Converging a Knowledge-Based Scoring Function: DrugScore^{2018}

Jonas Dittrich[†], Denis Schmidt[†], Christopher Pfleger[†], Holger Gohlke*[†‡]

[†]Mathematisch-Naturwissenschaftliche Fakultät, Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[‡]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC) & Institute for Complex Systems -Structural Biochemistry (ICS 6), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany

Running title: Converging a knowledge-based Scoring Function: DrugScore^{2018}
Keywords: Scoring, ranking, screening, docking, PDB, CASF, drug design, structure-based, protein-ligand complex

*Corresponding Author: Holger Gohlke
Address: Universitätsstr. 1, 40225 Düsseldorf, Germany.
Phone: (+49) 211 81 13662; Fax: (+49) 211 81 13847
E-mail: gohlke@uni-duesseldorf.de

1 Abstract

We present DrugScore^{2018}, a new version of the knowledge-based scoring function DrugScore, which builds upon the same formalism used to derive DrugScore, but exploits a training data set of nearly 40,000 X-ray complex structures, a highly diverse and the by far largest dataset ever used for such an endeavour. About 2.5 times as many pair potentials than before now have a data basis required to yield smooth potentials, and pair potentials could now be derived for eight more atom types, including interactions involving halogen atoms and metal ions highly relevant for medicinal chemistry. Probing for dependence on training data set size and quality, we show that DrugScore^{2018} potentials are converged. We evaluated DrugScore^{2018} in comprehensive scoring, ranking, docking, and screening tests on the CASF-2013 dataset, allowing for a comparison with >30 other scoring functions. There, DrugScore^{2018} showed similar or improved performance in all aspects when compared to either DrugScore, DrugScore^{CSD}, or DSX and was, overall, the scoring function showing a most consistently good performance in scoring, ranking, and docking tests. Applying DrugScore^{2018} as objective function in AutoDock3 in a large-scale docking trial, using 4,056 protein-ligand complexes from PDBbind 2016, reproduced a docked pose to within 2 Å RMSD to the crystal structure in >75% of all dockings. These results are remarkable as the DrugScore^{2018} potentials were derived from crystallographic information only, without any further adaptation using binding affinity or docking decoy data. DrugScore^{2018} should thus be a competitive scoring and objective function for structure-based ligand design purposes.

2 Introduction

In computational chemistry, scoring functions (SFs) aim on evaluating ("score") interactions between binding partners such as in protein-ligand complexes. SFs are frequently used in molecular docking,[1] virtual screening,[2, 3] and other drug discovery applications.[4] Since the early 1990s, a myriad of SFs has been published.[5-19] These SFs can be classified into four categories.[20] I) SFs based on force fields derived from physical principles to describe interatomic repulsion and attraction forces;[21] II) empirical approaches that use measured binding affinities to weight physics-based terms; III) knowledge-based SFs that use experimental structural data to derive statistical, potential of mean force-related preferences; IV) machine learning approaches that consider some or all of the above features and other chemical and structural descriptors of proteins and ligands.

In 2000, we introduced the knowledge-based SF DrugScore.[22, 23] Here, distance-dependent pair potentials were derived from crystal structures of 1,376 protein-ligand complexes, taken from the Protein Data Bank (PDB).[24] DrugScore was the first knowledge-based SF that included both distance-dependent pair potentials and solvent-accessible surface (SAS)-dependent singlet preferences of protein and ligand atoms. Initially, DrugScore was used for scoring given protein-ligand complexes[25-28] and later successfully applied also as objective function in molecular docking[29-33] using AutoDock3[34] as a docking engine. In 2005, pair potentials were derived following the DrugScore formalism but using structural information from the Cambridge Structural Database[35] (CSD) resulting in DrugScore^{CSD}.[36] The large amount of available small-molecule organic and metal-organic crystal structures in the CSD database allowed deriving new potentials for atomic interactions, which were so far not available or underrepresented in DrugScore. Later on, the DrugScore formalism was applied to derive pair potentials to score RNA-ligand complexes[37] (DrugScore^{RNA}) and protein-protein complexes[38] (DrugScore^{PPI}).The latest DrugScore variant is DrugScore eXtended[39] (DSX), which uses 68 atom types as defined by fconv[40] instead of the 18 Sybyl atom types used originally, and newly defined solvent accessible surface-dependent potentials; the addition of novel knowledge-based torsion angle potentials enables one to use DSX as a stand-alone tool for energy minimization of ligand poses.

As with all statistical approaches, the accuracy and scope of knowledge-based SFs strongly depends on the amount of available data for derivation. Over the last years, the number of structures deposited in the PDB increased exponentially due to the improvement of methods for determining biomolecular structures.[41] Currently, the PDB contains more than 143,000 entries with structural information, including >130,000 protein-complexes. On this account,

other popular knowledge-based SFs, which have been originally developed before or around the year 2000, have been re-derived in the past, including the SFs PMF [12, 14] by Muegge *et al.* and SMoG[42, 43] by Shakhnovich *et al.*. Still, the number of structures used for the re-derived PMF04[14] is closer to the number of structures used to derive DrugScore potentials[22] than to the number of available protein-ligand complexes today. Likewise, the knowledge-based part of the hybrid (knowledge-based and empirical) SF SMoG2016 finally was derived from 1,038 protein-ligand complexes, as the SMoG approach only benefited marginally from an increased amount of data.

In this study, we investigated the influence of quality and quantity of structural data used to derive distance-dependent pair potentials for protein-ligand interactions in the context of the DrugScore approach on the performance of this knowledge-based SF. In particular, we aimed to see if converged knowledge-based potentials can be obtained with the current amount of structural data. As a training set for what will be termed DrugScore^{2018} potentials, we used nearly 40,000 protein-ligand complex structures taken from the PDB and, hence, the by far largest dataset ever used for such an endeavour. The dataset was assessed with respect to ligand diversity and drug-likeness as well as a potential bias caused by overrepresented protein classes. The performance of DrugScore^{2018} was evaluated in comprehensive scoring and docking experiments building upon established test sets, allowing for a thorough comparison with more than 20 stand-alone (or part of a software suit) SFs and more than ten SFs implemented in docking programs: (I) We used the "Comparative Assessment of Scoring Functions" (CASF)-2013 test set,[44, 45] which consists of four tests that challenge the *scoring* (correlation of measured and computed affinities), *ranking* (preference of high- over low-affinity ligands), *docking* (reproduction of ligand poses), and *screening* (enrichment of known ligands over decoys) capabilities of SFs and (II) we used DrugScore and DrugScore^{2018} potentials as objective functions in a large-scale docking experiment with 4,056 protein-ligand complex structures taken form the "Refined Set" of PDBbind 2016[46, 47] dataset. Overall, our results reveal that DrugScore^{2018} potentials are converged with respect to training data set size and quality and should be a competitive scoring and objective function for structure-based ligand design purposes.

3 Methods

Deriving DrugScore[2018] Potentials. Protein-ligand complexes for the dataset to derive DrugScore[2018] potentials were downloaded from the PDB (date Feb. 12th, 2017). PDB entries with a resolution larger than 2.5 □ were excluded from the data set in order to reduce the amount of imprecise structures, as done in the original approach.[22] OpenEye's OEChem toolkit[48] (version 2.1.4) was used to assign Sybyl atom types and to separate each PDB entry into the protein (including crystal water), ligand, and other components. Ions were treated as part of the receptor unless they were explicitly part of the ligand. We evaluated distances between ligand atoms and their respective receptor atoms (consisting of protein, ions, and other ligands including cofactors) within a radius of 6 □ around the ligand atoms to emphasize specific interactions formed by a ligand with neighboring residues in the binding site; this limit is short enough not to involve a water molecule as mediator of a ligand-to-protein interaction.[22] To reduce a potential bias of the derived potentials with respect to certain ligands, complexes with ligands that occur more frequently than 500 times within the PDB were excluded. Furthermore, ligands with <10 and >100 heavy atoms were excluded, as were ligands with missing (unresolved) atoms. Finally, not more than four identical ligands per PDB structure were allowed for deriving DrugScore[2018] potentials. However, excluded ligands may still be part of the receptor structure.

Potentials were derived as described for DrugScore potentials.[22, 23] The compilation of distance-dependent pair potentials between ligand atoms *i* and protein atoms *j* grouped by their atom types, $T(i)$ and $T(j)$, respectively, is based on an inverse Boltzmann approach and a formalism developed by Sippl[49] (eq. (1)),

$$\Delta W_{T(i),T(j)}(r) = -\ln \frac{g_{T(i),T(j)}(r)}{g(r)} \qquad (1)$$

where $g_{T(i),T(j)}(r)$ denotes the normalized radial pair distribution for atom types $T(i)$ and $T(j)$, calculated from their respective occurrence frequency in the distance interval $[r; r + dr)$ with dr = 0.1 Å, and $g(r)$ is the normalized mean radial pair distribution function for a distance between two atoms in the interval of $[r; r + dr)$, incorporating all non-specific information common to all atom pairs in the protein-ligand complexes (eqs. (2,3)):

$$g_{T(i),T(j)}(r) = \frac{N_{T(i),T(j)}(r)/4\pi r^2 dr}{\sum_r (N_{T(i),T(j)}(r)/4\pi r^2 dr)} \;;\; g(r) = \frac{\sum_{T(i)} \sum_{T(j)} g_{T(i),T(j)}(r)}{\|T(i)\| \cdot \|T(j)\|} \qquad (2,3)$$

Scoring Protein-Ligand Complexes. The total score describing the protein-ligand interaction is the sum of all occurring atom-atom interactions, i.e., interactions between atoms i of the ligand L and atoms j of the protein P (eq. (4)):

$$\Delta W = \sum_{i \in L} \sum_{j \in P} \Delta W_{T(i),T(j)}(r) \tag{4}$$

In this work we neglect the contribution of the SAS-dependent singlet preferences originally derived for DrugScore, as in subsequent studies predominantly the pair potentials were used as scoring[23, 50] and objective[29-33] function. Therefore, for scoring with DrugScore, DrugScore^{CSD}, and DrugScore^{2018}, we only applied pair potentials derived from the respective dataset (CSD and PDB). For scoring with DSX, we only considered pair potentials derived from the PDB (DSX^{PDB}::PAIR). For a more comprehensive description of the theory behind DrugScore, please see here[22, 23, 36-38].

