ACS OMEGA

Article

Cite This: ACS Omega 2018, 3, 5704−5714

A Diverse Benchmark Based on 3D Matched Molecular Pairs for Validating Scoring Functions

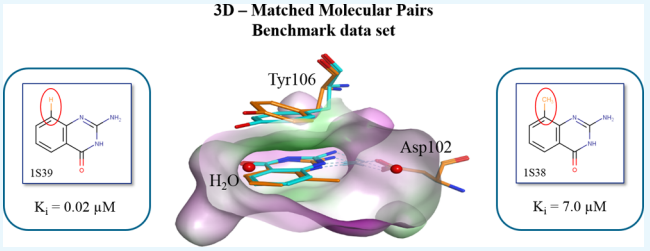
Lena Kalinowsky,[†,§] Julia Weber,[†,§] Shantheya Balasupramaniam,[‡] Knut Baumann,*[,‡] and Ewgenij Proschak*[,†]

[†]Institute of Pharmaceutical Chemistry, Goethe University Frankfurt, Max-von-Laue Str. 9, Frankfurt am Main D-60438, Germany

[‡]Institute of Medicinal and Pharmaceutical Chemistry, University of Technology of Braunschweig, Beethovenstr. 55, Braunschweig D-38106, Germany

Supporting Information

ABSTRACT: The prediction of protein−ligand interactions and their corresponding binding free energy is a challenging task in structure-based drug design and related applications. Docking and scoring is broadly used to propose the binding mode and underlying interactions as well as to provide a measure for ligand affinity or differentiate between active and inactive ligands. Various studies have revealed that most docking software packages reliably predict the binding mode, although scoring remains a challenge. Here, a diverse benchmark data set of 99 matched molecular pairs (3D-MMPs) with experimentally determined X-ray structures and corresponding binding affinities is introduced. This data set was used to study the predictive power of 13 commonly used scoring functions to demonstrate the applicability of the 3D-MMP data set as a valuable tool for benchmarking scoring functions.



INTRODUCTION

Since the 1980s, a variety of docking and scoring methods have been developed, which are used for three main purposes: the prediction of the bioactive conformation of a known active ligand, virtual screening to identify new ligands for a specific target, and the prediction of binding affinities for a series of related compounds.[1] In a recently published comparative assessment of scoring functions, 20 commercially and freely available scoring functions were evaluated in terms of "docking power", "ranking power", and "scoring power" using a diverse test set of 195 protein−ligand complexes.[2,3] The docking power evaluates the ability to identify the active binding mode among a decoy set of ligand binding poses. The ranking power evaluates the ability to rank known ligands according to their binding affinities. The scoring power evaluates the ability to generate scores that are (preferably) linearly correlated with the experimental binding data. Li et al. showed that the evaluated functions performed better in the docking power test than in the scoring/ranking power test.[2,3] These results support the common assumption that the "docking" problem has been solved for the case of rigid receptors, whereas the "scoring" problem still remains a major challenge.[4] Unfortunately, current scoring functions are still far from being able to accurately predict the binding free energy of a protein−ligand complex. Additionally, the inclusion of solvation and rotational entropy contributions as well as protein reorganization energy in the calculation of the binding free energy remains critical.[5−8] Furthermore, most of the scoring functions assume the binding affinity to consist of the sum of several independent terms, which often leads to scores that correlate with the molecular size rather than with binding affinity.[4,9]

To demonstrate the predictive power and to investigate the strengths and weaknesses of scoring functions, several benchmark test sets have been developed.[10−12] These data sets are characterized by their high diversity in terms of protein families, ligand chemotypes, and binding affinities. The high diversity is well suited for the evaluation and comparison of the global performance of docking and scoring software. However, understanding the local behavior of a scoring function, for example, how well it can differentiate between similar molecules, is almost impossible with these data sets. Here, a novel benchmark data set based on matched molecular pairs (MMPs) was developed to study the local behavior of scoring functions. MMPs are defined as molecules that differ in one well-defined transformation associated with a change in an arbitrary molecular property (transformation effect).[13] The PDBbind core set[14,15] forms the basis of the diverse data set containing 99 co-crystallized MMPs (3D-MMPs) stored together with the transformation effect on the binding affinity of the corresponding ligands. The assembled 3D-MMP data set was used to investigate whether the scoring functions can correctly differentiate between chemically related compounds (i.e., the pairwise ranking power was assessed). Therefore, the 3D-MMPs were scored in the respective crystal structures without any posing (i.e., the position

Received: August 16, 2017
Accepted: January 22, 2018
Published: May 28, 2018

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

Table 1. Prediction Accuracy (in %) of the 13 Scoring Functions Calculated at UF (First Value) and UB (Second Value)[a]

	3D-MMPs with |Δaffinity| ≥ 0.5		3D-MMPs with |Δaffinity| ≥ 1.0		3D-MMPs with |Δaffinity| ≥ 2.0	
scoring function[b]	w/ water[c]	w/o water[d]	w/ water[c]	w/o water[d]	w/ water[c]	w/o water[d]
Affinity dG	55.2/56.9	**63.8/74.1**	56.4/59.0	61.5/**71.8**	73.3/73.3	66.7/73.3
London dG	56.9/56.9	53.4/60.3	53.8/56.4	56.4/61.5	60.0/73.3	66.7/73.3
Alpha HB	55.2/58.6	51.7/60.3	59.0/64.1	61.5/**66.7**	73.3/73.3	60.0/73.3
ASE	60.3/60.3	55.2/58.6	56.4/56.4	53.8/56.4	66.7/66.7	46.7/53.3
GBVI/WSA dG	46.6/58.6	48.3/55.2	43.6/61.5	48.7/56.4	53.3/66.7	46.7/60.0
ChemScore	48.3/46.6	56.9/51.7	43.6/53.8	61.5/53.8	60.0/60.0	53.3/53.3
GoldScore	55.2/50.0	**62.1**/48.3	53.8/56.4	61.5/59.0	53.3/60.0	53.3/60.0
ChemPLP	53.4/50.0	53.4/50.0	53.8/53.8	56.4/53.8	66.7/60.0	60.0/60.0
ASP	55.2/50.0	55.2/53.4	51.3/48.7	53.8/51.3	60.0/60.0	60.0/73.3
AutoDock	48.3/–	50.0/–	53.8/–	51.3/–	66.7/–	60.0/–
AutoDock Vina	37.9/–	46.6/–	38.5/–	43.6/–	46.7/–	53.3/–
DSX	–/–	58.6/–	–/–	61.5/–	–/–	66.7/–
X-Score	–/–	**69.0**/–	–/–	**66.7**/–	–/–	73.3/–
Consensus	58.6/53.4	**63.8/62.1**	59.0/53.9	59.0/59.0	66.7/60.0	60.0/60.0

[a]The ability of the scoring functions to predict the direction of a transformation effect (positive or negative) with and without the consideration of water is shown. Results significantly different from chance (i.e., 50%) are in bold. [b]Scoring functions tested for the prediction accuracy of the transformation effect. The prediction accuracy of the majority vote over all scoring functions for each 3D-MMP is listed as "Consensus". [c]Scoring under consideration of water. [d]Scoring without consideration of water. σ_{crit}: [60.3% 64.1% 73.3%] for n: [58 39 15]; Note that σ_{crit} is the critical value, whereas n represents the respective subset size; The calculations of the critical value are provided in the Supporting Information; values are given for subsets 1−3.

of the small molecule was not changed) to focus on scoring and to exclude the influence of posing (i.e., the placement algorithm). Thirteen well-established scoring functions were included in the study covering a broad range of different scoring technologies. Not included were the recent machine-learning−based scoring functions. It has been shown that the machine-learning part may greatly improve the scoring and ranking power. Setting up the machine-learning part of the scoring functions needs a training data set whose source also commonly is the PDBbind database.[16−21] Hence, the complexes of the data set proposed here may already be known to the respective machine-learning−based scoring function, which would bias their results in the benchmark. Although it cannot be ruled out that some or all of the complexes were used to parametrize one or several of the studied scoring functions, the influence of being included in the training set of a machine-learning−based scoring function on the resulting scoring power is expected to be far greater than in cases of classically parametrized scoring functions. In the former case, the machine-learning−based scoring function simply needs to recall the result of the respective complex. As a result, this initial analysis of the scoring power was restricted to classically parametrized scoring functions.

