JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

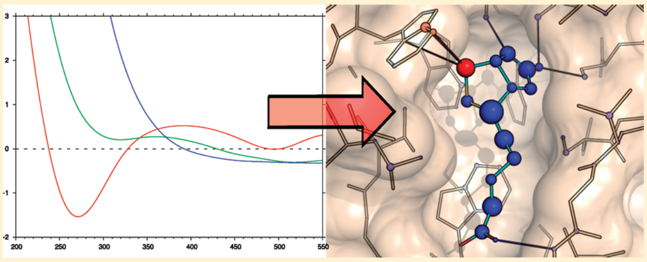
pubs.acs.org/jcim

DSX: A Knowledge-Based Scoring Function for the Assessment of Protein−Ligand Complexes

Gerd Neudert and Gerhard Klebe*

Department of Pharmaceutical Chemistry, Philipps-Universität Marburg, Marbacher Weg 6, D-35032, Germany

Supporting Information



ABSTRACT: We introduce the new knowledge-based scoring function *DSX* that consists of distance-dependent pair potentials, novel torsion angle potentials, and newly defined solvent accessible surface-dependent potentials. *DSX* pair potentials are based on the statistical formalism of DrugScore, extended by a much more specialized set of atom types. The original DrugScore-like reference state is rather unstable with respect to modifications in the used atom types. Therefore, an important method to overcome this problem and to allow for robust results when deriving pair potentials for arbitrary sets of atom types is presented. A validation based on a carefully prepared test set is shown, enabling direct comparison to the majority of other popular scoring functions. Here, *DSX* features superior performance with respect to docking- and ranking power and runtime requirements. Furthermore, the beneficial combination with torsion angle-dependent and desolvation-dependent potentials is demonstrated. *DSX* is robust, flexible, and capable of working together with special features of popular docking engines, e.g., flexible protein residues in AutoDock or GOLD. The program is freely available to the scientific community and can be downloaded from our Web site www.agklebe.de.

INTRODUCTION

Supported by an increasing number of deposited crystal structures of relevant drug targets, structure-based virtual screening has become an important aspect in modern drug research. Molecular docking is used to generate reasonable geometries of protein−ligand complexes from huge compound libraries. Popular docking programs, such as AutoDock,[1−3] DOCK,[4] eHiTS,[5] FlexX,[6] Glide,[7,8] GOLD[9,10] and Surflex,[11] use different approaches to solve the ligand placement problem,[12] and all methods are able to generate near-native binding geometries.[13−16] However, it has also been shown that the position of the pose closest to the experimental pose is distributed rather randomly among all generated poses, when they are ordered with respect to the docking score.[17] Hence, a reliable evaluation of computed protein−ligand complexes is still one of the most challenging problems in a virtual screening scenario.

We will distinguish between three different tasks a scoring function should accomplish: (i) If a distinct native binding mode for a compound exists, the function should identify the pose closest to the native conformation among a huge number of generated poses for this compound. (ii) If a set of different ligands binding to the same protein is given, a reliable scoring function must be able to rank the ligands according to their binding affinities. (iii) If a series of arbitrary protein−ligand complexes is given, the linear correlation between predicted scores and binding affinities should be as high as possible.

As proposed by Cheng et al.,[18] we will refer to the first criterion as “docking power”, the second as “ranking power”, and the third as “scoring power” (or synonymously “affinity prediction”). A scoring function that is perfect with respect to scoring power would also be perfect with respect to ranking power and superior in docking power. However, considering the level of simplification of the underlying biophysics, current scoring functions are far-off from being perfect in affinity prediction. Especially protein flexibility, desolvation effects, and entropic considerations involving torsional, translational, and rotational degrees of freedom are not sufficiently accounted for. Computationally expensive methods like free energy perturbation,[19] thermodynamic integration,[20] or MM-PBSA calculations[21] can yield accurate results but are not applicable in a high-throughput virtual screening campaign, as sufficient conformational sampling would be computationally too demanding. Nevertheless, scoring power has no direct relevance for rescoring. Instead, relative ranking of a sample of candidate ligands with respect to their affinity to one given target is required and not the absolute affinity (which can be determined experimentally for the most promising candidates). Most likely, a near-native pose is a prerequisite to yield correct ranking. In consequence, high ranking power is rather useless without high docking power. Therefore, a scoring function should be adequate for both, docking and ranking, or at least the task can be split into a combination of two functions, each tailored for one goal. One could argue that sufficient scoring power is required to predict

Received: June 16, 2011
Published: August 24, 2011

cross reactivities, but even in this case ranking with respect to a reference compound could be satisfying.