Assessment of the Training Data Set. The data set for deriving DrugScore^{2018} potentials was evaluated with respect to ligand diversity and the influence of the largest protein cluster on the potentials. To assess ligand diversity, extended-connectivity fingerprints[51] with a radius of two atoms (ECPF4) were generated for each ligand in the training data set, and then all pair-wise Tanimoto similarities were calculated using the OEChem toolkit.[48]

The coverage of the ligand chemical space[52] by the training data set was compared to that by the ChEMBL23 database[53] by means of a principal component analysis (PCA) of the space spanned by the molecular quantum numbers (MQNs).[54] MQNs are a set of 42 1D-descriptors (e.g., atom counts, bond counts, polarity counts, and topology counts), which can be used to map and visualize the chemical space covered by molecules of different databases.[54, 55] MQNs were calculated for all ligands in the training data set and the ChEBML23 database[44] using the MQN software of the Reymond group.[54] Unlike ref.[54], the 42 MQN descriptors were normalized separately to zero-mean and unit variance. The principal components were calculated from the normalized data, and the two sets of ligands were compared in terms of projections onto the first two principal components.

To test for a potential bias due to frequently represented protein families, the used PDB structures were first clustered by their Pfam-IDs.[56] The largest cluster consisted of 1,041 complexes with a single protein kinase domain. The redundant occurrence of certain proteins does not necessarily induce bias in the DrugScore potentials *per se*, as only interactions with the first shell of residues in the binding site are evaluated. Yet, kinase ligands are inherently structurally similar and bind to similar binding sites. Consequently, certain interaction distances

might be overrepresented. Hence, the complexes of the largest cluster were removed from the data set, the potentials were re-derived and compared to those derived from the complete set. The difference of the DrugScore potentials is quantified by a normalized root-mean-square error. This parameter is referred to as Potential Deviation (PD) in this work. The PD between two sets of potentials is calculated according to eq. (5),

$$PD = \frac{\sqrt{\frac{\sum_r \left(W_{T(i),T(j)}(r)_{ref} - W_{T(i),T(j)}(r)_{subset}\right)^2}{n}}}{\max\left(W_{T(i),T(j)}(r)_{ref}\right) - \min\left(W_{T(i),T(j)}(r)_{ref}\right)} \quad (5)$$

where $W_{T(i),T(j)}(r)_{ref}$ is the reference potential for the ligand atom i of type $T(i)$ and the protein atom j of type $T(j)$ as a function of the distance r for the interval of 2 Å to 6 Å, $W_{T(i),T(j)}(r)_{subset}$ is the corresponding potential derived from a subset, and *n* is the number of distance bins. This value is normalized by the range of values ($\max\left(W_{T(i),T(j)}(r)_{ref}\right) - \min(W_{T(i),T(j)}(r)_{ref})$) covered by the reference potential. A PD value of zero indicates identical potentials.

DrugScore^{2018} potentials are derived from the PDB, which includes structures of the PDBbind Refined Set (2016) and the CASF-2013 set that are later used to asses DrugScore^{2018}'s predictive power. To preclude a potential bias in the evaluation, DrugScore^{2018} potentials were additionally derived from the training data excluding all complexes containing proteins present in the PDBbind Refined Set (2016) or the CASF-2013 set. If the potentials without these protein complexes are almost indistinguishable from the DrugScore^{2018} potentials including them, a bias due to the overlap of training and test set can be neglected, and the potentials derived from the whole set can be used in the evaluation.

Convergence of DrugScore Potentials. To assess the convergence of the DrugScore^{2018} potentials as a function of the number of protein-ligand complexes used for derivation, potentials were derived from subsets with quantities of 1000, 5000, 10,000, 20,000, and 30,000 training structures, and the corresponding PD was calculated (eq. (5)). To generate the subsets, structures were randomly chosen from the training data set. For each subset of given quantity, ten independent bootstrap[57] sets were sampled allowing for replacement of chosen structures. The PD values of ten bootstrap samples were averaged, and the standard error (SEM) was calculated.

CASF-2013 Test Set. The distance-dependent pair-potentials of DrugScore^{2018} were evaluated on the CASF-2013 test set provided by Li *et al.*.[44, 45] All atom types of the structures in the CASF-2013 set were assigned using the same routine as described for the training set. The CASF-2013 set provides different tests, each assessing different features and abilities of a SF, such as their scoring, ranking, docking, and screening performance. Moreover, it establishes a basis to compare the performance of different SFs on equal terms. The CASF-2013 set consists of 65 different proteins with known binding ligands, their native and generated docking poses. For detailed explanations, further information on the test set and on the tested SFs, see refs. [44, 45]. In this study, we applied four tests, which are defined in ref.[45] and briefly described below.

Scoring Power Test. This test assesses the correlation of the SF's score for the native pose of the ligand within the complex to the experimentally determined binding energy. This correlation is quantitatively evaluated by Pearson's correlation coefficient R and standard deviation σ between predicted and experimentally determined binding affinities. We also calculated Spearman's rank correlation coefficient R_S, which is the nonparametric analogue of R, and therefore is considered more robust.[58] The calculations were done using the respective functions of the SciPy[59] python module (version 1.1.0).

Ranking Power Test. This test evaluates the SF's ability to rank known ligands of the same target protein by their binding affinities. The test set consists of 65 groups of complexes each formed by a protein with three different ligands. If a SF ranks all three ligands according to their known binding affinity, this is considered a "high-level" success, and if the SF identifies the best binder regardless of the ranking of the medium and poorest binder, this is considered a "low-level" success.

Docking Power Test. Unlike the tests before, this test does not focus on the reproduction of experimental affinities but on the capability to identify the correct ligand pose. For each protein-ligand complex in the test set, the SF scores up to 100 binding poses. These poses were created by Li *et al*. using GOLD[7] (version 5.1), Surflex[60, 61] as implemented in the SYBYL software (version 8.1), and the docking module of the MOE software package (version 2011). The native binding pose was also included in this set to ensure that there is at least one correct binding pose present. The similarity of the best scored and the native pose is expressed by the root mean square deviation (RMSD) (eq. (6))

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2}{n}} \quad (6).$$

Here n is the total number of atoms within the ligand molecule, x_i, y_i, z_i and x_i', y_i', z_i' are the Cartesian coordinates of the i^{th} atom in two binding poses. In order to yield correct results for symmetric ligands, Li *et al.* calculated a property-matched RMSD, called RMSD^{PM}, which matches atom pairs between two binding poses by atom types instead of atom IDs. The RMSD^{PM}, which is provided for each docking pose, was used to evaluate the docking power of the SFs. In this test, the SF is considered successful when it scores a docking pose with an RMSD^{PM} < 2 Å compared to the crystal structure as best pose ("Top1") or within the first two ("Top2") or three ("Top3") best scored poses.

Screening Power Test. This test assesses the SF's ability to distinguish between binding and non-binding ligands. The screening power is evaluated in a cross-docking experiment. All 195 (= 65 x 3) ligands were docked into each of the 65 proteins, yielding 12,675 possible protein-ligand combinations in total. For each of the protein-ligand pairs, up to 50 representative ligand binding poses were generated using above-mentioned docking tools. For each protein, all poses are scored with the SF of interest. Subsequently, ligand molecules are ranked according to the best score of any of its poses. The success of a SF is measured by counting the total number of true binders among 1%, 5%, and 10% of the top ranked molecules.

Docking Into DrugScore^{2018} Fields. The Refined Set of the PDBbind 2016[46, 47] was chosen as large-scale test set for the evaluation of DrugScore^{2018} as an *objective* function in docking. This test differs from the above Docking Power Test in that there, DrugScore^{2018} was evaluated as a *scoring* function for given docked ligand poses. The Refined Set contains 4,056 selected protein-ligand complexes with known binding affinities covering a pK_i range of ten units. To avoid any bias towards the input crystal structure, a single low-energy conformation was additionally generated for each ligand using OpenEye's OMEGA (version 2.5.1.4).[62, 63] The crystal pose of each ligand and the low-energy conformation were re-docked into the receptor crystal structure using AutoDock3[34] (AD3) in combination with the standard energy function of AD3 or the DrugScore and DrugScore^{2018} potentials as objective functions. Additionally, the ligand crystal structures were re-docked using AutoDock4 (AD4).[64]

Structure Preparation. The ligands and proteins were prepared using modified AutoDockTools[64] python scripts for AD3 and AD4 so that (I) all input charges of the ligand were preserved and (II) non-polar hydrogens were merged by adding their charge to the carbon

atom to which they were bonded and removed from the molecule (united-atom model). Kollman charges[65] were added to the atoms of the receptor. Additionally, non-polar hydrogens and lone pairs of the receptor were merged. For docking using DrugScore^{2018} potentials, all atom types of the input protein and ligand mol2 files were assigned using OpenEye's OEChem toolkit as before. The grid calculation was carried out with AutoGrid3 (for AD3), AutoGrid4 (for AD4), or the DrugScore software using the original and new potentials. The standard grid spacing of 0.375 Å was chosen for all dockings. When docking with AD3/4, the standard grid box of 40 x 40 x 40 units (15 x 15 x 15 Å3) was placed such that the ligand is centered in the box. For larger ligands (~10% of cases), AutoDockTools expanded the box dimensions by multiples of the grid spacing (0.375 Å) such that the ligand fits the box. DrugScore automatically ensures a 5 Å margin around the ligand along each axis. 95% of the resulting DrugScore grid boxes are at least as large by volume as the grid box created by AutoDockTools and ~92% of the boxes are at least 10% bigger. Due to the larger configurational space, a positive bias towards DrugScore can be excluded when the docking results are compared, as the smaller grid boxes should favor the original AutoDock software.