RESULTS

A diverse benchmark data set of 99 3D-MMPs associated with 33 diverse target clusters is assembled. The detailed composition of the data set is described in the Supporting Information (Table S1). For each target cluster, three 3D-MMPs are selected. The transformation effect on the binding affinity of the corresponding ligands is calculated as follows: first, the logarithm (base 10) of the affinity data is taken. Second, the difference in the logarithmically transformed data is computed, where the identity of the minuend and subtrahend is stored. The thus-obtained difference is referred to as Δaffinity.

Various scoring functions are used to score the data set of 99 3D-MMPs, and most of them are available within the commercially available software MOE 2014.09[22] (London dG, ASE, Affinity dG, Alpha HB, GBVI/WSA dG) and GOLD Suite 5.2.2[23−30] (ASP, ChemPLP, ChemScore and GoldScore). To ensure that the results are not artifacts from the preparation and processing of the 3D-MMPs, the computations are carried out independently in two different laboratories at the Goethe University Frankfurt (UF), Germany, and the University of Technology Braunschweig (UB), Germany. In addition to the nine aforementioned scoring functions, UF also evaluated the freely available docking tools AutoDock 4.2.6[31,32] and AutoDock Vina 1.1.2,[33] as well as the independent scoring functions X-Score[34] and DSX.[35] For some scoring functions, small values are optimal, whereas for others, large values indicate a high binding affinity. The latter scoring functions are multiplied by a negative one so that all of the scoring functions are commensurate. The difference in scores for a particular pair is referred to as Δscores. To compute this value, the minuend and subtrahend are the same as those used for computing Δaffinity. Because of this preprocessing, a positive Δscore represents an increase in the predicted binding affinity caused by the transformation, whereas a negative Δscore represents a decrease in the predicted binding affinity. The consensus for the predictions (i.e., the pairwise ranking power) was determined as the majority result of each single ranking over all scoring functions.

The protein−ligand complexes are scored by the different scoring functions with and without considering the water molecules in the active site (except for the DSX and X-Score, where water is not parameterized). The analysis is carried out on different subsets. First, a subset containing only those transformations with an absolute transformation effect (i.e., |Δaffinity|) of at least 0.5 log units (n = 58) is selected to eliminate the impact of experimental uncertainties.[36] This subset is referred to as subset 1. Out of subset 1, subsets with absolute transformation effects of at least 1.0 (n = 39) and 2.0 log units (n = 15) are used to study whether increasing the activity difference improves the results. These subsets are referred to as subsets 2 and 3, in which subset 3 is a subset of subset 2.

The ability of the scoring functions to predict the direction of a transformation effect (positive or negative) is examined first (Table 1). This pairwise test between the two members of the

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

respective 3D-MMPs basically amounts to checking the sign of Δscore and Δaffinity. If the affinity is improved by the molecular transformation, the score of the transformed molecule should also improve. Because there are only two possible outcomes (correct vs incorrect prediction), the sign test can be used to assess the statistical significance. If the number of correct predictions is larger than σ_{crit}, the scoring function yields a prediction accuracy that is significantly different from chance (i.e., 50%). It should be noted that this does not tell anything about the mechanism that led to the prediction accuracy better than chance.[37] Overall, X-Score reaches the highest prediction accuracy of 69.0% in subset 1. No scoring function reaches a prediction accuracy significantly different from chance in subset 3. Significant results are only obtained when water is not considered. The prediction accuracy is also computed for the 9 scoring functions and the various subsets studied at UB (Table 1, second value). Only Affinity dG, London dG, Alpha HB, and consensus scoring achieve statistically significant results, with the best prediction accuracy in subset 1 of 74.1% (Affinity dG). The prediction accuracy for 3D-MMPs with activity constants smaller than 1 μM and 3D-MMPs after geometry optimization can be found in the Supporting Information (Tables S2−S4).

Furthermore, the relationship between the prediction accuracy and the size of the molecules is analyzed (Table 2). The number of heavy atoms is used to characterize the molecular size. 3D-MMPs with (approximately) identical binding affinity are excluded from the analysis (n = 6). Previous studies have shown that the additive character of many scoring functions leads to higher scores as the size of the molecules increase[9] and that simply using the molecular size leads to better results than using a scoring function.[4] In fact, Δaffinity increases in nearly 72% of the 3D-MMPs when the molecular size increases (not counting molecules of equal size). Owing to the latter fact, the relationship between score improvements and size increases could not directly be studied as it has been done previously because affinity acts as a confounder in this case. Put differently, size increases and concordant score increases may actually be caused by an increase in the affinity and may not be governed just by a molecular size bias of the scoring function. To control for the confounder, three subgroups were built: (1) 3D-MMPs where the larger molecule is a more potent binder (n = 51), (2) 3D-MMPs where the smaller molecule binds more potently (n = 20), and (3) 3D-MMPs of equal molecular size (n = 22). The prediction accuracy of the direction of the transformation effect is then analyzed in these three subgroups. The analysis is restricted to complexes without water, as this yields better results on an average, and to the entire set and subset 1 ($|\Delta affinity| \geq 0.5$) as even in subset 1 the subgroups are already rather small. The results based on the scorings at UB are shown in Table 2. It can be seen that most scoring functions perform better in the case where the larger molecule is a more potent binder, whereas the opposite case is more difficult to predict correctly. In the case of equally sized molecules, there is no clear trend, but many scoring functions do not perform better than chance. For some scoring functions such as London dG and Affinity dG, there is clear preference for assigning better score to a larger molecule. They perform very well on the subset where the larger molecule is more affine, whereas they perform badly in the opposite case. As opposed to this, GBVI/WSA dG and ChemScore perform equally good (poor) irrespective of the molecular size.

Table 2. Prediction Accuracy (in %) of the Scoring Functions for All 3D-MMPs (First Value) and Subset 1 (Second Value) Calculated at UB[a]

scoring function	larger mol. is more affine (n = 51/30)	smaller mol. is more affine (n = 20/14)	equal size (n = 22/14)
Affinity dG	88.2/86.7	35.0/35.7	59.1/85.7
London dG	78.4/80.0	25.0/28.6	63.6/50.0
Alpha HB	76.5/73.3	45.0/50.0	50.0/42.9
ASE	58.8/56.7	35.0/48.9	77.3/78.6
GBVI/WSA dG	58.8/56.7	55.0/50.0	63.6/57.1
ChemScore	54.9/56.7	55.0/57.1	40.9/35.7
GoldScore	49.0/53.3	40.0/35.7	50.0/50.0
ChemPLP	60.8/63.3	40.0/35.7	45.5/35.7
ASP	64.7/63.3	25.0/21.4	59.1/64.3
Consensus	78.4/76.7	35.0/35.7	59.1/57.1

[a]The ability of the scoring functions to correctly predict the more affine molecule of the pair in different subgroups is shown. The subgroups were built on the heavy atom count of the ligands.

