0.11	0.16
FRED-HR Hybrid	0.57	0.93	0.91	0.39	0.54	0.72	0.70	0.40	0.32	0.40	0.20	0.32
FRED Hybrid	0.56	0.91	0.92	0.40	0.53	0.70	0.69	0.43	0.29	0.36	0.20	0.28
Glide HTVS	0.76	0.39	0.40	0.93	0.40	0.34	0.40	0.45	0.38	0.26	0.28	0.03
Surflex Ring	0.55	0.54	0.53	0.40	0.93	0.40	0.36	0.73	0.36	0.53	0.28	0.46
FRED-HR CG	0.50	0.72	0.70	0.34	0.40	0.93	0.90	0.45	0.41	0.31	0.19	0.23
FRED CG	0.55	0.70	0.69	0.40	0.36	0.90	0.92	0.43	0.39	0.26	0.15	0.14
Surflex	0.47	0.40	0.43	0.45	0.73	0.45	0.43	0.91	0.39	0.49	0.32	0.34
ICM	0.26	0.32	0.29	0.38	0.36	0.41	0.39	0.39	0.92	0.35	0.32	0.11
FlexX	0.28	0.40	0.36	0.26	0.53	0.31	0.26	0.49	0.35	0.96	0.19	0.41
Ph-dock	0.11	0.20	0.20	0.28	0.28	0.19	0.15	0.32	0.32	0.19	0.93	0.13
Dock	0.16	0.32	0.28	0.03	0.46	0.23	0.14	0.34	0.11	0.41	0.13	0.91

(b)	Glide SP	FRED-HR Hybrid	FRED Hybrid	Glide HTVS	Surflex Ring	FRED-HR CG	FRED CG	Surflex	ICM	FlexX	Ph-dock	Dock
Glide SP	0.83	0.07	0.15	0.51	0.33	0.03	0.04	0.37	-0.03	0.30	-0.08	0.00
FRED-HR Hybrid	0.07	0.66	0.60	0.02	0.21	0.52	0.52	0.09	0.11	0.18	0.22	0.09
FRED Hybrid	0.15	0.60	0.69	-0.05	0.21	0.49	0.53	0.10	-0.01	0.17	0.11	0.01
Glide HTVS	0.51	0.02	-0.05	0.78	0.28	-0.06	-0.07	0.19	0.08	0.12	0.15	0.05
Surflex Ring	0.33	0.21	0.21	0.28	0.76	0.15	0.16	0.57	0.32	0.39	0.29	-0.10
FRED-HR CG	0.03	0.52	0.49	-0.06	0.15	0.67	0.63	0.12	0.04	0.12	0.22	0.00
FRED CG	0.04	0.52	0.53	-0.07	0.16	0.63	0.68	0.16	0.03	0.11	0.25	0.01
Surflex	0.37	0.09	0.10	0.19	0.57	0.12	0.16	0.77	0.25	0.26	0.26	-0.09
ICM	-0.03	0.11	-0.01	0.08	0.32	0.04	0.03	0.25	0.84	0.24	0.35	0.06
FlexX	0.30	0.18	0.17	0.12	0.39	0.12	0.11	0.26	0.24	0.91	0.01	0.16
Ph-dock	-0.08	0.22	0.11	0.15	0.29	0.22	0.25	0.26	0.35	0.01	0.80	-0.11
Dock	0.00	0.09	0.01	0.05	-0.10	0.00	0.01	-0.09	0.06	0.16	-0.11	0.79

^a See Table VII of the Supporting Information for these same results, including all early enrichment metrics listed in Table 3. Diagonal elements are self correlation.

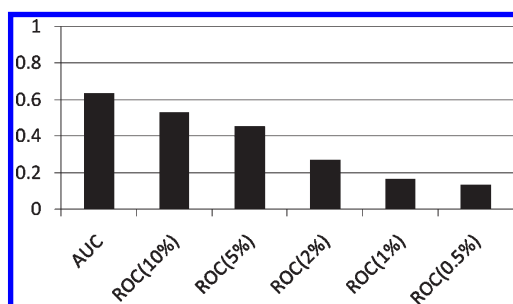


Figure 6. Probability that the relative average performance of two methods can be distinguished to 95% confidence from the DUD data set using global (AUC) and early enrichment [i.e., ROC(*x*%) metrics].

function must be correct; however, for pose prediction, only the elements of the scoring function that vary between poses need be correct. Thus, it is easier for a scoring function to be good at pose prediction than virtual screening because errors in any element of the scoring function always degrade virtual screening performance, while not all errors affect pose prediction.