Docking Protocol. Following established procedures,[32, 33] the docking protocol considered 100 independent runs for each ligand using an initial population size of 100 individuals, a maximum number of 27.0×10^3 generations, a maximum number of 5.0×10^6 energy evaluations, a mutation rate of 0.02, a crossover rate of 0.8, and an elitism value of 1. For sampling, the Lamarckian Genetic Algorithm (LGA) was chosen in all approaches. In general, the standard settings of AutoDock are used, except for the maximum number of energy evaluations, which is doubled in this experiment to focus more on the accuracy of the objective function than on the efficiency of the sampling algorithm. In order to assess a convergence of the docking solutions, the generated docking poses were clustered as to their RMSD with AutoDock, using the standard cut-off value of 2 Å. Fewer clusters of docking poses indicate a higher convergence of the proposed docking solutions. The docking is considered successful when the best scored pose is within 2 Å RMSD of the native pose. In contrast, the best pose by RMSD was not considered for evaluation, as this would not be known in a prospective approach.

4 Results and Discussion

Assessment of the Training Data Set. The quality of a knowledge-based SF benefits from a large and diverse data set for derivation that is representative of the protein and ligand application domains. Overall, 39,683 selected protein-ligand complexes were used to derive DrugScore^{2018}, covering complexes of 24,570 different proteins and, therefore, representing nearly 20% of all available entries of protein-ligand complexes in the PDB. As to the protein application domain, 75% of the proteins have been classified as enzymes, where the distribution of enzyme classes in the data set closely resembles the distribution within the PDB (identical for EC1, deviation < 3.3% for EC2 and EC3, and deviation < 0.3% for EC4 to EC6). The remaining 25% of protein-ligand complexes include receptors, antibodies, chaperones, transporters, and membrane proteins. As to the ligand application domain, only ligands with 10–100 heavy atoms where considered in the data set, equivalent to molecular masses between ~80–600 Da, i.e., the ligands cover fragment-like to drug-like molecules with respect to size. The ligands are structurally diverse, shown by the fact that >99% of all pairwise ECFP4-Tanimoto similarities are <0.2 (Figure 1A). At the same time, the chemical space covered by the ligands of the data set considerably overlaps with that of the ChEMBL23 database, as demonstrated by a projection of the respective ligands onto the first two principal components of the MQN[43] space (Figure 1B). This is also true when the first nine PCs are taken into account, which explain 60% of the variance in total (Figures S2 and S3). As the ChEMBL23 database contains more than 1.7 million entries of predominantly drug-like chemical compounds based on information extracted from more than 65,000 publications,[66] we considered this database a reliable and adequate reference. This overlap of the chemical space covered by both datasets is remarkable considering that the ChEMBL23 database contains 43 times more molecules than our training set. We conclude that the training data set used to derive DrugScore^{2018} potentials is diverse with respect to proteins and ligands, and representative of the relevant chemical space of small-molecule ligands.

The composition of the training data set is not uniform with respect to protein superfamilies and families. In order to probe to what extent this may bias the derived DrugScore^{2018} potentials, we removed all entries that contained only a protein kinase domain, i.e., the largest cluster of proteins with respect to their Pfam-ID. Potentials re-derived from the reduced data set show, on average, a normalized root mean square deviation (NRMSD), referred to as PD, of <1% from the DrugScore^{2018} potentials derived on the full data set ($\overline{\mathrm{PD}} = 0.0079 \pm 0.0010$; eq. (5)). Similarly, when excluding all structures containing proteins present in the PDBbind Refined Set or containing proteins of the CASF-2013 test set (4,635 protein-ligand complexes in total)

used for evaluating DrugsScore^{2018}'s predictive power, the average NRMSD is ~2% ($\overline{PD} = 0.0242 \pm 0.0025$; eq. (5)). As both of the potentials derived from such reduced data sets are virtually indistinguishable from the DrugScore^{2018} potentials, the latter, derived from the full data set, can be used for testing DrugScore^{2018}'s performance without having to expect a pronounced bias in the results. We note already now that, to our knowledge, the overlap between training data set and data sets used to evaluate a SF's predictive power has not been assessed for other SFs[43, 45, 67] with which we will compare DrugScore^{2018}.

Improved Coverage of Atom Types in DrugScore^{2018}. When DrugScore was first introduced in 2000, the types and quantity of protein-ligand complexes in the training data set allowed deriving interactions between 17 Sybyl atom types. Due to statistical reasons, rare atom types such as N.2 (sp^2-hybridized nitrogen, e.g. in imines) and N.ar (nitrogen in aromatic rings, e.g. in pyridine), or S.2 (sp^2-hybridized sulfur, e.g. in thiocarbonyls), S.3 (sp^3-hybridized sulfur, e.g. in cysteine (thiol group) or methionine (thioether)), S.O (sulfoxide sulfur, e.g. in a thionyl group), and S.O2 (sulfone sulfur, e.g. in sulfonyl groups) were combined to yield sufficiently populated potentials.[22] In DrugScore^{2018}, the 39,683 protein-ligand complexes allowed deriving distance-dependent pair potentials for 25 Sybyl atom types. The additional atom types are I, C.1 (sp-hybridized carbon, e.g. in alkynes), N.1 (sp-hybridized nitrogen, e.g. in a nitrile group), N.2, N.4 (protonated sp^3-hybridized nitrogen, e.g. protonated amino groups), S.2, S.O, and S.O2.

We previously found that potentials derived from ≥ 500 pair interactions (i.e., on average, 10 pair interactions per distance bin) are sufficiently smooth to yield reliable results.[36] The number of pair-potentials with ≥ 500 pair interactions increased from 117 in DrugScore to 289 in DrugScore^{2018} (Figure 2). Particularly, the DrugScore^{2018} potentials now cover also three halogen-mediated interactions and 14 metal-mediated interactions (Figure 2). Panel C and D of Figure 2 exemplarily show how potentials of more rarely occurring interactions (cation-π (N.4 *versus* C.ar) and, likely, σ-hole-mediated interactions (I *versus* O.2)) benefit from an increased amount of structural data. In conclusion, the increased amount of data used to derive DrugScore^{2018} potentials now also allows scoring of interactions that are highly relevant for medicinal chemistry, including those involving halogen atoms[68-71] and metal ions.[72]

Convergence of DrugScore^{2018} Potentials. The convergence of the DrugScore^{2018} potentials was assessed by also deriving the potentials from multiple, randomly sampled subsets with sizes between ~1/40 and ~3/4 of the whole training data set. Convergence was measured

in terms of the PD from DrugScore^{2018}. As the PD calculation is based on bootstrapping with replacement, it yields values >0, including when the complete dataset is used for deriving potentials (see also next paragraph). Potentials of frequently occurring interactions, such as C.3-C.3 (hydrophobic interactions of sp^3-hybridized (aliphatic) carbons), O.3-O.2 (interactions of h-bond donors (e.g. oxygen in hydroxyl groups)) and acceptors (e.g. oxygen in carbonyl groups), and C.ar-C.ar (e.g. π-π stacking of aromatic carbons), are converged already with 5,000 complexes (PD < 1.8%; Figure 3), whereas rare interactions, such as N.1-C.cat (e.g. interaction between sp-hybridized nitrogen (e.g. in a nitrile group) and the carbon of guanidine group of an arginine sidechain) or S.3-Met (interactions of sp^3-hybridized sulfur atoms and metals (e.g. zinc, magnesium) within the binding site of the protein) require up to four-fold more structural data (20,000 complexes; PD < 5.4%; Figure 3). Hence, in general, DrugScore^{2018} potentials are converged when >20,000 complexes are used in the training data set.

Furthermore, the statistical error of the potentials was estimated by bootstrapping with replacement. With the full training data set, the median PD amounts to 3.1% of the range of DrugScore^{2018} potential values (Figure 3A), with 46 (158) potentials showing NRMSDs < 1.5% (3.0%) (Figure 3B). Very similar results were obtained for potentials derived from the same training data set, but now only considering structures with a resolution ≤2.0 Å (23,087 complexes; PD_{median} = 3.8%). Figure 3A shows that the deviation from the final potentials for the smaller subsets converges to the deviation estimated by bootstrapping on the full training data set, i.e. the potentials are within the estimated error of the potentials derived from the whole training set. This error most likely results from the inherent uncertainty of the used experimentally determined protein-ligand complexes. Together with the above analyses on the convergence of DrugScore^{2018} potentials, our results, therefore, indicate that a derivation of pair-potentials from more, and higher-quality, complexes will most likely not result in a qualitative change of potential values or an improvement in their precision, except for non-typical pairs of atom types. However, the influence of such pairs on a SF's overall performance can be expected to be low, though.