To examine the validity of the underlying setup, UB processed the CSAR-NRC benchmark data set with the same protocol as that used for the 3D-MMP data set and compared the results against the previously published results.[38,39] The first analysis attempts to reproduce the results of Smith et al.[38] In the aforementioned study, the correlation between the experimentally measured binding affinity data for 332 energy-optimized complexes of the CSAR-NRC data set, excluding the crystal structures of Factor Xa, and the corresponding scores is determined. ASE and Affinity dG are studied by Smith et al., as well as in this work. For ASE, the Pearson correlation is 0.61 (Figure S1A), and for Affinity dG, a value of 0.51 is obtained (Figure S1B). A one-to-one comparison is not possible because the authors of the previous study used pseudonyms for the scoring functions. However, the two values obtained in our analysis can be compared to the distribution of values previously published. As a minimum requirement, the determined correlation coefficients for ASE and Affinity dG are larger in magnitude than that of the weakest method of Smith et al., with a Pearson correlation coefficient of 0.35. The resulting correlation coefficient for Affinity dG is included in the confidence interval of the scoring functions ranked 12−16 (out of 17; Table 1 of ref 38), and that for ASE is included in the confidence interval for scoring functions ranked 3−16. Hence, the results obtained here are within the distribution of the previously published values, with the result for ASE being centrally located in that distribution, whereas that for Affinity dG is in the last third.

The second validation study attempted to reproduce the study of Corbeil et al.[39] The correlation between the experimentally determined binding affinity data for the energy-optimized complexes of the entire CSAR-NRC data set and the corresponding scores of GBVI/WSA dG is determined. The coefficient of determination R^2 of GBVI/WSA dG published by Corbeil et al. is 0.30, whereas that determined according to the protocol used at UB is 0.29 (Figure S2). Hence, the employed protocol used here yields results that are in close agreement with those of Corbeil et al.

The UF laboratory analyzed the transformations in which many scoring functions failed to predict the correct effect. Figure 1 shows that the substitution of a heterocycle led to a significant increase in the binding affinity of the corresponding ligand. All of the scoring functions predicted the opposite effect. Apparently, the hydrophobic interaction between fluorine and the side chains of methionine Met126 and leucine Leu123 led to a higher contribution to the binding affinity from the scoring functions

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

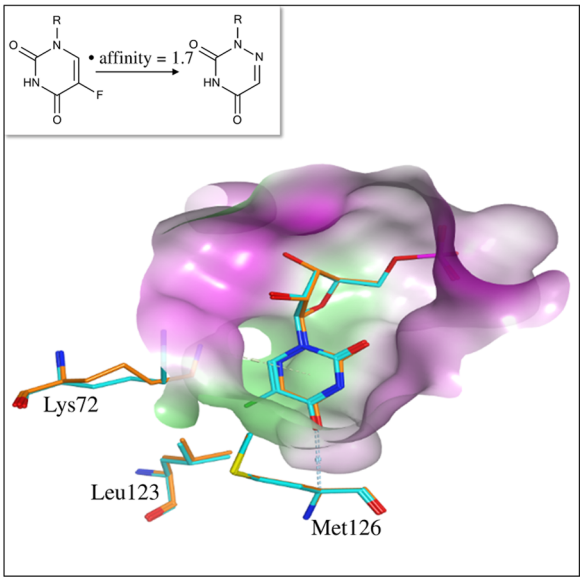


Figure 1. Heterocycle substitution. A 3D-MMP representing two orotidine 5′-monophosphate decarboxylase ligands is shown. The binding affinity increased by 1.7 orders of magnitude upon substitution of the heterocycle. 5-Fluoro-uridine-5′-monophosphate is shown in orange (PDBcode: 3G1V), and 6-aza uridine-5′-monophosphate is shown in blue (PDBcode: 1KM3). The surface of the binding pocket is colored by its lipophilicity (white: neutral, green: lipophilic, magenta: hydrophilic).

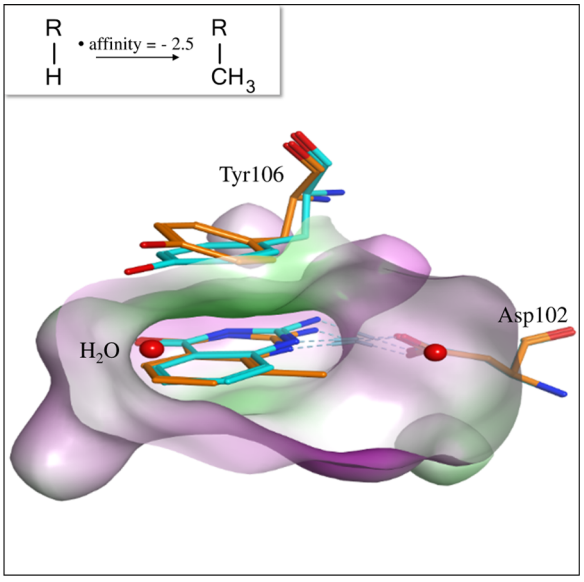


Figure 2. Methylation of an aromatic ring. A 3D-MMP representing two tRNA-guanine transglycosylase inhibitors is shown. The binding affinity decreased by 2.5 orders of magnitude upon methylation of the aromatic ring. The unsubstituted analogue is shown in blue (PDBcode: 1S39), and the methylated analogue is shown in orange (PDBcode: 1S38). The surface of the binding pocket is colored by its lipophilicity (white: neutral, green: lipophilic, magenta: hydrophilic).

than the electrostatic interaction between the electronegative heterocycle and the lysine Lys72 side chain. Furthermore, the Lys72 side chain undergoes considerable conformational change, which might lead to a significant change in the conformational strain of the protein, which is not considered by the most scoring functions, as well as the change in the conformational strain of the ligand caused by the exchange of the 6-carbon to nitrogen.

In Figure 2, the methylation of an aromatic moiety is shown to lead to a loss in binding affinity of more than 2 orders of magnitude. The side chain of Tyr106 undergoes a shift, which might weaken the aromatic interactions with the heterocycle. All of the scoring functions predicted a gain in binding affinity when the water molecules are not considered. Only two scoring functions (Affinity dG and Alpha HB) correctly predict the transformation direction when including water in the active site.

Figure 3 shows the substitution of pyrazole on a benzene ring. This substitution leads to a decrease in the binding affinity of more than 1 order of magnitude. The smaller heterocycle allows the ligand to move deeper into the binding site and changes the distance between the pyrazole nitrogen and the amino group of Lys38, leading to a more favorable hydrogen bond distance.[40] Only one scoring function (ASE) correctly predicts the transformation direction when considering water.

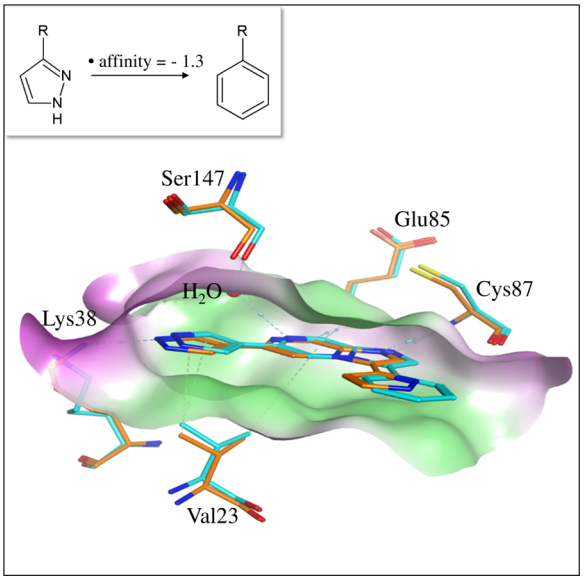


Figure 3. A 3D-MMP representing two checkpoint kinase 1 ligands is shown. The binding affinity decreased by 1.3 orders of magnitude upon substitution of a heterocycle to the benzene ring. The heterocycle analogue is shown in orange (PDBcode: 2XEZ), and the benzene analogue is shown in blue (PDBcode: 2XF0). The surface of the binding pocket is colored by its lipophilicity (white: neutral, green: lipophilic, magenta: hydrophilic).

