JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

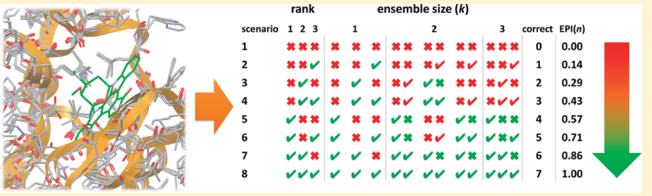
pubs.acs.org/jcim

The Ensemble Performance Index: An Improved Measure for Assessing Ensemble Pose Prediction Performance

Oliver Korb,* Patrick McCabe, and Jason Cole

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

ABSTRACT: We present a theoretical study on the performance of ensemble docking methodologies considering multiple protein structures. We perform a theoretical analysis of pose prediction experiments which is completely unbiased, as we make no assumptions about specific scoring functions, search paradigms, protein structures, or ligand data sets. We introduce a novel interpretable measure, the ensemble performance index (EPI), for the assessment of scoring performance in ensemble docking, which will be applied to simulated and real data sets.



INTRODUCTION

In the past, different measures have been proposed for assessing the performance of protein−ligand docking approaches with respect to pose prediction and virtual screening.

Pose Prediction. A common measure for the assessment of pose prediction performance is the root-mean-square deviation (rmsd) of atomic coordinates. The rmsd is calculated between the heavy atom coordinates of the predicted and the experimentally observed ligand structure. Yusuf et al.[1] introduced a different measure based on the real space *R*-factor (RSR), which compares the expected electron density as calculated from the predicted ligand conformation with the experimentally observed electron density. Another method called interactions-based accuracy classification (IBAC)[2] assesses the similarity of ligand poses based on the interactions formed by the ligand with the receptor.

Virtual Screening. In the area of virtual screening, several metrics have been published trying to characterize the ability of docking programs to discriminate between active and inactive compounds. Apart from enrichment factors, most of these measures are based on the area under an accumulation curve (AUC). AUC values of 0.5 and 1.0 characterize random and perfect classifiers, respectively. Examples are the receiver operator characteristic (ROC),[3] the Boltzmann-enhanced discrimination of ROC (BEDROC),[4] and robust initial enrichment (RIE).[5] While the ROC measure captures the discrimination performance across the whole database, the last two measures deal with the early recognition problem, i.e., identifying active ligands in the top percentages of the screened database.

Currently, the same measures are applied when using multiple protein structures in an ensemble docking protocol. In this approach, instead of performing a docking run considering binding site flexibility, multiple discrete protein structures are used to represent a structural ensemble. A ligand is then docked either sequentially into each individual ensemble structure or a time-efficient optimization methodology considering all ensemble structures at the same time is applied.[6−8] Starting from single-receptor docking results, the final ligand score can be calculated in different ways. Apart from selecting the ensemble structure with the best score across the whole ensemble, also a consensus score may be calculated, as shown by Paulsen et al.[9] In their study for each ligand either the average score over all ensemble members is calculated assigning the same weight to each member or a Boltzmann-weighted averaging scheme is used.

In this study we concentrate on the pose prediction problem in the context of ensemble docking. We introduce a new performance measure which is highly interpretable, as it corresponds to the fraction of correct pose prediction solutions expected from an exhaustive enumeration of all possible ensembles containing up to *n* protein structures.

METHODS

Assessing Scoring Performance in Ensemble Docking. In previous studies, the ability of scoring functions has been assessed with respect to predicting ligand conformations as well as binding affinities and to the closely related problem of discriminating between ligands and decoys in virtual screening experiments. However, when it comes to ensemble docking, another degree of freedom is added to the scoring problem. In this scenario, given a ligand and multiple protein structures, the scoring function also has to assign the best scores to correctly posed ligands. The pose prediction performance, i.e., in how many protein structures the ligand pose can be correctly predicted, is therefore dependent on the cross-docking performance and more importantly on the ranking of the docking solutions according to scoring function value.