Glide HTVS is competitive at virtual screening but the worst performer in structure reproduction. This result is more difficult to explain. As discussed by Warren et al.,¹ good structure

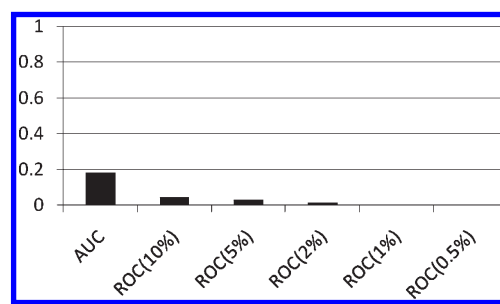


Figure 7. Probability that the relative single trial performance of two methods can be distinguished to 75% confidence using global enrichment (AUC) or early enrichment [ROC(*x*%) metrics].

reproduction performance would seem to be a prerequisite for good virtual screening. If a program incorrectly docks a ligand within the site, the score of the ligand will presumably be different than if it were docked correctly, and thus, its ranking in the virtual screen will be different than if it were correctly docked. One would expect that incorrectly ranking ligands in this way would degrade virtual screening performance. Nevertheless, the present results indicate that despite its poor structure reproduction performance Glide HTVS has good virtual screening performance.

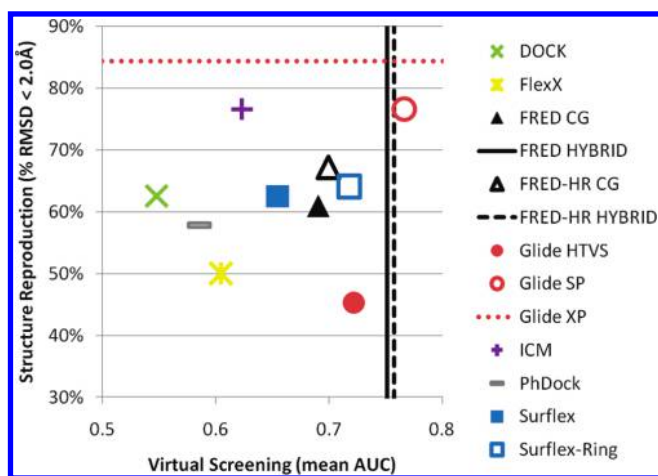


Figure 8. Structure reproduction vs virtual screening (AUC) performance. Methods for which one but not both values are available (i.e., FRED Hybrid and Glide XP results) are plotted as a line.

We posit two explanations for how Glide HTVS has good virtual screening and poor structure reproduction performance. First, if a ligand docked incorrectly had the same, or very similar, score to the correctly docked pose, virtual screening would not be degraded. While this explanation is possible, it seems unlikely that such a coincidence would occur on a large scale in a virtual screening run. Testing this hypothesis also requires knowledge of the correctly docked structure for all ligands, which was not available for the DUD data set.

The second possibility is that the scores of the ligands are determined primarily by the properties of the ligand and not by the interactions the ligand makes with the active site. For example, a score proportional to the number of atomic contacts a ligand makes with the site would be size biased and generally score larger ligands better than smaller ones. For a given ligand, different poses would have different scores, but the difference would likely be small compared to the size effect. While simplistic, this example illustrates how a scoring function could inadvertently be testing more for ligand properties than the interaction the ligand makes with the protein. As we do not have access to Glide, we cannot test this hypothesis for Glide HTVS. However, we have devised a simple way to test for this type of bias that is explained in detail in the next section and applied to FRED.

Null Hypothesis Testing. An underlying assumption of all docking programs is that active ligands can be distinguished from inactive ligands because active ligands will complement the shape and chemical functionality of the receptor site better than inactive ligands. Figure 5 shows that, while performance varies significantly from target to target, on average almost all docking methods are capable of distinguishing actives from decoys in a virtual screen. However, it is possible that the virtual screening performance observed in Figure 5 arises not because the actives complement the receptor site better than the decoys but rather because the actives and decoys differ in some systematic way not related to the protein structure that is recognizable to the scoring function. While the DUD data set decoys were designed to have similar properties to the actives, it would be useful to eliminate this possibility.