With dependence on the methodological background, scoring functions are often classified into three categories: (i) Force-field-based scoring functions[4,9,10] use classical molecular mechanical force fields to evaluate binding energy. (ii) Empirical scoring functions[3,7,22−30] decompose the total energy into several linear energy terms. The weighting of the individual terms is done by regression analysis using a training set with experimental binding affinities. (iii) Knowledge-based scoring functions[31−42] calculate the total score as sum of statistical potentials, which are derived from a database of known protein−ligand complexes.

This classification is rather crude and some scoring functions are difficult to assign to one of the three categories. For example, MotifScore[43] does not apply the usual decomposition into individual atom−atom terms but instead scores complete three-dimensional motifs.

Because they are trained by affinities, the key skills of empirical scoring functions should be ranking power and scoring power. As a shortcoming, their predictive power strongly depends on the similarity between important interactions in the complex under evaluation and important interactions in the training set complexes. Furthermore, they suffer from both, uncertainties in the structural data and experimental errors for the affinity data of the training set.

In contrast, knowledge-based functions do not rely on affinity data but exploit comprehensive crystallographic information. Thus, they are more general and their key skill should be docking power because the statistical potentials reflect native binding geometries. When only distance-dependent atom−atom potentials are used, they are also faster to compute than empirical functions. This is important, as docking power needs many more function evaluations compared to ranking power.

In this study, we present a new knowledge-based scoring function named *DSX* (DrugScore eXtended), whose pair potentials are based on the DrugScore formalism.[35,36] We extended this approach with respect to a more detailed atom type assignment and a modification to overcome a problem with the reference state. We also included statistically derived torsion-angle potentials, which allow for fast relaxation of docking poses and can improve docking power and ranking power. In addition, a new type of solvent-accessible-surface-dependent potential is introduced and a validation of *DSX* is presented based on the carefully prepared and publicly available data set of Cheng et al.[18]

The next section supplies the theoretical background for subsequent discussions about reference states, volume corrections, and the newly defined statistical potentials. It also clarifies inconsistencies in terminology and foundation of the formalisms that are found in the literature. Furthermore, differences between the most popular knowledge-based functions, namely, PMF,[41,42] ASP,[38] and DrugScore,[35,36] will be presented along with the implemented modifications and extensions in *DSX*.

THEORY

An idea that originally lead to knowledge-based potentials is based on the Boltzmann distribution

$$\frac{n(i)}{N} = \rho(i) = \frac{\mathrm{e}^{-E(i)/kT}}{Z(T)} \qquad (1)$$

where $n(i)$ is the number of particles in a set of states i with the energy $E(i)$, N the total number of particles in the system, T the absolute temperature, k the Boltzmann constant, and $Z(T)$ the partition function (or Boltzmann sum over states). The fraction $\rho(i)$ is a state-dependent density function, which is also a probability function. Equation 1 is the distribution function for the canonical ensemble, hence for a system in thermodynamic equilibrium with fixed temperature, volume, and number of particles. Rearrangement leads to an equation which is often referred to as the inverse Boltzmann law.

$$E(i) = -kT\ln(\rho(i)) - kT\ln(Z(T)) \qquad (2)$$

If the partition function is unknown, one can still calculate energy differences compared to a reference state, because $Z(T)$ is constant at constant temperature.

$$\begin{aligned}\Delta E = E(i) - E_{\mathrm{ref}} &= -kT\ln(\rho(i)) + kT\ln(\rho_{\mathrm{ref}})\\ &= -kT\ln\left(\frac{\rho(i)}{\rho_{\mathrm{ref}}}\right)\end{aligned} \qquad (3)$$