Ensemble Docking. We simulate ensemble docking by postprocessing single-receptor docking results. Given a set of *n* protein structures, the ligand is docked into each structure, and the scoring function value as well as the outcome of the pose prediction assessment are recorded. Any measure, for example, the rmsd or the RSR mentioned previously, can be used to assess the correctness of the predicted pose. The postprocessing step is illustrated in Figure 1 for $n = 3$ protein structures. In this example,

Received: June 17, 2011
Published: October 01, 2011

dx.doi.org/10.1021/ci2002796 | *J. Chem. Inf. Model.* 2011, 51, 2915–2919

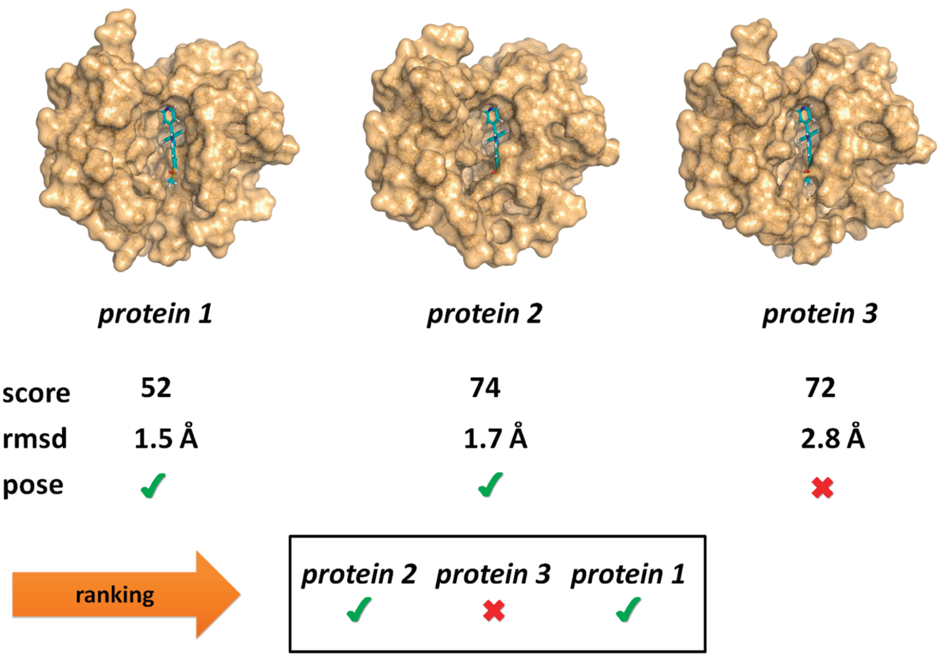


Figure 1. Illustration of the ensemble docking process. One ligand is docked into all available protein structures, and the predicted pose is compared to the experimentally observed structure. A pose is assessed as correct (tick) if the rmsd of the predicted compared to the experimentally observed structure is lower than 2 Å and as incorrect (cross), otherwise. Finally, the solutions are ranked according to decreasing scoring function values.

S	rank 1 2 3	ensemble size *k* = 1	ensemble size *k* = 2	ensemble size *k* = 3	correct	EPI
1	✗✗✗	✗ ✗ ✗ (0)	✗✗ ✗✗ ✗✗ (0)	✗✗✗ (0)	0	0.00
2	✗✗✓	✗ ✗ ✓ (⅓)	✗✗ ✗✓ ✗✓ (0)	✗✗✓ (0)	1	0.14
3	✗✓✗	✗ ✓ ✗ (⅓)	✗✓ ✓✗ ✗✗ (⅓)	✗✓✗ (0)	2	0.29
4	✗✓✓	✗ ✓ ✓ (⅔)	✗✓ ✓✓ ✗✓ (⅓)	✗✓✓ (0)	3	0.43
5	✓✗✗	✓ ✗ ✗ (⅓)	✓✗ ✗✗ ✓✗ (⅔)	✓✗✗ (1)	4	0.57
6	✓✗✓	✓ ✗ ✓ (⅔)	✓✗ ✗✓ ✓✓ (⅔)	✓✗✓ (1)	5	0.71
7	✓✓✗	✓ ✓ ✗ (⅔)	✓✓ ✓✗ ✓✗ (1)	✓✓✗ (1)	6	0.86
8	✓✓✓	✓ ✓ ✓ (1)	✓✓ ✓✓ ✓✓ (1)	✓✓✓ (1)	7	1.00