DISCUSSION

A diverse benchmark data set of 99 3D-MMPs was assembled (Figure 4). In this study, 13 commonly used scoring functions were evaluated, and it was shown that the concordance between the change in affinity and the change in score (i.e., use of the sign of the score differences as a predictor for the change in affinity) is rather mediocre. Even in subsets in which the pairs have large affinity differences, the results were not generally improved. This is rather unexpected because it should be easier for a scoring function to differentiate two molecules with very different affinities. Focusing on the more potent compounds (i.e., affinity < 1 μM) again eases the differentiation of the two molecules and consequently improves the results in many cases (see the Supporting Information, Table S2). However, there is no clear pattern with respect to the performance of the scoring functions. It should be noted that owing to a decreasing size of the subsets with larger affinity differences, larger values of the prediction

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

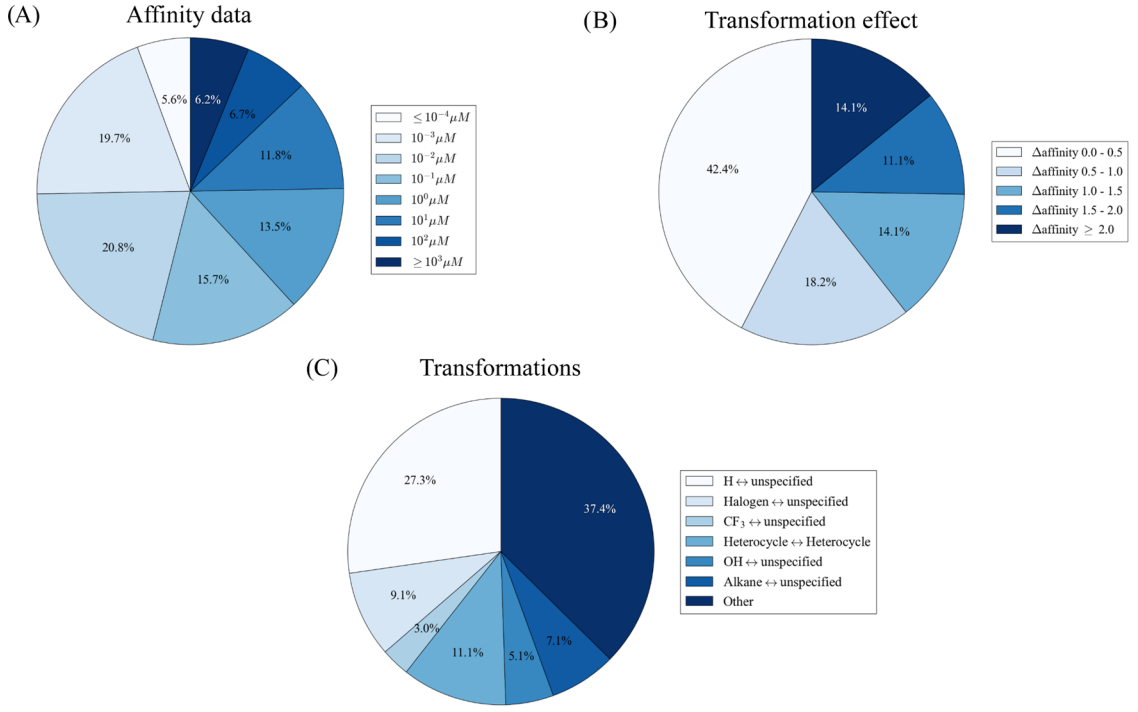


Figure 4. The profile of the data set. The affinity of the ligands was widely distributed from the sub-nanomolar to the two-digit-micromolar level, with a slight emphasis on the rather potent sub-micromolar ligands. In 42.4% of all of the cases, Δaffinity ranged from 0.0 to 0.5 log units. In this case, the ligands can be considered as almost equipotent. However, more than half of the transformations within a 3D-MMP are larger than 0.5 log units and can be regarded as markedly different.

accuracies are more likely just by chance. Hence, the critical values also increase, which renders the head-to-head comparison of differently sized subsets difficult.

In addition to the prediction accuracy of Δaffinity based on Δscore (Table 1), the relationship between the molecular size (expressed as the number of heavy atoms) and the prediction accuracy was studied (Table 2). This was done in subgroups because there is a strong relationship between molecular size and affinity. In 71.8% of the cases (51 out of 71), the larger molecule was more active, a trend well known from lead optimization. The remaining molecules showed either no difference in affinity (n = 6) or were of equal size (n = 22). It was shown that it is easier to correctly predict the more affine molecule in the 3D-MMP if it is the larger one and vice versa. Hence, there is a general preference to assign the larger molecule a better score irrespective of the actual affinity. However, the strength of this preference varies largely across different scoring functions and is absent for some scoring functions. The preference for larger molecules is strong for London dG, Affinity dG, and the consensus prediction and slightly less pronounced for Alpha HB (Table 2). Exactly these scoring functions yielded statistically significant results for the prediction accuracy at UB (Table 1). Hence, it is likely that these prediction accuracies turned significant owing to the strong preference for larger molecules given that the data set is composed of mainly 3D-MMPs where the larger molecule is more affine. Put another way, it is likely that the data set composition acted as confounder in these cases. This shows that a careful subgroup analysis is important to identify potential confounders. Unfortunately, the interesting subgroups where the more affine molecule is smaller and the subgroup of equally sized molecules is rather small in this data set. It would be beneficial to enlarge these groups in future releases of the benchmark set to eliminate this potential confounder.

The computations were carried out independently in two different laboratories to remove artifacts from preprocessing of the complexes and to strengthen the results. For preprocessing and processing the complexes, slightly different protocols were used deliberately. A detailed description of these steps can be found in the Materials and Methods section. Consequently, the obtained results vary, which is a phenomenon well known in the literature.[39] The prediction accuracy varies up to 15% in both directions. The median difference is about 5%. Taking a closer look at the data without considering water (i.e., columns 2, 4, 6, and 8 of Table 1), some trends can be seen. In the entire set and subset 1, the UFs results for GoldScore are better (and statistically significant). This effect could be due to a slight difference in the binding site definition at UF and UB; in particular, at UB, a slightly extended binding site was defined. Another striking difference is the better performance of all of the MOE scoring functions at UB, which can be traced back to the differently employed force fields (MMFF94x@UF vs Amber10:-EHT@UB).

The differences in the preparation of these data sets in terms of protonation and minimization of the protein−ligand complexes are difficult to control apart from the obvious changes in the settings (e.g., the force field). Yet, controlling the preprocessing steps is important, as these processes can lead to completely different scoring results.[41] Some examples were given above. In light of the sources of variability, it should be recalled that the comparison between UB and previous studies by Smith et al.[38] and Corbeil et al.[39] revealed no striking differences, supporting the validity of preprocessing as it was done here.