To verify that the protein structure is being used to distinguish active and inactive molecules we have designed a simple null

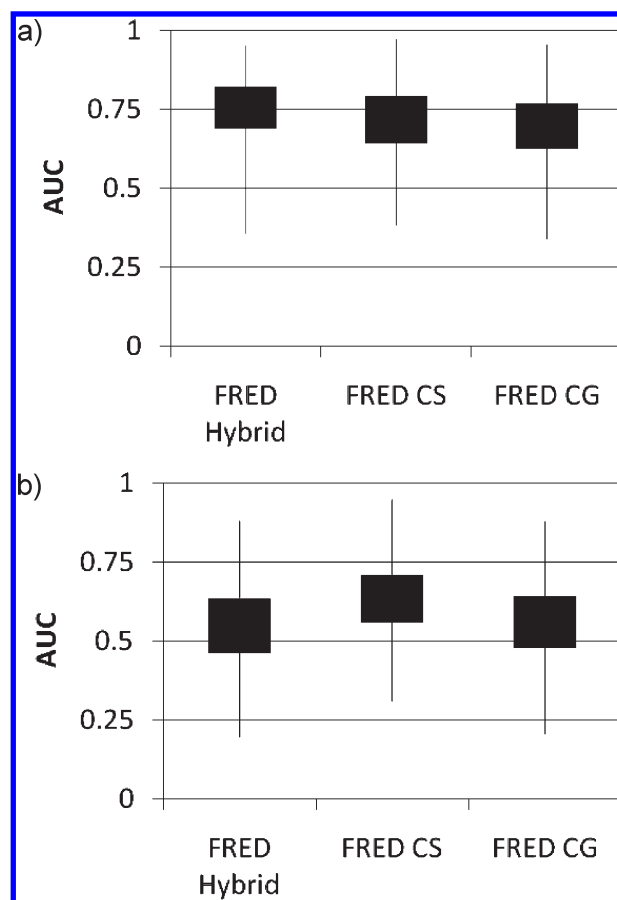


Figure 9. FRED virtual screening performance (AUC) for hybrid, Chemscore, and scoring when docking to (a) the correct protein structure and (b) the incorrect protein structure. The 95% confidence interval for the true mean performance of a given method is shown by the thick black bar. The 95% confidence interval for a given method's performance on a single target is shown by a black line.

hypothesis test by docking ligands of the DUD data set into a random protein target from the DUD data set (see Supporting Information for exact pairings). The hypothesis we are testing is that virtual screening requires the correct protein target. Because the active molecules in this test are not active against the target we are docking to, the predicted mean virtual screening should be random. If the predicted true mean performance is better (or worse) than random (AUC=0.5), then the docking method is distinguishing actives from decoys without using the structure of the protein.

Statistically significant virtual screening performance when docking to the incorrect protein structure is not necessarily indicative of an error or bias in the scoring function or data set; however, methods that show statistically significant enrichment docking to random protein structures are undesirable for use in drug development. For example, a hypothetical scoring function that perfectly calculated $\Delta G_{\text{binding}}$ could produce enrichment in some circumstances even when docking to the incorrect protein structure. Specifically $\Delta G_{\text{binding}}$ is defined as

$$\Delta G_{\text{binding}} = G_{\text{ProLig}} - G_{\text{Pro}} - G_{\text{Lig}} \quad (14)$$

G_{Lig} is a property of the ligand, while G_{ProLig} is the interaction between the protein and ligand. G_{ProLig} would be expected to be random when docking to the wrong protein (G_{Pro} is constant for

a given protein), but if G_{lig} were significantly higher for the actives than the decoys (i.e., if the actives are less soluble than the decoys), then positive enrichment could be obtained. However, the actives obtained in this way would be highly promiscuous binders, undesirable for drug development.

To perform the null hypothesis test, we dock the ligands using the randomized data set (see Table 1 of the Supporting Information) and analyze the results as we would any virtual screening tests. Because we do not have access to the programs tested by Cross et al.,¹² we could not perform this test for programs besides FRED. The results for FRED are shown in Figure 9. In addition to the Chemgauss and hybrid scoring functions described previously, these results include virtual screens using an addition scoring function, Chemscore.³⁰

Docking to the correct structure, Chemscore outperforms FRED's default scoring function Chemgauss (Figure 9a); however, Chemscore's predicted mean AUC is better than random (AUC = 0.5) when docking to the incorrect protein structure (Figure 9b). This shows that a portion of Chemscore's virtual screening performance comes from recognition of a property or properties of the active ligands that differs from the decoys. We hypothesize that this property is the number of rotatable bonds because Chemscore directly penalizes a molecule on the basis of the number of rotatable bonds, and the decoy molecules in the DUD data set have on average more rotatable bonds than the active molecules (5.5 vs 4.9 for decoys and actives, respectively). The predicted true mean performance of Chemgauss and hybrid scoring are not predicted to be better than random (with statistical significance), and we therefore conclude that these scoring functions are distinguishing actives from decoys using the information present in the correct protein structure.