Figure 2. Exhaustive enumeration of all possible scoring scenarios (S) and ensembles given one ligand and three protein structures ($n = 3$). The scoring function value of the solutions decreases from rank 1 to 3. A tick corresponds to a correct docking solution, while a cross corresponds to an incorrect one. In the ensemble enumeration a green solution (the ensemble begins with a tick) is correct, and a red one is incorrect (the ensemble begins with a cross). For each scenario and ensemble size k, the respective EPI_{subset} values are presented in parentheses. The last columns report the number of correct ensembles and EPI values, respectively.

one ligand is docked to the three different protein structures, and therefore, for each protein structure a scoring function value of the predicted ligand structure is available. There is no restriction on the type of protein structures used. In a second step, the correctness of each docking result is assessed, for example by calculating the rmsd or RSR measure mentioned before. In the example, a predicted pose is assumed to be correct if the rmsd is less than 2 Å. Note that the decision whether a pose is correct is exclusively up to the docking protocol used. While in the example only the rmsd of the docking solution is used, more sophisticated decision procedures may be employed. In the last step a ranking is constructed by ordering the pose prediction results, according to decreasing scoring function values.

This ranking step is sensitive to the accuracy of the docking/scoring methodology used. Imagine, for example, the methodology reliably produces single docking results within 5 scoring function units. Then, in the example above, the potential final rankings could either be 2, 3, 1 or 3, 2, 1, which could essentially result in different ensemble docking solutions. As the ranking sensitivity is potentially different for different docking programs, scoring functions, and target combinations, the analysis of this issue is beyond the scope of this publication.

Given the three ranked docking solutions in the example above, it is possible to construct up to seven ensembles containing at least one protein structure. More specifically, there are three ensembles containing exactly one protein structure (protein [1], [2], and [3]), three ensembles containing exactly two (proteins pairs [2, 3], [2, 1], and [3, 1]) and one ensemble containing exactly three protein structures (protein triple [2, 3, 1]). An ensemble is said to be correct if the top-ranked docking solution of the ensemble is also assessed as having a correct pose and to be incorrect otherwise. In the above example five out of the seven ensembles, i.e., ensemble [2], [1], [2, 3], [2, 1], and [2, 3, 1], are therefore correct (also see scoring scenario 6 in Figure 2). Thus, given the above input ranking derived from the three single docking calculations, about 71% of all ensembles which could be constructed out of the three protein structures would result in a correct docking solution.

While in this example the fraction of correct ensembles was derived by exhaustive enumeration, we will introduce a method which obviates the need for this enumeration. We propose a novel measure, the ensemble performance index, for the assessment of ensemble scoring performance. Given a set of protein structures, a ligand pose is predicted in each protein structure and compared to the experimentally observed ligand conformation. These results can then be ranked by the associated scoring function value such that the best-scored ligand pose is ranked first.