Evaluation of DrugScore^{2018} on the CASF-2013 Test Set. CASF-2013 provides different tests, each assessing different features and abilities of a SF, such as their scoring, ranking, docking, and screening performance, thereby establishing a basis to compare the performance of different SFs.[44, 45]

Scoring Power Test. The scoring power test evaluates the quality of the correlation of predicted binding scores with experimentally determined binding energies.[45] The results for the

DrugScore variants is shown in Figure 4. For DrugScore^{2018}, an increase in the scoring power compared to DrugScore and DrugScore^{CSD} by ~4% is observed (Figure 4), which is statistically significant ($p < 0.0001$ according to a Steiger test[73]). One reason for the small improvement in scoring power between DrugScore to DrugScore^{2018}, despite the 40-fold increase in the size of training data set, may be that the CASF-2013 test set comprises only 33 halogen-containing ligands and 23 complexes with a ligand-metal interaction out of 195 complexes, such that the particular improvement in the quality of these potentials in DrugScore^{2018} carries little weight. Indeed, when focusing on these 56 complexes, the increase in the scoring power on going from DrugScore to DrugScore^{2018} is about twice as large (Figure S4).

Compared to other academic and commercial SFs, DrugScore^{2018} ranks among the best three, when R_s is taken as a measure (Table 1). The only SFs for which a significantly ($p < 0.0001$) better R has been reported are BT-Score,[74] AutoDockHybrid,[67] and X-Score^{HM} [8] (when R is taken as a measure). Remarkably, SFs achieving better R/R_s values in this test use either multiple-linear fitting (AutoDockHybrid[67]) or belong to the machine-learning approaches (BT-Score[74]). In this case, the partial overlap of training and test data may contribute to that improvement, as shown for BT-Score in Figure 3 of ref. [74], where the resulting correlation of predicted versus experimental binding affinities depends on the presence of (close) derivatives of the structures in the training set. All SFs fall short of coming close to the theoretical R_{max} of 0.975 (assuming an uncertainty of the experimental values of 0.5 log units and considering the standard deviation of the experimental values of 2.25 log units)[75] (Table 1), indicating that they only provide an approximate estimate for binding affinities, as mentioned before.[76-78] Still, for non-trained (i.e., no data of binding affinities were used for the derivation of the potentials) SFs such as DrugScore^{2018}, the residual range of the regression line of ~1.8 log units (Figure 4) indicates that such SFs may be applied to distinguish potential binders from non-binders.[78-80]

Ranking Power Test. The result of the ranking power test is shown in panel A of Figure 5. The top four SFs, namely DrugScore, DrugScore^{2018}, ChemPLP,[19] and X-Score^{HM},[8] all achieve the same high-level success rate of 58.5% in this test, which details for how many proteins in the test set the known binding ligands are ranked in the correct order of their affinities. Regarding low-level success rates (i.e., the proportion of cases where the most affine ligand is on the top rank), even the first seven SFs, including all three investigated DrugScore variants, are comparable. Potential reasons for the very similar behavior of the SFs are the small test set size of only 65 protein/ligands combinations and/or the fact that the ligands to be distinguished per protein differ, on average, in their binding affinities by ~2.1 log units, which makes ranking possible also for less sensitive functions.

Docking Power Test. The result of the docking power test is shown in panel B of Figure 5. In 81.0% of all cases, DrugScore^{2018} scores a docking pose with an RMSD ≤ 2 Å as the best pose. Only one reference SF (ChemPLP[19] as implemented in GOLD) achieves similarly good results on the given test set. The best ranked pose by DrugScore^{CSD} and DrugScore lies in 79.5% and 71.8% of all cases within 2 Å RMSD of the correct pose, respectively. The best three SFs (two of which are DrugScore derivatives) in this test just differ slightly from each other. When the top two poses are considered, DrugScore^{2018}, DrugScore^{CSD}, and ChemPLP achieve success rates of 86.2%, 87.2% and 86.7%, respectively. When the best three scored poses are taken into consideration, the respective success rates increase to 88.2%, 89.2% and 89.7%.

Screening Power Test. In the Screening Power test, the top three performing SFs with regard to the enrichment factor are the commercial GlideScore (single (SP) and extra (XP) precision, with the single precision method outperforming the extra precision method) and ChemScore. GlideScore-SP achieves an $EF_{1\%}$ of 19.54% (Table 2). The performances of the DrugScore variants differ strongly in this test. DrugScore^{CSD} ranks at the sixth position with an $EF_{1\%}$ of 12.69% as the best DrugScore version. DrugScore^{2018} achieves an $EF_{1\%}$ of 6.92% and DrugScore one of 4.36%. Interestingly, SFs with high scoring, ranking, and docking power, such as X-Score^{HM} or DrugScore^{2018}, perform less good in screening trials and *vice versa* (e.g., Glide-SP/XP, LigScore2). This might be the reason why Ashtawy and Mahapatra designed task-specific SFs in their approach.[74] Furthermore, Glide-SP/XP has been trained on crystal ligands as well as decoys docked into a receptor.[5] Thus, the setup of the training (separation of ligand and decoys) resembles the setup of the application in the Screening Power test. This might explain the reduced false positive rate of these SFs in this test compared to DrugScore^{CSD} and DrugScore^{2018}.

Docking Into New DrugScore Fields. In addition to the CASF-2013 set, we evaluated DrugScore^{2018} as objective function in a large-scale docking trial using 4,056 protein-ligand complexes from PDBbind 2016[46, 47], a curated and highly diverse dataset. The overall docking performance is evaluated by the RMSD of the best scored pose with regard to the native pose, where an RMSD of ≤2 Å is considered to be a success. In Figure 6A, the cumulative frequency of the RMSD of the best scored docking pose for each docking run is plotted. Using the crystallographic ligand pose as input, AutoDock3 in combination with either DrugScore^{2018} or the built-in SF is able to reproduce this pose to within 2 Å RMSD in >75% of all dockings, while AutoDock4 surprisingly only achieves about 59.8% success rate. An overall better or at least very similar performance of AutoDock3 compared to AutoDock4 has been observed by

us[32] and others.[64] AutoDock3 in combination with DrugScore still has a 54.3% success rate. When the docking is performed using low-energy conformers as input, the success rate of AutoDock3 with DrugScore^{2018} or the built-in SF drops to 62.5% or 58.8%, respectively.

The observation of the drop in performance when using an input conformer that differs from the native pose is congruent with other studies.[81-84] One of those studies describes a similar assessment of ten SFs by docking as we did here for the DrugScore^{2018} potentials. The authors used the older and smaller PDBbind 2014 data set, which was further filtered by removing structures with non-standard residues, cofactors, and ions.[81] We did not apply such a filter on the larger PDBbind 2016 data set, which is why we consider our test set more challenging. Still, the similarity of the test sets and the generation of ligand conformers should allow a comparison of published docking performances with the performance of the DrugScore variants described here. According to Wang *et al.*, none of the freely available academic and commercially available docking tools was able to achieve a success rate of more than 60% in their test when the best scored pose is considered.[81] LeDock[85] performed best (57.4%) in the field of academic software, followed by rDock[86] (50.3%) and AutoDock Vina[87] (49.0%). The best commercially available docking software in their test was GOLD[7] (59.8%), followed by Glide-XP[88] (57.8%) and Glide-SP[5, 6] (53.8%). Thus, all of these tools performed worse than AutoDock3 with DrugScore^{2018} even on the putatively less challenging data set.

Besides the ability to reproduce crystallographic ligand poses, we also observed that AutoDock3 in combination with DrugScore^{2018} tends to generate less diverse docking solutions. In fact, more than 60% of all generated poses fell within the first cluster of the respective docking run, which is higher than AutoDock3 using different objective functions and substantially higher than AutoDock4 (Figure 6A, inlay). One example for a better and more converged docking is shown in Figure 6B**.** The docking of the PDB ID 3up2 complex using DrugScore^{2018} potentials (red) led to converged (all docking solutions were clustered into one cluster, i.e., all docking poses were within 2 Å RMSD to the cluster representative) docking solutions with an RMSD comparable to the native pose. This is a substantial improvement to docking with the initial DrugScore potentials (blue) and with standard AutoDock3 (yellow).

Altogether, in our setup, the combination of AutoDock3 and DrugScore^{2018} as objective function showed the best results with regard to the quality and convergence of generated docking solutions. The higher convergence in docking solutions indicates a smoother energy landscape of DrugScore^{2018}, which likely facilitates picking the right solution as small structural deviations are accommodated. Moreover, this combination of docking engine and objective

function has a higher success rate on a presumably more challenging dataset than any software tested in ref. [81].

5 Concluding remarks

We presented a new version of the knowledge-based SF DrugScore, DrugScore^{2018}, which builds upon the same formalism used to derive the original DrugScore almost 20 years ago, but exploits a training data set of X-ray complex structures almost 40 times larger than before. This training data set is highly diverse with respect to protein classes and ligands, and represents the relevant chemical space of small-molecule drug-like ligands. As a consequence, ~2.5 times as many pair potentials than before have a data basis of ≥500 pair interactions required to yield smooth potentials, and pair potentials could now be derived for eight more atom types, including interactions to halogen atoms and metal ions that are highly relevant for medicinal chemistry. Considering a normalized RMSD as a measure, we demonstrated that frequent protein-ligand interactions require 5,000 complex structures for derivation to become converged, infrequent ones 20,000. Furthermore, our results indicate that a derivation of pair-potentials from higher-quality complexes will most likely not result in a qualitative change of potential values or an improvement in their precision. Hence, we conclude that the DrugScore^{2018} potentials are converged, and further improvement would require modifications to the formalism, including, e.g., modified atom type definitions,[39] considering three-body interactions,[89] or adding terms accounting for not yet considered contributions to binding,[33] rather than additional structural data for derivation. Surprisingly, two other re-derived knowledge-based SFs, PMF04[14] and SMoG2016[43], did not find merit in using that many complex structures.