CONCLUSIONS

Using a docking program for hit discovery and lead optimization is worthwhile. Prospective virtual screening is predicted, with statistical certainty greater than 99%, to be better than random *on average* for all but one docking program examined. The average amount of enrichment varies from docking program to docking program, with the best programs having *on average* approximately a 75% chance of giving an active molecule a better score than an inactive one. The relative *average* performance of these methods is statistically significant in most cases, although programs with similar mean performance on the DUD data set are not statistically distinguishable in terms of mean predicted performance on prospective systems. There still remains room for improvement in even the best docking programs as ideally every active molecule would score better than an inactive molecule. Our analysis of FRED indicates that such improvements are likely to come in the form of better scoring functions rather than better docking algorithms. Nevertheless, despite the potential for improvement, current docking programs are capable of distinguishing actives from decoys at a rate that is statistically significantly better than random.

Lack of consistency in virtual screening and structure reproduction results is the Achilles' heel of docking programs. They tend to do very well on some systems and terribly on others. The worst virtual screening method (measured by AUC) is on average only slightly better than random; nevertheless, it is expected to outperform the best virtual screening method (measured by AUC) 1 in 7 times. The situation for structure reproduction is even worse, with the worst method expected to outperform the

best method 1 in 4 times. Inconsistency makes it difficult to evaluate and improve docking programs because variance in outcome due to random chance is generally much larger than variance in outcome between methods. This limitation can be overcome by using large enough test data sets to tease out small differences in the true mean performance, but assembling and testing such large data sets is difficult and time-consuming. Inconsistency also makes it difficult for modelers to trust docking programs. While they often perform well, they always perform poorly on some systems, and there is no way for a modeler to know if his or her new research target is one of the systems for which docking will work. In a sense, using a docking program is like the stock market: on average it pays off, but you can also lose it all.

Hybrid docking appears to be a promising way to improve virtual screening performance. FRED's hybrid docking method outperforms FRED's standard docking in virtual screening (using either global or early enrichment metrics), and the mean difference in performance is statistically significant to 95% confidence (except when using 1% and 0.5% early enrichment, which are only different to 90% confidence). Docking with the hybrid method is also faster than standard docking by approximately a factor of 2, and the ligand information used by FRED's hybrid method (the structure of a bound ligand within the active site) is commonly available to modelers when doing prospective virtual screening.

Virtual screening studies should test docking programs by randomizing the target proteins of the test data set, in addition to the normal analysis with the correct protein target, to verify that actives are recognized by the way they complement the shape and chemistry of the active site rather than by recognition of differences in properties between the active and decoy molecules of the test data set. This analysis was undertaken for FRED and revealed that one of FRED's alternate scoring functions, the commonly used Chemscore, was biased to favor the DUD actives over the DUD decoys, independent of the structure of the active site (although using the correct protein structure improved performance further). While the CPU cost of this testing is significant (each ligand database must be docked an additional time), the manpower should generally be minimal because the scripts and other machinery used to analyze the virtual screening results with the correct protein structure can generally be reused.

Global enrichment metrics are more capable of distinguishing the relative performance than early enrichment metrics, and "earlier" early enrichment metrics are also less able to distinguish the performance of two methods than "later" early enrichment metrics. When comparing relative average performance the global enrichment metric, AUC was capable of distinguishing any two methods with statistical certainty 61% of the time (statistical certainty being 95% confidence), while the earliest early enrichment metric, ROC(0.5%), could only distinguish two methods with statistical certainty 14% of the time. On the basis of these results, we recommend using global rather than early enrichment metrics to evaluate virtual screening performance.

APPENDIX A: R-CURVES

We analyze absolute virtual screening performance using a type of cumulative distribution function that we refer to as a reliability curve, or R-curve. An R-curve is a plot of the values of a metric (e.g., AUC, enrichment factor, etc.) versus the probability that any test result will be better than the corresponding value.