Ensemble Performance Index. Starting from the set of n protein structures, there are $2^n - 1$ ensembles consisting of at least one protein structure, and there are exactly $\binom{n}{k}$ ensembles of size k. A correct ensemble of size k by definition has a correct pose at the highest ranked position (lowest rank index) within that ensemble. Let us consider how an ensemble of size k is selected. If the first choice is from rank one in the original ranking, the remaining $k-1$ selections can be chosen from ranks 2 to n, a set of size $n-1$. Similarly, if the first selection at rank r is a correct pose, the remaining $k-1$ selections can only be made, due to the ordering within the ranking, from ranks $r+1$ to n, a set of size $n-r$. There are thus $\binom{n-r}{k-1}$ correct ensembles when the pose at rank r is correct. Now, due to the generation of subsets of size k, the first choice can only be made from ranks 1 to $n-k+1$. If the first choice is made from any position with higher rank index, then there are not enough elements remaining to generate a subset of size k. If the pose at rank r is incorrect, then the resulting ensemble is incorrect. Thus to count the total number of correct ensembles of size k, we must sum, over those ranks leading to a subset of size k, the number of correct ensembles for a particular rank r. Thus the number of correct ensembles for a specific ensemble size k is given by

$$N(k) = \sum_{r=1}^{n-k+1} I_r \binom{n-r}{k-1} \tag{1}$$

where I_r is a function which returns 1 if the pose at rank r is correct and 0 otherwise.

The fraction of correct ensembles can then be obtained by normalizing this value by the total number of ensembles of size k. This results in the ensemble performance index which depends on the maximum ensemble size n, i.e., the total number of protein structures, and on k, the size of the chosen subset:

$$\mathrm{EPI}_{\mathrm{subset}}(n,k) = \frac{N(k)}{\binom{n}{k}} \tag{2}$$

which ranges from 0 to 1. We will now derive a compact representation for the total number of correct ensembles N by summing over all possible ensemble sizes up to the maximum ensemble size n.

$$\text{Let } N = \sum_{k=1}^{n} N(k) = N(1) + N(2) + \ldots + N(n-1) + N(n)$$

Using eq 1 for $N(k)$ we get

$$\begin{aligned} N &= \sum_{r=1}^{n}\binom{n-r}{0}I_r + \sum_{r=1}^{n-1}\binom{n-r}{1}I_r + \ldots + \sum_{r=1}^{2}\binom{n-r}{n-2}I_r \\ &+ \sum_{r=1}^{1}\binom{n-r}{n-1}I_r = \binom{n-1}{0}I_1 + \binom{n-2}{0}I_2 + \ldots + \binom{1}{0}I_{n-1} \\ &+ \binom{0}{0}I_n + \binom{n-1}{1}I_1 + \binom{n-2}{1}I_2 + \ldots + \binom{2}{1}I_{n-2} + \binom{1}{1}I_{n-1} \\ &+ \ldots \\ &+ \binom{n-1}{n-2}I_1 + \binom{n-2}{n-2}I_2 + \binom{n-1}{n-1}I_1 \end{aligned}$$

Thus,

$$\begin{aligned} N = I_1 \sum_{r=0}^{n-1}\binom{n-1}{r} + I_2 \sum_{r=0}^{n-2}\binom{n-2}{r} + \ldots \\ + I_{n-1}\sum_{r=0}^{1}\binom{1}{r} + I_n\binom{0}{0} \end{aligned}$$

Using the well-known result:

$$\sum_{i=0}^{n}\binom{n}{i} = 2^n$$

N simplifies to

$$\begin{aligned} N &= I_1 2^{n-1} + I_2 2^{n-2} + \ldots + I_{n-1}2 + I_n \\ &= \sum_{r=1}^{n} I_r 2^{n-r} \end{aligned}$$

Thus, the total number of correct ensembles over all possible ensembles of size one or greater is given by

$$N = \sum_{r=1}^{n} I_r 2^{n-r} \tag{3}$$

In order to obtain the ensemble subset size-independent ensemble performance index, the total number of correct ensembles is normalized by the total number of ensembles of size one or greater, i.e., $2^n - 1$, resulting in

$$\mathrm{EPI}(n) = \frac{N}{2^n - 1} \tag{4}$$

Like the AUC measure used in virtual screening calculations,[3] this measure has a number of useful properties. The values of EPI range from 0 to 1, where a value of 0 indicates that none of the poses in the list of single protein pose prediction results were correct, while a value of 1 indicates that all poses were correct. Furthermore it accounts for the ranking of correct and incorrect pose prediction solutions obtained for the different protein structures in a sensible way. We will show that a value greater than 0.5 implies that the top-ranked pose is correct, and thus increasing the ensemble size k up to the maximum ensemble size n will necessarily converge to a correct ensemble (the correct, top-ranked pose will be selected in the largest ensemble of size n). If the top-ranked pose is wrong, the EPI value will be lower than 0.5, and the ensemble of maximum size n will be incorrect.