When evaluating DrugScore^{2018} in comprehensive tests on the established CASF-2013 dataset, the following results stood out: I) As to scoring, DrugScore^{2018} performs significantly better than DrugScore and DrugScore^{CSD} and ranks among the top three SFs. The residuals of predicted *versus* experimental binding affinities suggest that DrugScore^{2018} may be used to distinguish potential binders from non-binders. II) As to ranking, DrugScore^{2018} is in the top group of four SFs that all achieve the same performance. III) As to docking, DrugScore^{2018} is in the top group formed by two SFs. IV) As to screening, it is found that SFs with high scoring, ranking, and docking power, such as DrugScore^{2018}, perform less good in this test and *vice versa*. Finally, in a large-scale docking trial using 4,056 protein-ligand complexes from the PDBbind 2016, DrugScore^{2018} used as an objective function reproduces ligand poses to within 2 Å RMSD in >75% (63%) of the cases for crystallographic (low-energy) ligand conformations.

In summary, DrugScore^{2018} showed similar or improved performance in all aspects when compared to either DrugScore or DrugScore^{CSD} and was, overall, the SF showing the most consistently good performance in scoring, ranking, and docking tests. In our view, this result is

all the more remarkable as the DrugScore^{2018} potentials were derived from crystallographic information only, without any further adaptation using binding affinity or docking decoy data. DrugScore^{2018}, which is freely available from the authors for academic purposes, should thus be a competitive scoring and objective function for structure-based ligand design purposes.

6 Acknowledgements

We gratefully acknowledge the computational support provided by the “Center for Information and Media Technology” (ZIM) at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) on the supercomputer JURECA at Jülich Supercomputing Centre (JSC) (user ID: HKF7). We are grateful to OpenEye for an academic license.

7 Associated Content

The Supporting Information is available free of charge on the ACS Publications website.

Supplementary results on the comparison of characteristic DrugScore potentials, the chemical space and PCA of MQNs, and the improved scoring for metal and halogen containing complexes are provided. Supplementary figures show a comparison of selected pair potentials from DrugScore, DrugScore^{CSD}, and DrugScore^{2018} (Figure S1), the variance explained by the first nine principle components from the PCA of the molecules of the ChEMBL23 database represented by 42 descriptors (MQNs) (Figure S2), the projection of the chemical spaces covered by the ligands used for deriving DrugScore^{2018} pair potentials and the ChEMBL23 database (Figure S3), and the correlation of binding scores predicted by DrugScore, DrugScore^{CSD}, and DrugScore^{2018} with experimental binding constants for different subsets of the CASF-2013 test set (Figure S4). A supplementary table shows PDB entries with halogen- or metal-ligand interactions (Table S1).

8 References

(1) Huang, S.-Y.; Grinter, S. Z.; Zou, X., Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, 12, 12899-12908.

(2) Schneider, G.; Böhm, H.-J., Virtual screening and fast automated docking methods. *Drug Discov. Today* **2002**, 7, 64-70.

(3) Walters, W. P.; Stahl, M. T.; Murcko, M. A., Virtual screening - an overview. *Drug Discov. Today* **1998**, 3, 160-178.

(4) Böhm H.-J.; Stahl, M. The Use of Scoring Functions in Drug Discovery Applications. In *Reviews in Computational Chemistry (eds K. B. Lipkowitz and D. B. Boyd)*; Wiley-VCH: Hoboken, NJ, USA, 2003, pp 41-87.

(5) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, 47, 1750-1759.

(6) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, 47, 1739-1749.

(7) 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.

(8) Wang, R. X.; Lai, L. H.; Wang, S. M., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, 16, 11-26.

(9) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M., LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graph. Model.* **2005**, 23, 395-407.

(10) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E., Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.* **1995**, 8, 677-691.

(11) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W., Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, 14, 731-751.

(12) Muegge, I.; Martin, Y. C., A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, 42, 791-804.

(13) Muegge, I., A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspect. Drug Discov. Des.* **2000**, 20, 99-114.

(14) Muegge, I., PMF Scoring Revisited. *J. Med. Chem.* **2006**, 49, 5895-5902.

(15) Jain, A. N., Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, 10, 427-440.

(16) Böhm, H.-J., The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, 8, 243-256.

(17) Bohm, H. J., Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, 12, 309-323.

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

(19) 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.

(20) Liu, J.; Wang, R., Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, 55, 475-482.

(21) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. In *Adv. Protein Chem.*; Academic Press: 2003; Vol. 66, pp 27-85.

(22) Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, 295, 337-356.

(23) Gohlke, H.; Hendlich, M.; Klebe, G., Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, 20, 115-144.

(24) 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.

(25) Gohlke, H.; Klebe, G., Statistical potentials and scoring functions applied to protein–ligand binding. *Curr. Opin. Struct. Biol.* **2001**, 11, 231-235.

(26) Ashtawy, H. M.; Mahapatra, N. R. A Comparative Assessment of Conventional and Machine-Learning-Based Scoring Functions in Predicting Binding Affinities of Protein-Ligand Complexes. In 2011 IEEE International Conference on Bioinformatics and Biomedicine, 12-15 Nov. 2011, 2011; 2011; pp 627-630.

(27) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., 3rd, Assessing Scoring Functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004**, 47, 3032-3047.

(28) Wang, R.; Lu, Y.; Fang, X.; Wang, S., An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein−Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 2114-2125.

(29) Sotriffer, C. A.; Gohlke, H.; Klebe, G., Docking into Knowledge-Based Potential Fields: A Comparative Evaluation of DrugScore. *J. Med. Chem.* **2002**, 45, 1967-1970.

(30) Marek, L.; Hamacher, A.; Hansen, F. K.; Kuna, K.; Gohlke, H.; Kassack, M. U.; Kurz, T., Histone Deacetylase (HDAC) Inhibitors with a Novel Connecting Unit Linker Region Reveal a Selectivity Profile for HDAC4 and HDAC5 with Improved Activity against Chemoresistant Cancer Cells. *J. Med. Chem.* **2013**, 56, 427-436.

(31) Diedrich, D.; Hamacher, A.; Gertzen, C. G.; Alves Avelar, L. A.; Reiss, G. J.; Kurz, T.; Gohlke, H.; Kassack, M. U.; Hansen, F. K., Rational design and diversity-oriented synthesis of peptoid-based selective HDAC6 inhibitors. *Chem. Commun. (Camb.)* **2016**, 52, 3219-3222.

(32) Krüger, D. M.; Jessen, G.; Gohlke, H., How Good Are State-of-the-Art Docking Tools in Predicting Ligand Binding Modes in Protein–Protein Interfaces? *J. Chem. Inf. Model.* **2012**, 52, 2807-2811.

(33) Ben-Shalom, I. Y.; Pfeiffer-Marek, S.; Baringhaus, K. H.; Gohlke, H., Efficient Approximation of Ligand Rotational and Translational Entropy Changes upon Binding for Use in MM-PBSA Calculations. *J. Chem. Inf. Model.* **2017**, 57, 170-189.

(34) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, 19, 1639-1662.

(35) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Cryst. B* **2016**, 72, 171-179.

(36) Velec, H. F.; Gohlke, H.; Klebe, G., DrugScoreCSD - 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.

(37) Pfeffer, P.; Gohlke, H., DrugScoreRNA - Knowledge-Based Scoring Function To Predict RNA-Ligand Interactions. *J. Chem. Inf. Model.* **2007**, 47, 1868-1876.

(38) Kruger, D. M.; Gohlke, H., DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. *Nucleic Acids Res.* **2010**, 38, W480-486.

(39) 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.

(40) Neudert, G.; Klebe, G., fconv: format conversion, manipulation and feature computation of molecular data. *Bioinformatics* **2011**, 27, 1021-1022.

(41) Berman, H., The Protein Data Bank: a historical perspective. *Acta Cryst. A* **2008**, 64, 88-95.

(42) DeWitte, R. S.; Shakhnovich, E. I., SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **1996**, 118, 11733-11744.

(43) Debroise, T.; Shakhnovich, E. I.; Chéron, N., A Hybrid Knowledge-Based and Empirical Scoring Function for Protein-Ligand Interaction: SMoG2016. *J. Chem. Inf. Model.* **2017**, 57, 584-593.

(44) 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.

(45) 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.

(46) 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.

(47) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R., PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, 31, 405-412.

(48) *OpenEye Scientific Software, Santa Fe, NM 87507.*

(49) Sippl, M. J., Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **1995**, 5, 229-235.

(50) Gohlke, H.; Klebe, G., DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. *J. Med. Chem.* **2002**, 45, 4153-4170.

(51) Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, 50, 742-754.

(52) Reymond, J.-L., The Chemical Space Project. *Acc. Chem. Res.* **2015**, 48, 722-730.

(53) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P., The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, 42, D1083-D1090.

(54) Awale, M.; van Deursen, R.; Reymond, J. L., MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of Drug Bank, ChEMBL, Pub Chem, GDB-11, and GDB-13. *J. Chem. Inf. Model.* **2013**, 53, 509-518.

(55) Ruddigkeit, L.; Blum, L. C.; Reymond, J. L., Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2013**, 53, 56-65.

(56) Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G. A.; Tate, J.; Bateman, A., The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **2016**, 44, D279-D285.

(57) Efron, B., Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* **1979**, 7, 1-26.

(58) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P., *Numerical Recipes in C: The art of scientific computing, 2nd ed.* Cambridge University Press: New York, 1992.

(59) Jones, E.; Oliphant, T.; Peterson, P.; other SciPy: Open source scientific tools for Python. http://www.scipy.org (05.05.2018),

(60) Jain, A. N., Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, 46, 499-511.