In the theory of liquids,[44] free energies are calculated using radial distribution functions $g(r)$ corresponding to the fraction $(\rho(i))/(\rho_{\mathrm{ref}})$. The Helmholtz free energy $W_{\mathrm{ab}}(r)$ of two particles a and b in a homogeneous solvent is

$$\begin{aligned}W_{\mathrm{ab}}(r) &= -kT\ln(g(r)) = U_{\mathrm{ab}}(r) + \delta G_{\mathrm{ab}}\\ g(r) &= \frac{\rho_{\mathrm{ab}}(r)}{\rho_{\mathrm{ref}}(r)} = \frac{P_{\mathrm{ab}}(r)}{P_{\mathrm{ref}}(r)}\end{aligned} \qquad (4)$$

which is the reversible work spent or gained when transferring a and b from infinite separation to a distance r. In this case, the reference state is the ideal gas, thus $P_{\mathrm{ab}}(r)$ corresponds to the probability to find two particles in liquid at a distance r, while $P_{\mathrm{ref}}(r)$ corresponds to the probability to find them at the same distance in an ideal gas. Because $W_{\mathrm{ab}}(r)$ corresponds to the mean force acting on the two particles due to their interactions with the surrounding δG_{ab} and with each other $U_{\mathrm{ab}}(r)$, it is called a potential of mean force.

In analogy to eq 4, attempts were made to use potentials of mean force for protein folding prediction[45−48] and for scoring of protein−ligand complexes.[41] Here, the contact densities are calculated from the contact data found in the protein data bases like the PDB. However, it has been clearly pointed out that the derived statistical potentials are no potentials of mean force.[49,50] In essence, the radial distribution functions for the protein systems are derived for particles taken from different environments. That is, a and b have different interactions with their surrounding in different protein−ligand complexes. Thus, the δG_{ab} is different for each contact ab, and averaging this data cannot yield a density function that corresponds to the $g(r)$ used in eq 4. Furthermore, the $U_{\mathrm{ab}}(r)$ in eq 4 are additive but the δG_{ab} are not.[49] In consequence, a partition of the total free energy into pairwise atom−atom contributions is not valid. With the argument on the basis of eqs 2 and 3, the problem simply is that the distribution of atom−atom contact distances does not really follow the Boltzmann distribution and therefore they cannot be used to calculate energies based on this statistic. The reason is related to the problem of different environments. Two atoms a and b in a protein are not necessarily found at thermodynamic equilibrium distances even though the complete system might be at equilibrium because the intramolecular structure of both, protein and ligand, prevents a Boltzmann like distribution.

Taking all this into account, we should strictly avoid terms like "potential of mean force" or "energy" when we talk about statistical potentials. Koppensteiner and Sippl[50] even proposed to avoid the term "potential", instead "preference" or "quantity" should be used. As the term "potential" is rather popular and not necessarily linked to an energy function, we will also stick to this term in the following.

Given that the values computed by statistical potentials are not energies, we can drop the linear factor kT and replace the term "energy" by "score", which leads to the master equation for knowledge-based scoring functions.

$$\text{score}(i) = -\ln\left(\frac{\rho(i)}{\rho_{\text{ref}}}\right) \tag{5}$$

In the case of pairwise distance-dependent contributions, the total score for a given complex of protein atoms a_{p} and ligand atoms a_{l} is calculated as

$$\text{total score}_{\text{pair}} = \sum_{a_{\text{p}}}\sum_{a_{\text{l}}} \text{score}(p(a_{\text{p}}), l(a_{\text{l}}), r(a_{\text{p}}, a_{\text{l}})) \tag{6}$$

$$\text{score}_{\text{pair}}(p, l, r) = -\ln\left(\frac{\rho(p, l, r)}{\rho_{\text{ref}}}\right) \tag{7}$$

where $p(a_{\text{p}})$ and $l(a_{\text{l}})$ are the atom types and $r(a_{\text{p}},a_{\text{l}})$ is the distance of a_{p} and a_{l}. Equation 5 is not necessarily restricted to distance dependent atom—atom scores but can also be applied to many other structural features like bond or dihedral angles. Using Bayesian probability theory, one can obtain similar equations,[51] but the problem of deriving meaningful probability functions remains and the prerequisite of pairwise independence is not fulfilled. Finally, we should accept statistical potentials as a class of heuristics and only the experiment can tell us how meaningful they are.

Distance-Dependent Pair Potentials. Besides the choice of appropriate atom types and an appropriate data sample, the choice of a proper reference state is crucial for the quality of statistical potentials. In case of PMF[41,42] or ASP,[38] it has been selected as state of no interaction, referring to the analogy to potentials of mean force. In contrast, the reference state used in DrugScore is chosen as state of mean interaction. In principle, ρ_{ref} can be seen as a kind of weighting function for $\rho(i)$ to successfully apply eq 5. Another aspect often discussed is the volume correction for atom types i to account for the *de facto* available volume.