When lower values of the metric are better values (e.g., rmsd), the R-curve is simply the cumulative distribution function. If high values of the metric are better values, as is the case with most virtual screening metrics, the R-curve is one minus the cumulative distribution function or

$$R(x) = 1 - \int_{x_{\min}}^x P(x') \leftarrow dx' \quad (\text{A1})$$

where x is the value of the metric, P is the probability distribution of the metric, and $R(x)$ is the probability that any test result will be better than x . The probability distribution function, P , for a set of N test systems is the average of the probability distribution for each individual target, thus the R-curve equation becomes

$$R(x) = 1 - \frac{1}{N} \sum_i \int_{x_{\min}}^x P_i(x') \leftarrow dx' \quad (\text{A2})$$

When there is no error measuring each individual target, the probability function for each target, P_i , is the Dirac delta function at the measured value for that target, the integral of which is the Heaviside step function. Virtual screening uses a finite set of actives and decoys, and the associated error when measuring the performance of an individual target must be taken into account. If the error is normally distributed, the R-curve formula is

$$R(x) = 1 - \frac{1}{N} \sum_i \int_{x_{\min}}^x \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x' - x_i)^2}{2\sigma_i^2}\right) dx' \quad (\text{A3})$$

where x_i is the measured value of the metric on the i^{th} target, and σ_i^2 is the variance of that measurement due to the limited number of actives and decoys. When using AUC, an analytical estimate for this variance is available that depends only on the AUC and the number of actives and decoys (see eq 4 in the main text). Integrating the normal distributions we arrive at

$$R(x) = 1 - \sum_i \frac{1}{2N} \left[1 + \operatorname{erf}\left(\frac{x - x_i}{\sqrt{2\sigma_i^2}}\right) \right] \quad (\text{A4})$$

Equation A4 is the form of the R-curve we use when the variance of the error (σ_i^2) can be calculated analytically. If the variance of the error cannot be calculated analytically and the scores of the actives and decoys are known, the individual probability distributions can be calculated numerically.

The R-curve formula of eq A4 accounts for the error due to limited numbers of actives and decoys that are docked to each target; however, we must also account for the error due to the limited number of targets in the data set. (Error in this context refers to the confidence with which we can estimate the population mean performance from our sample, the DUD data set.) To account for this error, we recognize that any point on the R-curve, $R(x_o)$, is a binary response system [$R(x_o)$ is the probability that $x > x_o$, and $x > x_o$ can only have two results: true or false; hence, it is a binary response system]. The variance, for a single trial of a binary response system is $p(1 - p)$, where p is the mean value of the outcome, $R(x_o)$. Because we are interested in the variance of the mean rather than the variance a single trial, we divide by the number of samples. Assuming the error is normally distributed a 95% confidence, upper and lower bound R-curves can be calculated as follows

$$\text{CI95}(x) = R(x) \pm 1.96 \sqrt{\frac{R(x)(1 - R(x))}{N}} \quad (\text{A5})$$

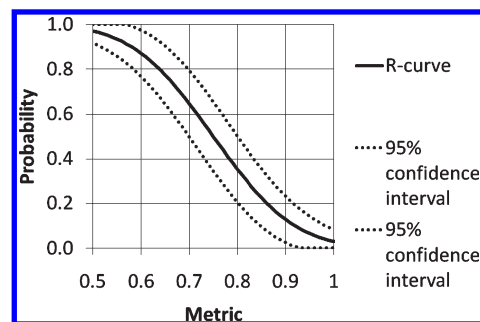


Figure A1. Example R-curve and 95% confidence interval.

where N is the number of test systems (in the case of the DUD data set, $N = 40$), and $R(x)$ is the R-curve we obtained on the retrospective test.

These R-curves (shown in Figure A1) enclose the range within which there is a 95% probability that the population R-curve lies (i.e., the R-curve we would obtain by testing an infinite number of virtual screening systems). By treating the upper and lower bound of the 95% confidence as normal R-curves, the 95% confidence interval for the median can be calculated (by the definition of median this is simply the value of the R-curve @ metric = 0.5). Equation A5A shows that the confidence interval of the probability of doing better than a given value varies by $1/(N)^{1/2}$, where N is the number of test systems. Note that eq A5A calculates the confidence interval of the probability of doing better than a given value metric value, not the confidence interval of the mean or median metric. The relationship between the number of test systems and confidence interval of a metric mean or median is more complex than $1/(N)^{1/2}$, as the shape of the R-curve is relevant as well. However, it is a reasonable rule of thumb that quadrupling the number of test systems will reduce the error predicting the true mean or median by half.