(61) Jain, A. N., Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, 21, 281-306.

(62) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T., Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, 50, 572-584.

(63) Hawkins, P. C. D.; Nicholls, A., Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, 52, 2919-2936.

(64) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J., AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, 30, 2785-2791.

(65) Singh, U. C.; Kollman, P. A., An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, 5, 129-145.

(66) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, 45, D945-D954.

(67) Tanchuk, V. Y.; Tanin, V. O.; Vovk, A. I.; Poda, G., A New, Improved Hybrid Scoring Function for Molecular Docking and Scoring Based on AutoDock and AutoDock Vina. *Chem. Biol. Drug Des.* **2016**, 87, 618-625.

(68) Gentry, C. L.; Egleton, R. D.; Gillespie, T.; Abbruscato, T. J.; Bechowski, H. B.; Hruby, V. J.; Davis, T. P., The effect of halogenation on blood–brain barrier permeability of a novel peptide drug. *Peptides* **1999**, 20, 1229-1238.

(69) Lu, Y. X.; Wang, Y.; Zhu, W. L., Nonbonding interactions of organic halogens in biological systems: implications for drug discovery and biomolecular design. *Phys. Chem. Chem. Phys.* **2010**, 12, 4543-4551.

(70) Hernandes, M. Z.; Cavalcanti, S. M. T.; Moreira, D. R. M.; de Azevedo, W. F.; Leite, A. C. L., Halogen Atoms in the Modern Medicinal Chemistry: Hints for the Drug Design. *Curr. Drug Targets* **2010**, 11, 303-314.

(71) Lu, Y.; Liu, Y.; Xu, Z.; Li, H.; Liu, H.; Zhu, W., Halogen bonding for rational drug design and new drug discovery. *Expert Opin. Drug Discov.* **2012**, 7, 375-383.

(72) Supuran, C. T., Carbonic anhydrase inhibitors. *Bioorg. Med. Chem. Lett.* **2010**, 20, 3467-3474.

(73) Steiger, J. H., Tests for Comparing Elements of a Correlation Matrix. *Psychol. Bull.* **1980**, 87, 245-251.

(74) Ashtawy, H. M.; Mahapatra, N. R., Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment. *J. Chem. Inf. Model.* **2018**, 58, 119-133.

(75) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A., The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **2012**, 55, 5165-5173.

(76) Stahl, M.; Rarey, M., Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, 44, 1035-1042.

(77) Lionta, E.; Spyrou, G.; K. Vassilatis, D.; Cournia, Z., Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, 14, 1923-1938.

(78) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, 3, 935-949.

(79) Shoichet, B. K., Virtual screening of chemical libraries. *Nature* **2004**, 432, 862-865.

(80) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I., Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discov.* **2003**, 2, 369-378.

(81) Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T., Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* **2016**, 18, 12964-12975.

(82) Feher, M.; Williams, C. I., Effect of Input Differences on the Results of Docking Calculations. *J. Chem. Inf. Model.* **2009**, 49, 1704-1714.

(83) Corbeil, C. R.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model.* **2009**, 49, 997-1009.

(84) Jain, A. N., Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput. Aided Mol. Des.* **2008**, 22, 201-212.

(85) Zhang, N.; Zhao, H., Enriching screening libraries with bioactive fragment space. *Bioorg. Med. Chem. Lett.* **2016**, 26, 3594-3597.

(86) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D., rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, 10, e1003571.

(87) 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.

(88) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, 49, 6177-6196.

(89) Andreani, J.; Faure, G.; Guerois, R., InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics* **2013**, 29, 1742-1749.

9 Figures

TOC Figure Workflow for deriving new DrugScore pair-potentials.

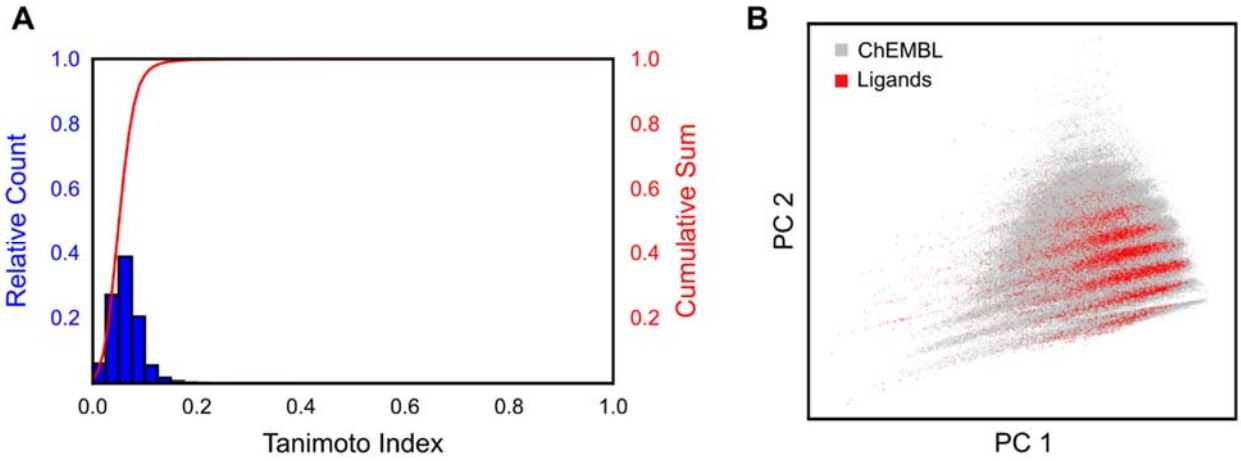

Figure 1. **(A)** Histogram of frequency (blue bars) and cumulative sum (red line) of Tanimoto indices for all pairs of ligands used for the new DrugScore^{2018} potentials. **(B)** The chemical space covered by the ligands (red) in comparison to the ChEMBL database (grey) as described by a PCA of 42 1D-descriptors (MQNs). PC1 and PC2 cover 16.1% and 13.4% of the variance, respectively.

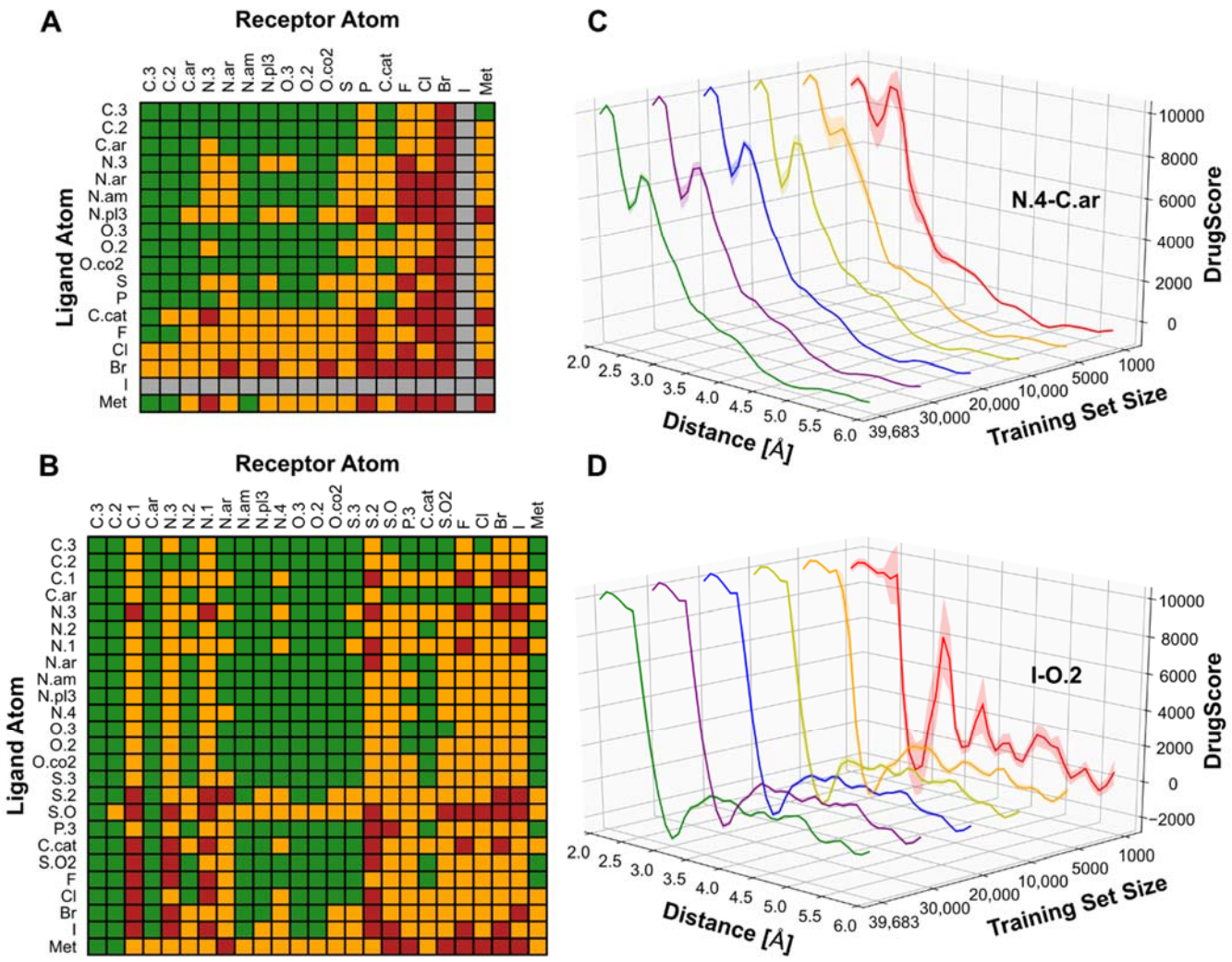

Figure 2. Number of pair interactions of the respective atom types of ligand and receptor found in the training sets for DrugScore (**A**) and DrugScore^{2018} (**B**). The colors depict if >500 (green), <500 (yellow), or no interactions (red) are found. Interactions not evaluated in DrugScore potentials are marked by grey rows/columns. In panel (**A**), the atom types S.2, S.O, and S.O2 were merged together with S.3, as was N.2 with N.ar. The influence of the number of structures used to derive DrugScore^{2018} potentials on the shape and error of the potential is shown exemplary for N.4-C.ar (**C**) and I-O.2 (**D**) interactions.