PMF, ASP, and DrugScore are the most popular knowledge-based scoring functions and have been evaluated on the test set used in this contribution. With respect to the changes in our new scoring function *DSX*, we will focus our comparison concerning reference state and volume correction to these functions. All are based on eq 7 but differ in the definition of the density functions.

In PMF, we have

$$\rho^{\text{PMF}}(p, l, r) = f(l, r)\frac{N(p, l, r)}{\Delta V(r)} \tag{8}$$

$$\rho_{\text{ref}}^{\text{PMF}} = \rho_{\text{ref}}^{\text{PMF}}(p, l) = \frac{\sum_{r}^{R} f(l, r)N(p, l, r)}{V(R)} \tag{9}$$

where $N(p,l,r)$ is computed from the database as the number of contacts between protein atom type p and ligand atom type l with a distance in the interval $[r,r + bin\_size[$. The contact numbers are normalized by the theoretically available volume $\Delta V(r)$ of the spherical shell corresponding to $[r,r + bin\_size[$. The factor $f(l,r)$ is a correction of the theoretically available volume due the space that is occupied by other ligand atoms (averaged from the database). R is the cutoff radius of 12 Å, and $V(R)$ is the volume of the corresponding sphere. Strictly speaking, in this case the density functions are not probability functions. However, an applied normalization will not change the value of the fraction. Here, the reference density is clearly dominated by long-range contacts and thus an approximation to a state of no specific interaction.

In ASP, we have

$$\rho^{\text{ASP}}(p, l, r) = \frac{N(p, l, r)}{\Delta V(r)f(l, r)f(p, r)} \tag{10}$$

$$\rho_{\text{ref}}^{\text{ASP}} = \rho_{\text{ref}}^{\text{ASP}}(p, l) = \left\langle \frac{N(p, l, r')}{\Delta V(r')f(l, r')f(p, r')} \right\rangle_{r'=6.0}^{r'=8.0} \tag{11}$$

where, in addition to a ligand volume correction, also a protein volume correction $f(p,r)$ is used. The angle brackets stand for the calculation of a mean value over all bins from 6 to 8 Å. As in PMF, the reference is chosen as a state of no specific interaction and the density functions are not probability functions. The cutoff distance used for scoring is 6 Å.

In DrugScore, we have

$$\rho^{\text{DS}}(p, l, r) = \frac{N(p, l, r)}{\Delta V(r)\sum_{r'} N(p, l, r')/\Delta V(r')} \tag{12}$$

$$\rho_{\text{ref}}^{\text{DS}} = \rho_{\text{ref}}^{\text{DS}}(r) = \frac{\sum_{p'}\sum_{l'} \rho(p', l', r)}{n_{\text{p}}n_{\text{l}}} \tag{13}$$

where n_{p} is the number of different protein atom types and n_{l} is the number of different ligand atom types. Here, the reference is selected as a state of mean interaction and the density functions are also probability functions. The latter fact is important, because averaging over all density functions without normalization would result in a reference dominated by contact types with high occurrence frequencies. As for ASP, the cutoff distance is 6 Å.

It is not important whether the reference is chosen as a state of no interaction or a state of mean interaction. Its main task is to weight $\rho(p,l,r)$ in the best achievable agreement with experimental evidence. In eqs 9 and 11, the reference depends on the contact type p_l; hence, it is constant for a given contact type. As a consequence, the weighting between two different contact types $p_1\_l_1$ and $p_1\_l_2$ is constant for all distances and the extrema in the potentials will always correspond to the extrema in the $\rho(p,l,r)$. In eq 13, the reference is solely a function of r. The weighting between different contact types is done by averaging over all possible contact types, but in contrast to PMF and ASP the weighting for short-range interactions of two given contact types may differ from the weighting of long-range interactions for the same types. As a result, the extrema of the DrugScore potentials can differ from the extrema of $\rho(p,l,r)$.