The main advantage of the 3D-MMP data set is that it is suited for the global benchmarking of scoring functions. The relatively small size and high diversity of the 3D-MMP data set make it a valuable tool for the evaluation of the strengths and weaknesses

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

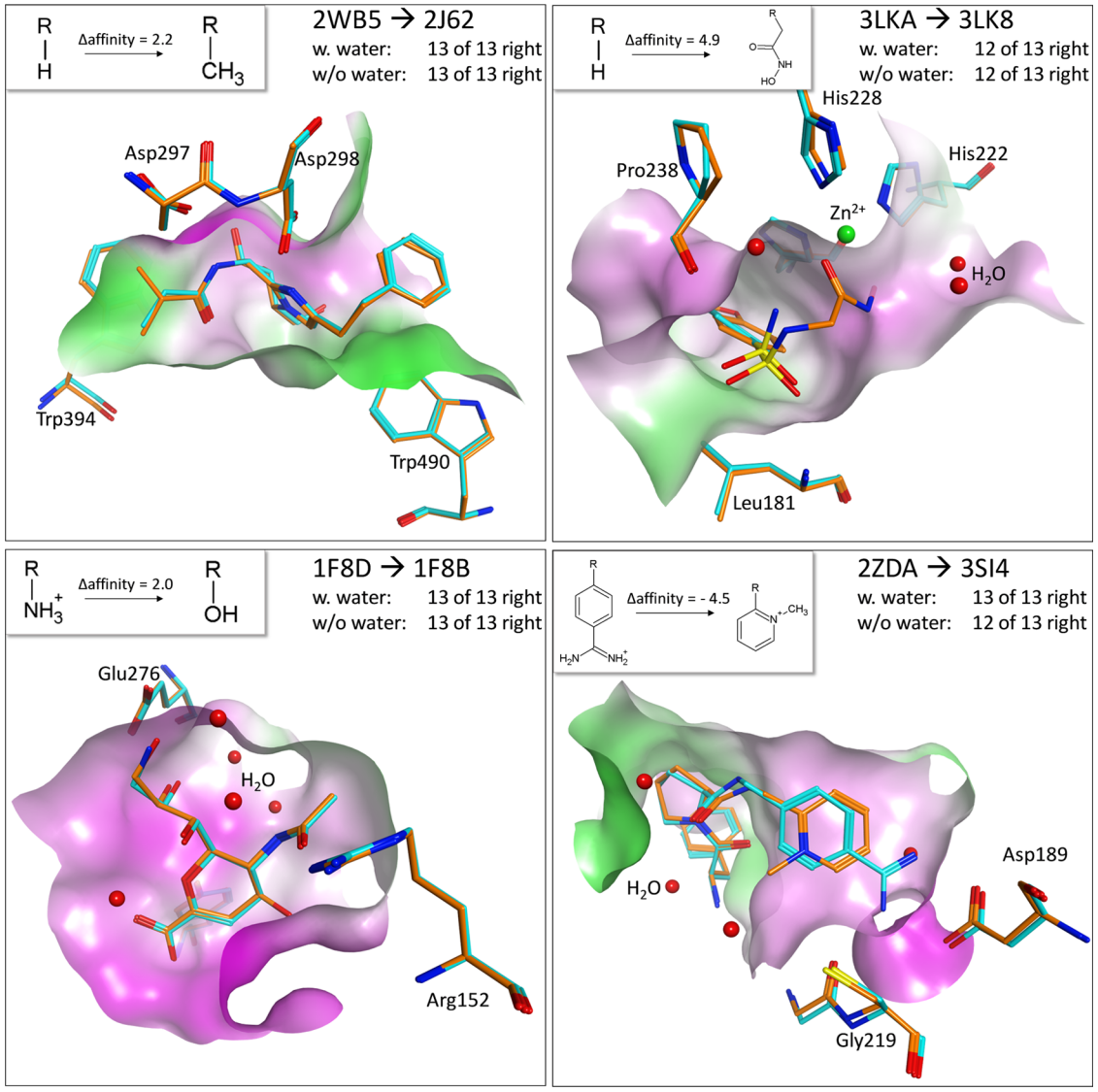


Figure 5. Four 3D-MMPs where at least 12 out of 13 functions were able to predict the correct transformation direction are shown. The binding affinity increased or decreased at least by 2.0 orders of magnitude after the shown substitutions. Ligands from PDBcodes 2WB5, 3LKA, 1F8D, and 2ZDA are shown in blue. Ligands from PDBcodes 2J62, 3LK8, 1F8B, and 3SI4 are shown in orange. The surfaces of the binding pockets are colored by their lipophilicity (white: neutral, green: lipophilic, magenta: hydrophilic).

of individual and consensus functions. Cases in which the scoring functions were not able to predict the correct sign change in the affinity of the transformation are shown in Figures 1−3, and these cases exemplify the current challenges of the available software tools.

Figure 5 shows the substitutions, which seem to be well represented by the terms of most scoring functions. The introduction of an additional lipophilic group leading to an increased activity is recognized by all of the scoring functions. Furthermore, the introduction of additional functional groups leading to the formation of novel, clearly defined interactions, such as metal chelation of ionic interactions, is well represented by all of the scoring functions. In addition, all of the scoring functions clearly recognize the exchange of a functional group by another one with more favorable directed interactions. However, the success in these cases might also arise from the inclusion of the shown examples into the training set of the scoring functions. These examples show that some interactions are very well represented by the scoring functions, whereas others are not. The 3D-MMP data set can provide a valuable basis for the identification of the strengths and weaknesses of the scoring functions, which will ultimately lead to improvements in docking and scoring software.

■ CONCLUSIONS

In this study, we compiled a novel data set of 3D-MMPs, which is suitable for benchmarking and fine-tuning of scoring functions. The data set comprises 99 3D-MMPs, which are highly diverse in terms of target proteins, chemical chemotypes of the co-crystallized ligands, binding affinities, and differences in binding affinities. Although the size of the data set is rather small, it offers the possibility to examine cases of erroneous and correct predictions manually. By this, it provides detailed insights into the capabilities of scoring functions and other methods for structure-based affinity comparison. In this study, we demonstrated the applicability of the 3D-MMP data set to performance evaluation of scoring functions and identified systematic and individual weaknesses of these. The 3D-MMP data set can be

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

used for further optimization and evaluation of algorithms for structure-based computer-aided drug design in the future.

MATERIALS AND METHODS

Basis. The PDBbind v2014[14,15] forms the basis of the diverse benchmark data set of 3D-MMPSs. The PDBbind provides a broad collection of binding affinity data for protein−ligand complexes embedded in the Protein Data Bank (PDB).[42] PDBbind v2014 comprises 44 569 complexes formed between proteins, ligands, and nucleic acids. Furthermore, three subdivided sets of different size and quality are available (general set, refined set, and core set).[14,15]

Clustering. All of the data were processed using the workflow management tool KNIME (Konstanz Information Miner, KNIME Analytics Platform 2.10.1, KNIME.com AG, 2014).[43] In the first step, PDB codes from the PDBbind general set were assigned to new cluster IDs, where one cluster corresponds to one target protein. This assignment was carried out using UniProtKB (UniProt Knowledgebase) numbers[44] via mapping on the UniProt homepage (www.uniprot.org). Due to ambiguity in the assignment of the PDB codes to UniProtKB numbers (i.e., one PDB code with two different UniProtKB numbers), the assignment had to be extended. By employing a KNIME workflow with the *SubGraph Extractor* node (available in the KNIME Labs), a connection network between the PDB codes and UniProtKB numbers was established. The *Object Inserter* node is used to define nodes (PDBcodes) and edges (UniProtKB numbers). *Row to Network* and *Node* Table nodes are then used to aggregate the PDBcodes corresponding to the same UniProtKB number forming a new cluster corresponding to a new cluster ID. To ensure correct clustering of the new cluster IDs, sequence alignment (sequence identity > 90%) was carried out. The new cluster IDs based on the PDBbind general set were used to assign the PDB codes from the PDBbind core set with the same newly created cluster IDs. Because the PDBbind core set is already a diverse set of target proteins, we used the target proteins from the core set to build up our new benchmark data set. Therefore, the 68 cluster IDs from the core set were used to collect all of the corresponding PDB codes from the general set. This results in 3958 PDB codes belonging to the 68 core set cluster IDs.