Proofs

1. A correct pose at rank one will result in an EPI value of greater than 0.5. A correct pose at rank 1 by definition implies $I_1 = 1 \Rightarrow N \geq 2^{n-1} \Rightarrow \mathrm{EPI}(n) \geq 2^{n-1}/(2^n - 1)$. Now $2^{n-1}/(2^n - 1) > 1/2 \Leftrightarrow 2^n > 2^n - 1$, which is true for $n = 0, 1, 2, \ldots$ Hence, with a correct pose at rank 1, $\mathrm{EPI}(n) > 1/2$■.
2. An incorrect pose at rank 1 and correct poses from ranks 2 to n result in an EPI value of less than 0.5. In this case $I_1 = 0$, $I_2 = I_3 = \ldots = I_n = 1$

$$\therefore N = \sum_{r=2}^{n} 2^{n-r} = \sum_{r=0}^{n-2} 2^r$$

Summing this series we get $N = 2^{n-1} - 1$. Now $\mathrm{EPI}(n) = (2^{n-1} - 1)/(2^n - 1) < 1/2 \Leftrightarrow 2^n - 2 < 2^n - 1 \Leftrightarrow 2^n < 2^n + 1$, which is true for $n = 0, 1, 2, \ldots$ Hence, with an incorrect pose at rank 1, $\mathrm{EPI}(n) < 1/2$■.