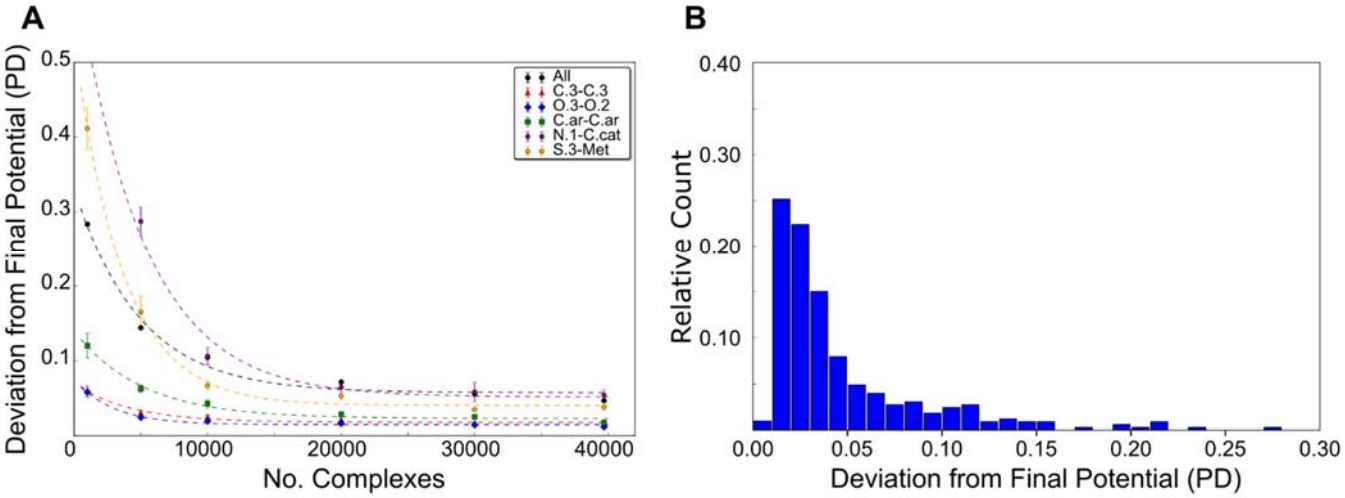

Figure 3. **(A)** Convergence of the distance-dependent DrugScore[2018] pair potentials with respect to the number of complexes in the training set. The mean PD of the pair potentials with corresponding error bars representing the standard error of the mean (SEM) was derived from bootstrapping experiments using six subsets of the training set. **(B)** Distribution of mean PD values obtained for the bootstrapping of the complete data set.

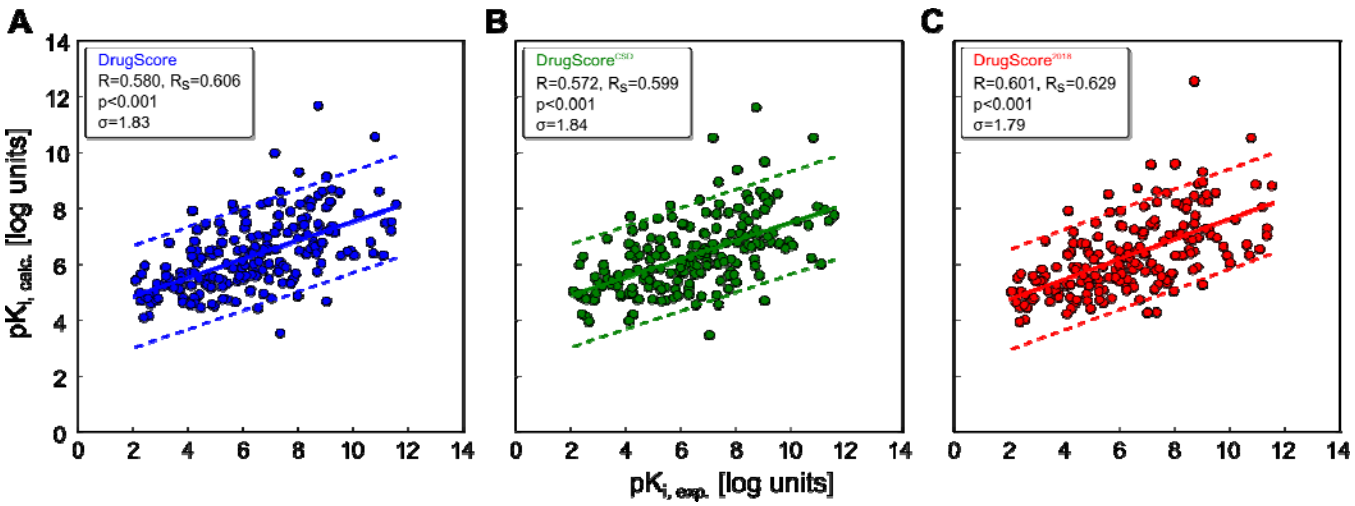

Figure 4. Comparison between predicted binding affinities using DrugScore (**A**), DrugScore^{CSD} (**B**), and DrugScore^{2018} (**C**) and those from experiments for the CASF-2013 dataset. p$K_{i,\,calc.}$ values were obtained by linear regression of DrugScore values to experimental data. For DrugScore^{2018}, values of 3.0 for the intercept and $-1.4*10^{-2}$ for the slope were found (in the case of a fixed intercept of 0, the slope (scaling coefficient) is $-2.5*10^{-2}$). These slope values are close to the one found previously for DrugScore.[50] The dashed lines indicate the residual range (± standard deviation **σ**) of the regression line. The Pearson correlation coefficient (*R*), the Spearman rank correlation (*Rs*), and *p* are given in the insets.

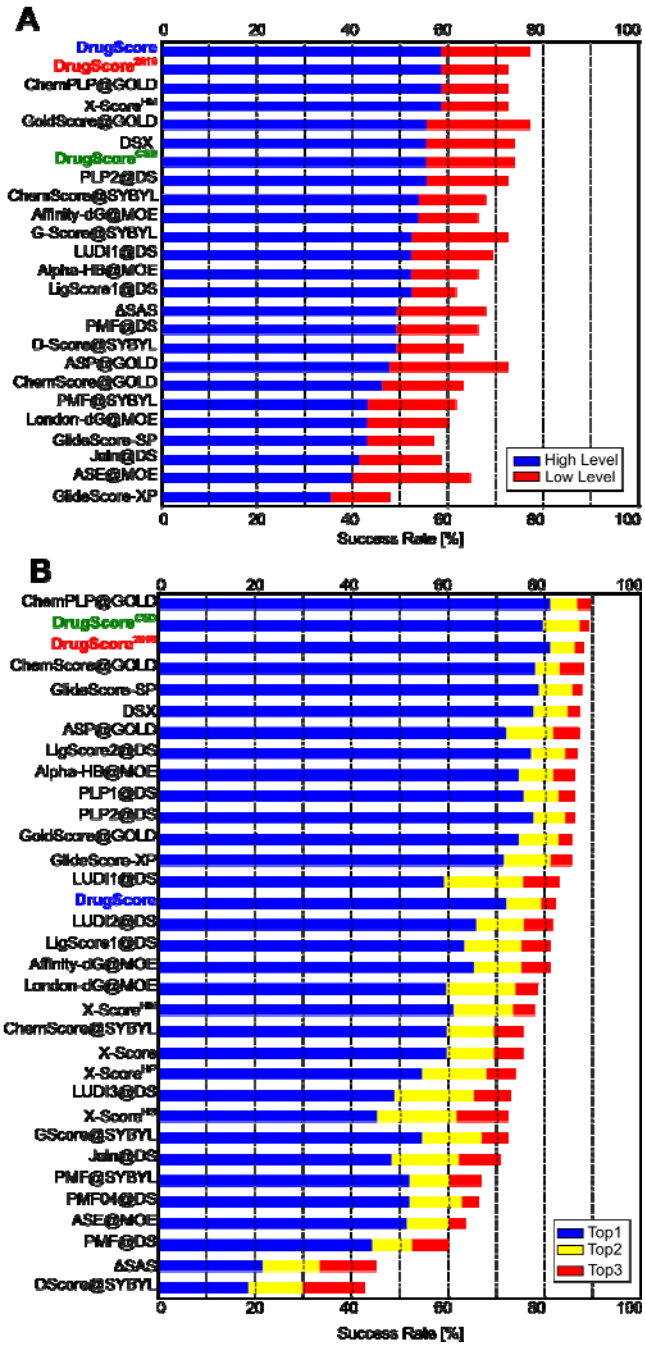

Figure 5. Bar plots of the results of the *Ranking Power* test (**A**) and *Docking Power* test (**B**) for the CASF-2013 dataset. The success rate of the *Ranking Power* test is defined by the SF's ability to score either all known binders in the correct order with respect to their known binding affinities ("High Level", blue), or to identify the best binder ("Low Level", red). The success rate of the *Docking Power* test is defined by the ability to best score a docking solution compared to the crystal ligand. The results are divided into three categories: A solution with RMSD ≤2 Å to the crystal ligand is ranked best ("Top1", blue), is within the first two ranked solutions ("Top2",

yellow), or within the first three ranked solutions ("Top3", red). Except for the DrugScore variants, values were taken from ref. [45].