3D-MMP Generation. MMPs were generated using the available *Matched Pairs Detector* node (provided by Erl Wood Cheminformatics) in KNIME.[45−47] Sequence alignment was applied (sequence identity > 90%) to ensure affiliation to the same target protein (cluster ID) within one MMP. By calculating the root-mean-square deviation (RMSD) from the maximum common substructure (MCS), the orientation and location of the ligands within one MMP were analyzed. An RMSD < 1 Å of the MCS was necessary to provide comparability of the ligands within one MMP. In cases where the RMSD was larger than 1 Å, structural alignment (superposition) was conducted. Therefore, a superposition of the proteins was carried out, and a rotation matrix for each protein was obtained. These rotation matrices were used to rotate the ligands in the same manner. Afterward, the RMSD was calculated again, and the MMPs with an RMSD < 1 Å were collected. This alignment of complex coordinates led to 3D-MMPs. Restriction rules for the collected 3D-MMPs were implemented in the next step. The maximum size of a cyclic substituent was limited to nine nonhydrogen atoms, and a noncyclic substituent was limited to five nonhydrogen atoms. The common core in the molecules was restricted to be at least 50% of the size of the entire molecule. Furthermore, the measured binding affinities within a 3D-MMP were of the same experimental type (K_i, K_d, or IC_{50}). 3D-MMPs with IC_{50} values were only accepted if they were obtained from the same publication. This rule was applied to avoid inaccuracies because of the high dependence of the assay on the conditions. To achieve diversity with respect to the target and the binding affinity effect (Δaffinity), three 3D-MMPs were selected for each target cluster with the smallest, largest, and mean Δaffinity. After applying all of the substituent restriction and diversity rules, a diverse data set of 99 3D-MMPs, corresponding to 33 target clusters, was obtained.

Quality Assessment of the Data Set. To ensure the high quality of the data set, all 99 3D-MMPs X-ray structures (178 protein−ligand complexes) were reviewed. The resolutions of the protein−ligand X-ray structures were used for an initial global assessment of quality. A mean resolution of 1.90 Å was achieved in our data set. However, this value only describes the theoretical limit on the precision of the model. It does not provide any quality information about the specific parts of the structure, such as the ligand or binding pocket. The additional global quality measures are the refinement *R*-factors (R_{work}, R_{free}). R_{work} is a measure of the difference between the measured data and the model-predicted data. Therefore, R_{work} values can be used to estimate the model quality. Before refinement, a random subset of the data is collected (R_{free} reflections), which is later used for cross-validation to avoid overfitting the data.[48] These global quality measures (resolution, R_{work} and R_{free}) can be found in the Supporting Information (Table S5). For local quality assessment specific to the ligand and the binding pocket, the electron density of all of the ligands was examined using the electron density score for multiple atoms ($EDIA_m$). The EIDA values quantify the electron density fit of an atom by calculating a weighted sum over an oversampled electron density grid in the proximity of the atom.[49−51] According to Friedrich et al.,[49−51] an $EDIA_m$ value above 0.8 indicates a satisfying fit of a molecular structure on the observed electron density. The $EDIA_m$ values for the data set can be found in the Supporting Information (Table S6). In cases in which an $EDIA_m$ value of 0.8 was not achieved, the electron density maps were manually examined (http://www.ebi.ac.uk/pdbe/eds), and polder maps were generated. A polder map is an OMIT map that excludes the bulk solvent around an omitted region. In our case, the omitted region is the ligand. Polder maps are a helpful tool to visualize weak electron densities around the ligand or other regions of interest. The polder maps were generated using the software Phenix.[52] A detailed table including the electron density and polder maps associated with structures with an $EDIA_m$ value smaller than 0.8 can be found in the Supporting Information (Table S7). Using the global and local quality measures mentioned above, it can be concluded that our 3D-MMP data set fulfills the conventional quality criteria for a high-quality benchmark data set.

Data Set Preparation at Goethe University Frankfurt (UF). Complex Preparation. All of the protein−ligand complexes were downloaded from the PDB. In every case, chain A was used for all further applications. The protonation of proteins and ligands was carried out in a KNIME workflow using the available *Protonate 3D* node (provided by MOE; default settings).[53] Water was taken into consideration to analyze the influence during the predictive power investigation. Therefore, all of the predictions were performed twice, with and without the consideration of water.

Scoring Procedure. Thirteen commonly used scoring functions were evaluated in the study conducted at UF. These

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

included five scoring functions provided in MOE 2014.09[22] (London dG, ASE, Affinity dG, Alpha HB, and GBVI/WSA dG), four scoring functions provided in the software package GOLD Suite 5.2.2[23−30] (ASP, ChemPLP, ChemScore, and GoldScore), the freely available docking tools, AutoDock 4.2.6[31,32] and AutoDock Vina 1.1.2,[33] and the independent scoring functions, X-Score[34] and DSX.[35] Seven of these scoring functions can be considered as empirical scoring functions, three as force-field-based, and another three as knowledge-based. A brief description of these scoring functions can be found in the Supporting Information.

The MOE scoring was realized through the MOE extensions in KNIME, specifically the *complex scoring* node. Receptors were read in as .mol2, and ligands as .sdf. The *complex scoring* node was executed to yield scores for each complex using the five scoring functions implemented in MOE.

The GOLD scoring was realized through a KNIME workflow. GOLD docking or, in our case, re-scoring is based on a configuration file (gold.conf) that contains all of the necessary information. For each scoring function and each complex, a gold.conf file was generated. The gold.conf file is built up of information pointing to the corresponding ligand file (.mol) and protein file (.pdb), the binding site (protein atoms within 5 Å of the ligand), and the scoring function. It also contains information for only the re-scoring of a ligand with no advanced ligand minimization. These gold.conf files were used to run the GOLD re-scoring for all four scoring functions. The scores were extracted from the output files (solutions.rescore.log) for each complex and each scoring function.

Re-scoring with AutoDock 4.2.6 was carried out in a KNIME workflow. Python scripts for all of the preparation steps were available in the AutoDock Tools (ADT) provided in the MGLTools package.[54] AutoDock 4.2.6 required the receptor and ligand file to be written in PDBQT format. The PDBQT format had additional partial charges and AutoDock atom types to the normal PDB format. Receptor and ligand files were prepared using the *prepare_receptor4.py* and *prepare_ligand4.py* Python scripts. Next, a grid parameter file was prepared for AutoGrid 4, which precalculates the grid maps of the interaction energies later used by AutoDock 4.2.6 to determine the total interaction energy for a protein−ligand complex. The grid parameter files were prepared using the *prepare_gpf4.py* Python script. The prepared grid parameter files and receptor PDBQT files were then used to run AutoGrid 4 and generate grid maps (.glg). To run AutoDock 4.2.6, a docking parameter file (.dpf) was needed. This file was generated using the *prepare_dpf4.py* Python script. The obtained docking parameter files had to be modified to only perform re-scoring with AutoDock 4.2.6. Therefore, we used a Python script to remove all of the lines responsible for docking in the .dpf file and append the parameter *epdb*. After adding this parameter, AutoDock 4.2.6 was used to calculate the energy of the ligand provided in the PDBQT ligand file. In the final step, AutoDock 4.2.6 was run using the modified docking parameter file to yield the .dlg result file. As a score, the estimated free energy of binding was extracted from the generated .dlg file using a Python script.