dx.doi.org/10.1021/ci2002796 |J. Chem. Inf. Model. 2011, 51, 2915–2919

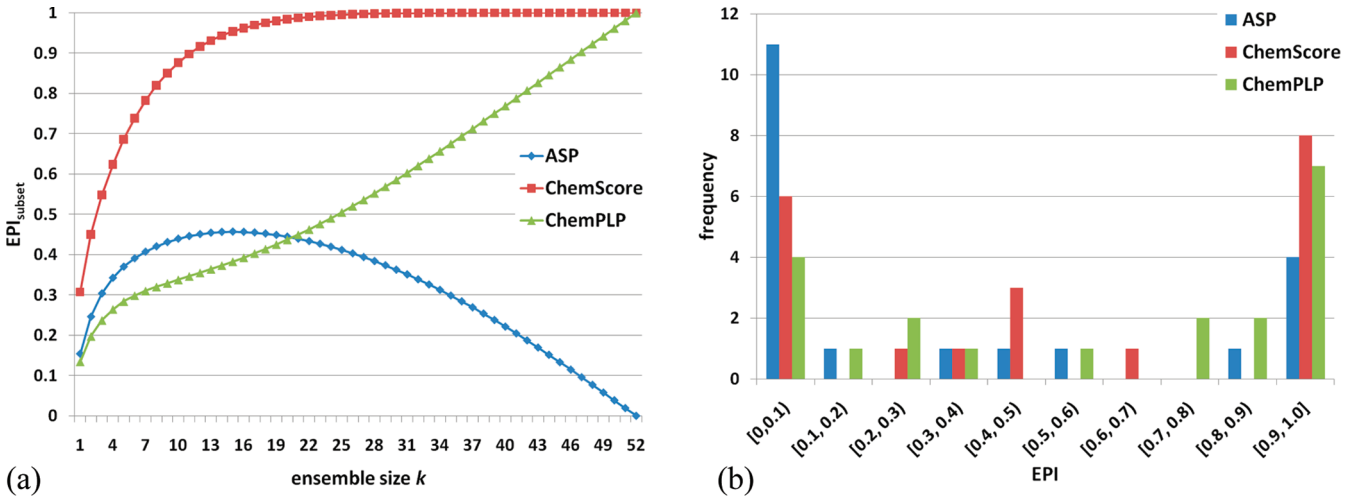


Figure 3. Comparison of scoring functions ASP, ChemScore, and ChemPLP when using $n = 52$ cyclin-dependent kinase 2 protein structures. (a) Distribution of EPI_{subset} values for ligand staurosporine (PDB code 1aq1). (b) Histogram of EPI values obtained for 20 cyclin-dependent kinase 2 ligands.

The interpretation of this measure is intuitive, as changing an incorrect pose to a correct one or improving the rank of a correct pose by at least one will always increase the EPI value. When applying the opposite operations, the EPI value will decrease.

Proof

Changing an incorrect pose to a correct one at rank k increases the EPI value.

This is simply due to I_k changing from 0 to 1 in N in eq 4, and thus N increases by 2^{n-k}■.

Higher EPI values indicate a higher pose prediction performance when using ensemble docking. Although the number of possible ensembles, $2^n - 1$, grows exponentially with n, EPI values can be easily calculated and offer an objective way of comparing the performance of different scoring functions. Furthermore, they are comparable across different targets and also different numbers of protein structures.

Note that compared to the single-ligand case, calculating the number of correct solutions for multiple ligands docked to the same set of protein structures is nontrivial, as not every ligand will be predicted correctly in the same protein structures. We therefore recommend analysis of the distribution of EPI values obtained for multiple ligands docked with different scoring functions.

■ DISCUSSION

Simulated Results. Figure 2 illustrates the outcome of an exhaustive enumeration of all possible scoring scenarios given one ligand and three protein structures. In total there are $2^3 - 1 = 7$ ensembles of size one or greater and $2^3 = 8$ scoring scenarios. In each scenario, the solutions are sorted in descending order with respect to their scoring function value from left to right and are either marked as correct (tick) or incorrect (cross) with respect to the pose prediction assessment. Thus, in the case where the full ensemble of size three is selected, the first (highest scoring) solution will always be selected. This will result in an incorrect solution being selected in scenarios 1−4 and a correct solution being selected in scenarios 5−8. Note that in the latter scenarios, the actual number of correct solutions is less important compared to a single correct solution being top ranked. The last column reports the EPI value for each scoring scenario. As derived above, scenarios 1−4 result in EPI values lower than 0.5, while for scenarios 5−8, due to a correct solution in the top rank, EPI values higher than 0.5 are observed. Each scoring scenario can be seen as a different scoring function applied to the same ligand and the same set of protein structures. Like the AUC values used in virtual screening experiments, the EPI values can thus be used for comparing the performance of different scoring functions.

Docking Results. In order to show the usefulness of the measure, we applied it to assess pose prediction results of cyclin-dependent kinase 2 obtained for three different scoring functions available in GOLD.[10]

We used the cyclin-dependent kinase 2 data set of the Astex non-native set,[11] as it has been used for the assessment of pose prediction performance in the context of ensemble docking. From the set of 72 ligand-bound protein structures, 20 ligands (protein data bank (PDB)[12] codes 1aq1, 1ckp, 1di8, 1e1x, 1e9h, 1fvt, 1jsv, 1oiq, 1p2a, 1pxj, 1vyz, 1w0x, 1y8y, 1y91, 1ykr, 2b54, 2btr, 2c68, 2c6i, 2duv) have been extracted, and their ligand geometries were recreated using Corina.[13] GOLD version 5.0 was then used to dock each Corina-generated ligand 25 times into the 52 remaining noncognate protein structures (autoscale 1.0, 20 genetic algorithm runs, no early termination) using the scoring functions ASP,[14] ChemScore,[15−17] and ChemPLP.[18] As all protein structures in the test set are superimposed, the binding site was defined based on the ligand given in the Astex diverse set[19] (PDB code 1ke5). All protein residues within 6 Å from any ligand heavy atom were considered in the docking calculation. For each experiment the top-ranked pose over all 25 runs was saved and assessed in terms of pose prediction accuracy. An experiment was successful if the heavy-atom coordinates of the predicted and experimentally determined ligand pose had an rmsd of less than 2 Å.