In the preceding discussion, we show that a median value of a test result can be obtained from an R-curve by taking the metric value at the point on the R-curve where the probability is 0.5. This is by definition because when the probability is 0.5, half of the measured values will be greater than the given value of the metric. The mean value of the metric can also be obtained from an R-curve by taking the area under the R-curve and adding the minimum value of the metric.

■ ASSOCIATED CONTENT

S Supporting Information. A complete set of figures and tables for early enrichment results, a list of the OMEGA and FRED options used for each mode, and more. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mcgann@eyesopen.com.

Present Addresses

[†]OpenEye Scientific Software, 222 3rd Street, Suite 3120, Cambridge, MA 02142, United States.

■ ACKNOWLEDGMENT

The author thanks Jason Cross, who helpfully supplied the structure reproduction rmsd and virtual screening AUC values

for the docking program he tested and reported on in a previous work.¹²

REFERENCES

- (1) Warren, G. L.; Andrews, C. W.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Pieshoff, C. E.; Head, M. S. A Critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5913.
- (2) Hawkins, P. C. D.; Skillman, G. A.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.
- (3) Cleves, A. E.; Jain, A. N. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 147–159.
- (4) Fingerprints: Screening and Similarity. Daylight Chemical Information Systems, Inc. <http://www.daylight.com/dayhtml/doc/theory/theory.finger.html> (accessed November 2, 2010).
- (5) Vidal, D.; Thormann, M.; Pons, M. LINGO: An efficient holographic text-based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386–393.
- (6) Grant, J. A.; Haigh, J. A.; Pickup, B. T.; Nicholls, A.; Sayle, R. A. Lingos, finite state machines, and fast similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 1912–1918.
- (7) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A. Fast method of molecular shape comparison: A simple application of a gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653–1666.
- (8) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A. Shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.
- (9) BROOD. OpenEye Scientific Software. <http://www.eyesopen.com/products/applications/brood.html> (accessed November 2, 2010).
- (10) Salam, N. K.; Nuti, R.; Sherman, W. Novel method for generating structure-based pharmacophores using energetic analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2356–2368.
- (11) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.
- (12) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455–1474.
- (13) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular docking using shape descriptors. *J. Comput. Chem.* **1992**, *13*, 380–397.
- (14) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175–1189.
- (15) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein–ligand docking. *Proteins: Struct., Funct., Bioinf.* **1999**, *37*, 228–241.
- (16) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: Docking and Scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.
- (17) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM. A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506.
- (18) Totrov, M.; Abagyan, R. Flexible protein–ligand docking by global energy optimization in internal coordinates. *Proteins: Struct., Funct., Bioinf.* **1998**, *Suppl. 1*, 215–220.
- (19) Joseph-McCarthy, D.; Thomas, B. E. I. V.; Belmarsh, M.; Moustakas, D.; Alvarez, J. C. Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins: Struct., Funct., Bioinf.* **2003**, *51*, 172–188.
- (20) Joseph-McCarthy, D.; McFadyen, I. J.; Zou, J.; Walker, G.; Alvarez, J. C. Pharmacophore-Based Molecular Docking: A Practical Guide. In *Virtual Screening in Drug Discovery*, 1st ed.; Alvarez, J. C., Shoichet, B., Eds.; CRC Press: Boca Raton, FL, 2005; Vol. 1, pp 327–347.
- (21) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499–511.
- (22) Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281–306.
- (23) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein–ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.
- (24) Jain, A. N. Scoring noncovalent protein–ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427–440.
- (25) Lee, H. S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.; Yoon, S. Optimization of high throughput virtual screening by combining shape-matching and docking methods. *J. Chem. Inf. Model.* **2008**, *48*, 489–497.
- (26) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph. Model.* **2003**, *21*, 449–462.
- (27) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499–2510.
- (28) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76–90.
- (29) Verkivker, G. M.; Bouzida, D.; Gehlaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand–protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731–751.
- (30) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions. I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
- (31) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
- (32) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
- (33) DUD: A Directory of Useful Decoys. <http://dud.docking.org/> (accessed November 2, 2010).
- (34) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein–ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.
- (35) Hanley, J. A.; McNiel, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.
- (36) Nicholls, A. What Do We Know? Simple Statistical Techniques That Help. http://cups2010.eyesopen.com/attachments/0000/6947/ant_on_basic_statistics.pdf (accessed November 2, 2010).
- (37) Nicholls, A. What do we know and when do we know it?. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.