AutoDock Vina 1.1.2, as the successor of AutoDock 4.2, used the same receptor and ligand format (PDBQT). Additionally, only a configuration text file (config.txt) was needed to run AutoDock Vina 1.1.2. This configuration file contains the receptor and ligand PDBQT file, coordinates for the center grid points (coordinates taken from the corresponding .gpf file generated in the AutoDock 4.2.6 workflow) and the number of grid points in each direction ($x = y = z = 40$). After generating the configuration files for all of the complexes, re-scoring with AutoDock Vina 1.1.2 was achieved by using the flag "--score_-only". The resulting scores were extracted from the generated .log files.

The independent scoring function X-Score required a receptor PDB file and a ligand .mol2 file. Using a Python script, re-scoring with X-Score was run in KNIME using the flag "-score". The resulting .log files were used to extract three single scores (HPScore, HMScore, and HSScore). The average of these three scores gave the final X-Score score.

The independent scoring function DSX required a receptor PDB file and a ligand .mol2 file. DSX was run using a simple Java script in KNIME. The resulting text files were used to extract the final DSX scores.

Data Set Preparation at TU Braunschweig (UB). Complex Preparation. The complexes with and without considering water and their energy-optimized forms were processed as follows. The complexes were prepared with MOE 2013.08.[55] Amber10:EHT was chosen as the force field, using the reaction field electrostatics as the solvent model. Each protein was loaded into MOE, and the structural discrepancies detected by the structure preparation tool of MOE were fixed. The default settings for protonation were adjusted to physiological conditions (i.e., T [K] = 310.15, pH = 7.4, salt-conc. [mol/L] = 0.9). Subsequently, partial charges based on the selected force field and hydrogens were added. These prepared proteins were saved as .pdb. The ligands were prepared in the same manner and saved as .mol2 and into a Molecular Database file (.mdb). Energy-optimized complexes were obtained by loading the prepared proteins and their corresponding ligands into MOE, where an active site limited energy minimization was conducted. Upon convergence, proteins and ligands were saved separately, as described previously.

Scoring Procedure. At UB, all of the scoring functions of MOE (London dG, ASE, Affinity dG, Alpha HB, and GBVI/WSA dG) and GOLD Suite 5.2.2[23−30] (ASP, ChemPLP, ChemScore, and GoldScore) were evaluated. The MOE scoring was carried out using KNIME version 2.11.2[56] with the MOE extensions for KNIME (*knimoe* 2.2.0). Proteins were read in as .pdb, and ligands as .mdb. Once the ligands were assigned to their corresponding proteins, the complex scoring node was executed to yield the scores of the five MOE scoring functions for each complex. The force field of the complex scoring node in knimoe was changed to Amber10:EHT before scoring.[57]

GOLD scoring was realized via command line. Therefore, four gold.conf files (one per GOLD scoring function) for each prepared protein−ligand pair were generated. In addition to containing the respective scoring function, each gold.conf file pointed to the corresponding protein and ligand. Furthermore, they contained default settings for re-scoring, with the following changes from the default settings. The binding site was defined as all atoms within 10 Å of the ligand, and the re-scoring options "perform local optimization", "retrieve rotated protein atom positions (if available)", and "replace score tags in file" were disabled.

Moreover, a subset of 3D-MMPs was composed of only more potent compounds in which their complexes had binding affinities of less than 1 μM (n = 54 3D-MMPs, subset 4). This subset was divided again into 3D-MMPs with $\log_{10}$ differences in the affinity equal to or greater than 0.5 (n = 25 3D-MMPs, subset 5) and 1.0 (n = 13 3D-MMPs, subset 6) to result in six subsets.

CSAR-NRC Data Set. The CSAR-NRC data set[38] was downloaded (CSAR-NRC HiQ from www.csardock.org) and

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

processed in the same manner as the complexes in the 3D-MMP data set. The CSAR-NRC data set contains "343 high-quality, protein−ligand crystal structures" and was used by Smith et al.[38] for a benchmark exercise in 2010. This data set was also made available in an energy-minimized form. Participants scored both sets of crystal structures of the CSAR-NRC data set with varying parameters. The results were published using pseudonyms so that only the distribution of the figures of merit could be analyzed.

Each crystal structure in the CSAR-NRC data set was loaded into MOE separately. The ligands were extracted from the complexes and saved into a Molecular Database file (.mdb). Proteins and ligands were prepared and saved in the same way as the proteins and ligands in the 3D-MMP data set. Subsequently, the prepared proteins and their corresponding ligands underwent energy minimization based on the Amber10:EHT force field and were saved as .pdb (protein) and .mdb (ligands). Scoring of the CSAR-NRC data set was conducted in an identical manner to the 3D-MMP data set, as described in the previous section. Because the correlation measures between the scores and affinities published by Smith et al. were restricted to the 332 complexes in the entire CSAR-NRC data set by excluding the crystal structures of Factor Xa, only that subset was used here. From the scoring functions tested at UB, ASE, Affinity dG, and ChemScore were part of the core methods of the benchmark exercise in 2010, which included a total of 17 scoring functions. Within this benchmark exercise, these three scoring functions were tested on the minimized complexes because they better correlated with the experimental data. The use of pseudonyms made a one-to-one comparison impossible. Nevertheless, it was checked whether the correlations of ASE and Affinity dG for the 332 minimized complexes at least outperformed the weakest method of Smith et al. by considering the Pearson correlation coefficients to ensure that the observed results are of sufficient technical quality (i.e., are not artifacts of the employed protocol for preparing and processing the data). Furthermore, Corbeil et al.[39] used the CSAR-NRC data set as a test set to validate the GBVI/WSA dG scoring function. Before applying the GBVI/WSA dG scoring function on the CSAR-NRC data set, Corbeil et al. minimized its complexes based on the MMFF94x force field with reaction field electrostatics. Corbeil et al. used the entire CSAR-NRC data set, including the Factor Xa protein−ligand complexes. For comparison, we also considered the entire energy-minimized CSAR-NRC data set and determined the R^2 accordingly.

Scoring Function Analysis at UF and UB. For a detailed analysis, the 3D-MMPs were divided into different subsets. The subsets comprised 3D-MMPs with $\log_{10}$ differences in the affinity equal to or greater than 0.5 (n = 58 3D-MMPs, subset 1), 1.0 (n = 39 3D-MMPs, subset 2), and 2.0 (n = 15 3D-MMPs, subset 3). The scoring results were extracted and analyzed in two different ways. The predicted transformation effect was calculated by subtracting the single scores for each 3D-MMP (Δscore). The ability of the scoring functions to predict the direction of a transformation effect (positive or negative) with and without the consideration of water (prediction accuracy) was analyzed first (Table 1). This was done for all of the subsets. Furthermore, it was investigated as to what extent the number of heavy atoms (as a surrogate of molecular size) affected the prediction accuracy (Table 2). 3D-MMPs with the same affinity were removed from this analysis (n = 6), which led to a reduced number of 3D-MMPs in the entire data set (n = 93) and in subset 1 (n = 58). Finally, the consensus was determined as the majority vote of all of the scoring functions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b01194.

Description of the scoring functions, detailed composition of the 3D-MMP data set, detailed data on quality the ligand electron density, tables with correlations and critical values (PDF)

3D-MMP data set is available for download (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: k.baumann@tu-braunschweig.de. Phone: +49 531 391 2750 (K.B.).

*E-mail: proschak@pharmchem.uni-frankfurt.de. Phone: +49 69 798 29301 (E.P.).