Figure 3a shows the EPI_{subset} values for ligand staurosporine (PDB code 1aq1) when using all three scoring functions. The EPI values for ASP, ChemScore, and ChemPLP are 0.4, 0.99, and 0.52, respectively. Note that in the single-receptor case, ASP correctly predicts the ligand in 8 out of the 52 protein structures, while ChemPLP does so in only 7 cases. However, for ChemPLP the highest scoring protein−ligand complex across the 52 protein structures is correct, while for ASP it is incorrect.

This explains why ChemPLP converges to a correct ensemble of size 52 and ASP does not. As stated previously, this behavior can be predicted from the EPI values of 0.4 and 0.52 for ASP and ChemPLP, respectively. ChemScore predicts the ligand correctly in 16 out of the 52 protein structures and scores correct solutions on average higher than incorrect ones. Compared to the other scoring functions, ChemScore shows the highest ensemble docking performance observed for this ligand. A histogram of EPI values obtained for all 20 ligands is presented in Figure 3b. ASP, ChemScore, and ChemPLP reach EPI values greater than 0.5 for 6, 9, and 12 ligands, respectively. Overall, the latter two scoring functions seem to be the most suitable ones for this target when used in the context of ensemble docking.

CONCLUSIONS

We have introduced a novel measure, the ensemble performance index (EPI), for assessing the scoring performance in the context of pose prediction when using ensemble docking. Ensemble subset size-dependent EPI_{subset} as well as ensemble subset size-independent EPI formulas have been derived which assess cross-docking and the discrimination performance between correctly and incorrectly predicted ligand poses at the same time. The values are highly interpretable as they represent the fraction of ensembles with a correctly predicted ligand pose that would result from an exhaustive enumeration of all possible ensembles. EPI values greater than 0.5 additionally imply that the highest scoring protein−ligand complex was predicted correctly, and thus the ensemble of maximum size, i.e., containing all protein structures, will result in a correct pose prediction too. The opposite is the case for values below 0.5. The usefulness of the measure has been demonstrated for a simulated and a real-world data set. It allows for an objective comparison of different scoring functions in the context of ensemble docking. We strongly encourage researchers to use this measure in future comparisons of scoring functions.

AUTHOR INFORMATION

Corresponding Author

*E-mail: korb@ccdc.cam.ac.uk. Telephone: +44-1223763923.

ACKNOWLEDGMENT

The authors thank Dr. Colin Groom for carefully reading the manuscript and for helpful discussions. O.K. was funded through a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

REFERENCES

(1) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An Alternative Method for the Evaluation of Docking Performance: RSR vs RMSD. *J. Chem. Inf. Model.* **2008**, *48*, 1411–1422.

(2) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J.-Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlén, M.; Stouten, P. F. W. Assessment of Docking Poses: Interactions-Based Accuracy Classification (IBAC) versus Crystal Structure Deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871–881.

(3) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(4) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(5) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395–1406.

(6) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377–395.

(7) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-Dimensional Docking: A Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking. *J. Med. Chem.* **2009**, *52*, 397–406.

(8) Huang, S.-Y.; Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 399–421.

(9) Paulsen, J. L.; Anderson, A. C. Scoring Ensembles of Docked Protein:Ligand Interactions for Virtual Lead Optimization. *J. Chem. Inf. Model.* **2009**, *49*, 2813–2819.

(10) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.

(11) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214–2225.

(12) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899–907.

(13) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev.* **1993**, *93*, 2567–2581.

(14) Mooij, W. T. M.; Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 272–287.

(15) 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.

(16) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503–519.

(17) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609–623.

(18) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84–96.

(19) 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.