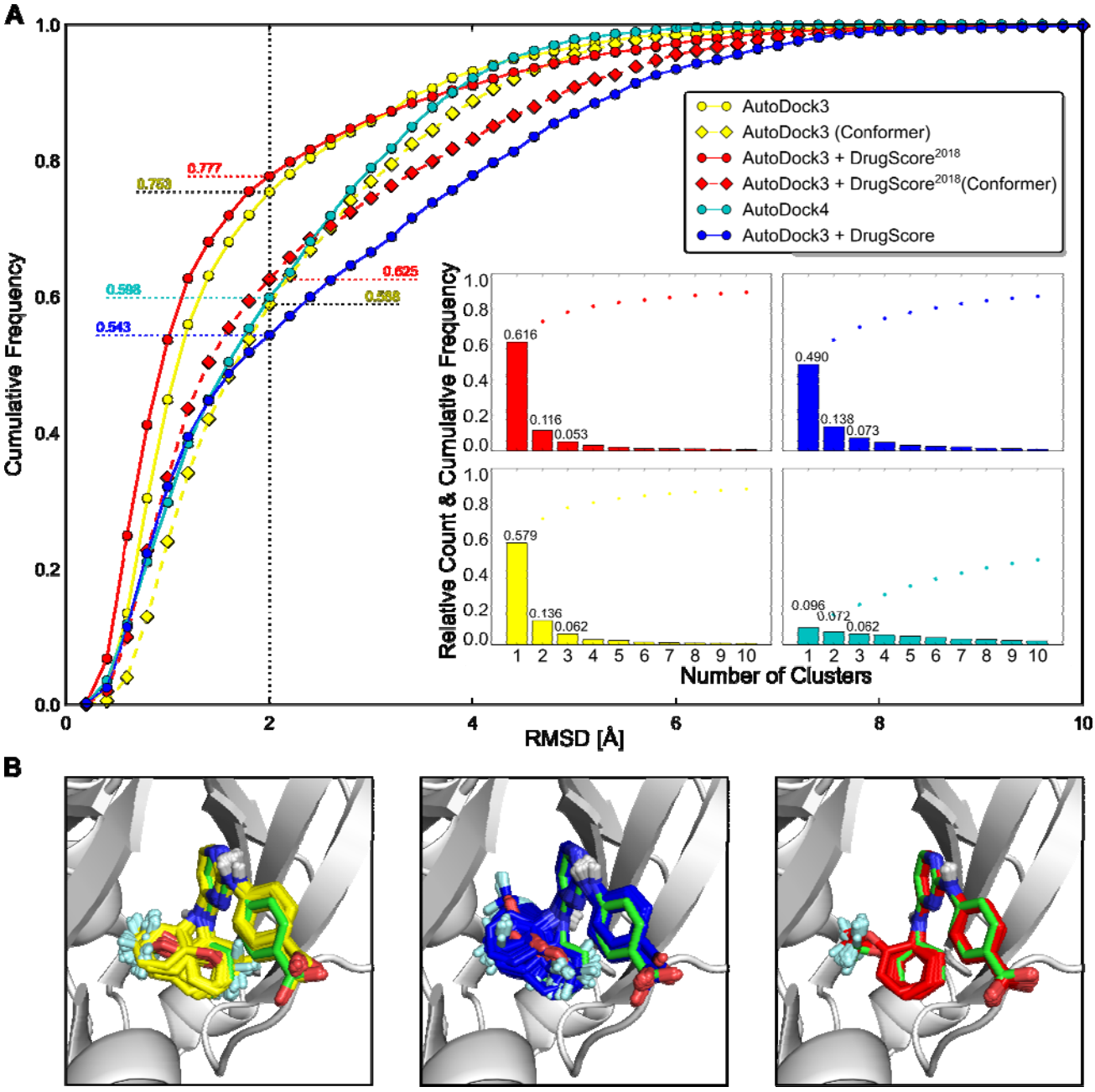

Figure 6. **(A)** Cumulative frequency of RMSD values of the best ranked docking pose against the crystal pose using either the ligand conformation of the crystal structure (dot marker, solid line) or a low-energy conformer(diamond marker, dashed line) as input. 4,056 protein-ligand complexes taken from the PDBbind Refined Set were used for docking. The vertical dotted line indicates an RMSD of 2.0 Å, which is considered a "good" docking solution. The inlay figures show the relative and cumulative frequency of the number of clusters of generated docking solutions in the docking runs. **(B)** Examples of docking convergence, comparing the built-in SF of AutoDock3, DrugScore, and DrugScore^{2018} (from left to right), for the protein-ligand complex PDB ID 3up2. The initial conformation of the crystal ligand is shown in green and the docking solutions in yellow (AD3), blue (DrugScore), and

red (DrugScore^{2018}), respectively. The chosen docking is exemplary for the enhanced prediction accuracy , as only DrugScore^{2018} predicts the orientation of the trifluoromethoxy group correctly.

10 Tables

Table 1. Results of the CASF-2013 *Scoring Power* test, sorted by Spearman's rank correlation.

Scoring function[a]	**# Scored complexes**	**R**[b]	**R_S**[c]	**SD**[d]
BT-Score[74] [e]	195	0.825	n.a.[e]	n.a.
AutoDockHybrid[67]	195	0.635	0.638	1.76
DrugScore^{2018}	**195**	**0.601**	**0.629**	**1.79**
X-Score^{HM}	195	0.614	0.626	1.78
ΔSAS	195	0.606	0.624	1.79
ChemPLP@GOLD	195	0.579	0.614	1.84
DSX	**195**	**0.584**	**0.610**	**1.82**
ChemScore@SYBYL	195	0.592	0.610	1.82
DrugScore	**195**	**0.580**	**0.606**	**1.83**
DrugScore^{CSD}	**195**	**0.572**	**0.599**	**1.84**
SMoG2016[43] [e]	195	0.570	n.a.[e]	1.68
PLP1@DiscoveryStudio	195	0.568	0.586	1.86
ASP@GOLD	195	0.556	0.578	1.88
PLP2@DiscoveryStudio	195	0.558	0.571	1.87
GScore@SYBYL	195	0.558	0.557	1.87
DScore@SYBYL	195	0.526	0.556	1.92
ChemScore@GOLD	189	0.536	0.544	1.90
Alpha-HB@MOE	195	0.511	0.526	1.94
ASE@MOE	195	0.544	0.522	1.89
GoldScore@GOLD	189	0.483	0.498	1.97
LUDI2@DiscoveryStudio	195	0.451	0.494	2.01
LigScore2@DiscoveryStudio	190	0.456	0.493	2.02
Affinity-dG@MOE	195	0.482	0.489	1.98
LUDI3@DiscoveryStudio	195	0.487	0.488	1.97
LUDI1@DiscoveryStudio	195	0.444	0.474	2.02
Jain@DiscoveryStudio	191	0.408	0.445	2.05
GlideScore-SP	169	0.452	0.402	2.03
PMF@DiscoveryStudio	194	0.364	0.364	2.11
PMF@SYBYL	191	0.221	0.364	2.20
LigScore1@DiscoveryStudio	192	0.348	0.345	2.13
GlideScore-XP	164	0.277	0.308	2.18
London-dG@MOE	195	0.242	0.277	2.19
PMF04@DiscoveryStudio	188	0.200	0.244	2.22

[a] All data was taken from ref. [45] except for BT-Score[74], AutoDockHybrid[67], SMoG2016[43], and the three DrugScore variants.
[b] Pearson correlation coefficient.
[c] Spearman correlation coefficient.
[d] Standard deviation of predicted pK_i values; in log units.
[e] Data not available.

Table 2. Results of the CASF-2013 *Screening Power* test.

Scoring function[a]	Enrichment Factor (EF)[b]		
	Top 1%	Top 5%	Top 10%
GlideScore-SP	19.54	6.27	4.14
ChemScore@GOLD	18.9	6.83	4.08
GlideScore-XP	16.81	6.02	4.07
LigScore2@DiscoveryStudio	15.9	6.23	3.51
ChemPLP@GOLD	14.28	5.88	4.31
DrugScore^{CSD}	**12.69**	**4.86**	**3.34**
LUDI1@DiscoveryStudio	12.53	4.28	2.8
ASP@GOLD	12.36	6.23	3.79
DSX	**8.46**	**4.05**	**2.88**
Affinity-dG@MOE	8.21	4.15	3.19
London-dG@MOE	8.08	3.36	2.51
GoldScore@GOLD	7.95	4.52	3.16
PLP1@DiscoveryStudio	6.92	4.28	3.04
DrugScore^{2018}	**6.92**	**3.44**	**2.57**
Jain@DiscoveryStudio	5.9	2.51	1.8
PMF@SYBYL	5.38	2.21	1.9
ChemScore@SYBYL	5.26	2.38	2.18
Alpha-HB@MOE	4.87	3.23	1.32
PMF04@DiscoveryStudio	4.87	2.87	2.63
ASE@MOE	4.36	2.35	1.59
DrugScore	**4.36**	**1.82**	**2.03**
X-ScoreHM	2.31	2.14	1.41
D-Score@SYBYL	2.31	1.79	1.46
G-Score@SYBYL	1.92	1.26	1.44
ΔSAS	1.41	1.28	1.12

[a] All data was taken from ref. [45] except for the three DrugScore variants.
[b] The results are divided into three categories: the enrichment factor (EF) is calculated for the top 1%, 5% and 10% of the data set.