ORCID

Knut Baumann: 0000-0001-9459-0045

Ewgenij Proschak: 0000-0003-1961-1859

Author Contributions

§L.K. and J.W. contributed equally.

Author Contributions

The manuscript was written through contributions of all of the authors. All of the authors have given approval to the final version of the manuscript. K.B. and E.P. share the senior authorship.

Funding

This research was supported by the German Research Foundation (DFG; Sachbeihilfe PR 1405/2-2; Heisenberg-Professur PR-1405/4-1; Sonderforschungsbereich SFB 1039 Teilprojekt A07).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Wahram Andrikyan for support during quality check of the electron densities and Dr. Guido Kirsten, Chemical Computing Group, for the adaptation of the KNIME node to the Amber force field.

■ ABBREVIATIONS

3D-MMPs, 3D matched molecular pair; MMPs, matched molecular pairs; PDBbind, protein data bank bind; MOE, molecular operation environment; GOLD, genetic optimization for ligand docking; UF, Goethe University Frankfurt; UB, University of Technology Braunschweig; DSX, DrugScoreX; mol, molecule; CSAR-NRC, Community Structure Activity Resource-National Research Council of Canada; R^2, coefficient of determination; PDBcode, protein data bank code; tRNA, transfer ribonucleic acid; PDB, protein data bank; KNIME, Konstanz Information Miner; UniProtKB, UniProt Knowledgebase; RMSD, root-mean-square deviation; MCS, maximum common substructure; K_i, inhibition constant; K_d, dissociation constant; IC_{50}, half maximal inhibitory concentration; R_{free}, free R-factor; R_{work}, work R-factor; EDIAm, electron density score for multiple atoms; .mol, MDL Molfile; .pdb file, protein data bank file; ADT, AutoDock Tools; MGLTools, molecular graphics laboratory tools; .dpf file, docking parameter file; .dlg file, docking log file; .gpf file, grid parameter file; PDBQT, protein

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

data bank, partial charge (Q), atom type (T); .mol2 file, Tripos Mol2 file; HPScore, hydrophobic pair score; HMScore, hydrophobic match score; HSScore, hydrophobic surface score; .mdb file, molecular database file; CSAR-NRC HiQ, Community Structure Activity Resource-National Research Council of Canada High Quality

REFERENCES

(1) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.

(2) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* **2014**, *54*, 1700−1716.

(3) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717−1736.

(4) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore(CSD)-Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of near-Native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.* **2005**, *48*, 6296−6303.

(5) Hunter, C. A. Quantifying Intermolecular Interactions: Guidelines for the Molecular Recognition Toolbox. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 5310−5324.

(6) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.

(7) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

(8) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 45−55.

(9) Schulz-Gasch, T.; Stahl, M. Scoring Functions for Protein-Ligand Interactions: A Critical Perspective. *Drug Discovery Today: Technol.* **2004**, *1*, 231−239.

(10) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX Incremental Construction Algorithm for Protein- Ligand Docking. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 228−241.

(11) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein-Ligand Interaction. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 457−471.

(12) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(13) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *49*, 6672−6682.

(14) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(15) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.

(16) Zilian, D.; Sotri, C. a. SFCscoreRF: A Random Forest-Based Scoring Function for Improved A Ffi Nity Prediction of Protein − Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923−1933.

(17) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Substituting Random Forest for Multiple Linear Regression Improves Binding Affinity Prediction of Scoring Functions: Cyscore as a Case Study. *BMC Bioinf.* **2014**, *15*, 291.

(18) Durrant, J. D.; Mccammon, J. A. NNScore 2.0: A Neural-Network Receptor À Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51*, 2897−2903.

(19) Wang, C.; Zhang, Y. Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.* **2017**, *38*, 169−177.

(20) Li, H.; Leung, K. S.; Wong, M. H.; Ballester, P. J. Improving Autodock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inf.* **2015**, *34*, 115−126.

(21) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(22) *Molecular Operating Environment (MOE)*, 2014.09; Chemical Computing Group Inc.: QC, Canada, 2014.

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

(24) 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.

(25) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein − Ligand Docking Using GOLD. *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 609−623.

(26) CCDC Gold Suite 5.2.2., Cambridge, United Kingdom.

(27) Mooij, W. T. M.; Verdonk, M. L. General and Targeted Statistical Potentials for Protein-Ligand Interactions. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 272−287.

(28) 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.

(29) 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.

(30) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible Docking Using Tabu Search and an Empirical Estimate of Binding Affinity. *Proteins: Struct., Funct., Bioinf.* **1998**, *33*, 367−382.

(31) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. Software News and Update a Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **2007**, *28*, 1145−1152.

(32) Morris, G. M.; Huey, R.; et al. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.

(33) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(34) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

(35) Neudert, G.; Klebe, G. DSX: A Knowledge-Based Scoring Function for the Assessment of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2731−2745.

(36) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data - a Statistical Analysis. *PLoS One* **2013**, *8*, No. e61007.

(37) Matthews, R.; Wasserstein, R.; Spiegelhalter, D. The ASA's P-Value Statement, One Year on. *Significance* **2017**, *14*, 38−41.

(38) Smith, R. D.; Dunbar, J. B.; Ung, P. M. U.; Esposito, E. X.; Yang, C. Y.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Combined Evaluation across All Submitted Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 2115−2131.

(39) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in Docking Success Rates due to Dataset Preparation. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 775−786.

(40) Matthews, T. P.; McHardy, T.; Klair, S.; Boxall, K.; Fisher, M.; Cherry, M.; Allen, C. E.; Addison, G. J.; Ellard, J.; Aherne, G. W.; et al. Design and Evaluation of 3,6-Di(hetero)aryl imidazo[1,2-A]pyrazines

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714

as Inhibitors of Checkpoint and Other Kinases. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 4045−4049.
(41) Schulz-Gasch, T.; Stahl, M. Scoring Functions for Protein−ligand Interactions: A Critical Perspective. *Drug Discovery Today: Technol.* **2004**, *1*, 231−239.
(42) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(43) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kotter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. *KNIME: The Konstanz Information Miner*; Springer, 2007.
(44) Magrane, M.; Consortium, U. UniProt Knowledgebase: A Hub of Integrated Protein Data. *Database* **2011**, *2011*, No. bar009.
(45) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.
(46) Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; et al. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of hERG Inhibition, Solubility, and Lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1872−1886.
(47) Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.
(48) Warren, G. L.; Do, T. D.; Kelley, B. P.; Nicholls, A.; Warren, S. D. Essential Considerations for Using Protein-Ligand Structures in Drug Discovery. *Drug Discovery Today* **2012**, *17*, 1270−1281.
(49) *The Protein Plus Server*. http://proteinsplus.zbh.uni-hamburg.de/ (accessed March 2, 2017).
(50) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 529−539.
(51) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-Ray Structures. *J. Chem. Inf. Model.* **2017**, *57*, 2437−2447.
(52) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; et al. PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr., Sect. D: Struct. Biol.* **2010**, *66*, 213−221.
(53) Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins: Struct., Funct., Bioinf.* **2009**, *75*, 187−205.
(54) The Scripps Research Institute Molecular Graphics Laboratory. MGLTools. http://mgltools.scripps.edu/downloads.
(55) *Molecular Operating Environment (MOE)*, 2013.08; Chemical Computing Group Inc.: Montreal, QC, 2016.
(56) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Koetter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. *Studies in Classification, Data Analysis, and Knowledge Organization*; Springer: Heidelberg, 2007.
(57) *Op_score.svl, Scientific Vector Language (SVL) Source Code*; Chemical Computing Group Inc.: Montreal, QC, 2016.

DOI: 10.1021/acsomega.7b01194
ACS Omega 2018, 3, 5704−5714