An advantage of the DrugScore reference state is the implicit inclusion of a volume correction. At short distances, we generally find fewer contacts than theoretically expected. This is due to the inaccessibility of space actually occupied by other ligand or

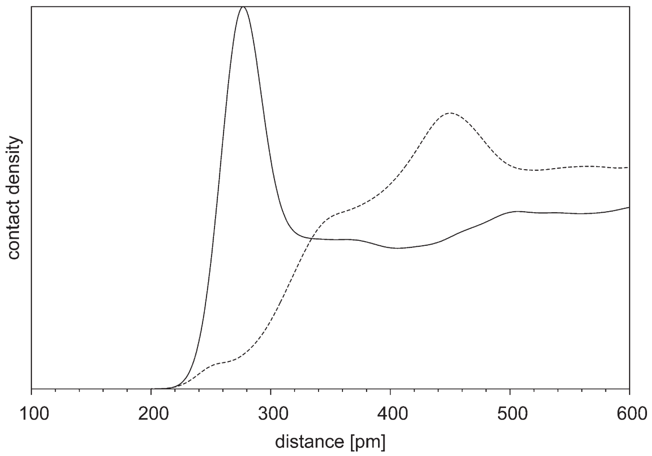


Figure 1. Density functions for two contact types processed from the CSD. O.3oh_O.carb (solid line) contact between the hydroxyl oxygen and carbonyl oxygen and O.3et_O.carb (dotted line) contacts between the oxygen in an aliphatic ether and a carbonyl oxygen.

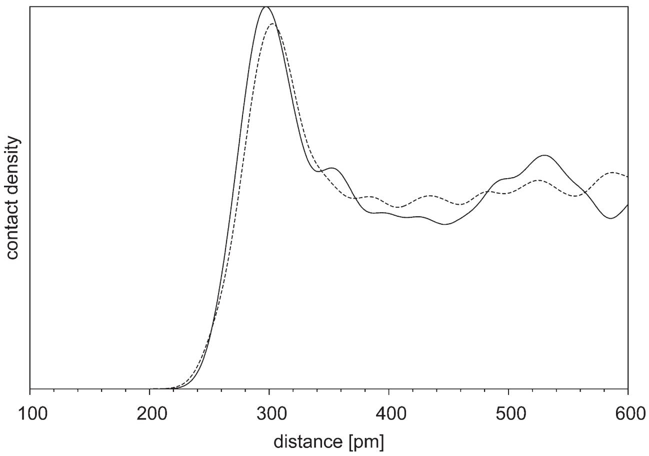


Figure 2. Density functions for two contact types processed from the CSD. N.3p_O.co2 (solid line) contact between a primary sp^3 nitrogen and oxygen in deprotonated carboxylates and N.3p_O.3oh (dotted line) contact to a hydroxyl oxygen.

protein atoms, but it also implies that the reference state obtains lower values at short distances. Thus, the ratio $(\rho(p,l,r))/(\rho_{\text{ref}}(r))$ does not change in the mean. This implicit correction is an average correction for all atom types and it is sufficient as long as the available volume for particular atom types is not significantly different from the averaged value. However, Mooij and Verdonk[38] demonstrated that there are considerable deviations from the mean value in the case of protein atoms. Thus, also for DrugScore-like pair potentials, an explicit volume correction seems to be necessary when deriving contact data from protein complexes and we will investigate its influence in the Results and Discussion section.

A putative disadvantage of the DrugScore reference state is the fact that any incorrect or erroneous density function will influence all resulting potentials, or more generally speaking, there is only one reference function that will affect and therefore determine the quality of all potentials. The latter observation becomes even more important due to another problem of eq 13 which we will discuss in the next section.

DSX Pair Potentials. *DSX* pair potentials are based on eqs 12 and 13, but in contrast to DrugScore, *DSX* does not apply Sybyl atom types but atom types defined by fconv.[52] The importance of the utilized atom-type set on the quality and reliability of statistical potentials has been shown in previous studies,[38,53] and the choice of appropriate types is not trivial. In case of the Sybyl types, one major concern regards the missing differentiation between oxygens with and without donor functionality (both O.3). In Figure 1, two fconv-type-based density functions derived from the Cambridge Structural Database (CSD)[54] are shown. With the use of Sybyl types, O.3oh_O.carb and O.3et_O.carb would be merged into one single density function O.3_O.2. With dependence on the occurrence frequencies of hydroxyl and ether oxygens (both assigned as O.3), information about the hydrogen-bond interaction would be lost. In other cases, it is not really obvious whether differences in contact densities have to be expected. As already mentioned, one problem with the analogy to potentials of mean force is that particles present in different environments should also be treated as particles of different types. Thus, the more atom types we differentiate with respect to their environment, the more this problem will be reduced. The degree of differentiation is mainly limited by the available contact information in the knowledge base. For *DSX*, we started with the 158 fconv atom types[52] and excluded all hydrogen atom types as well as unusual metal types. We also merged some atom types with low occurrence frequencies (see Methods). However, even if we assume that all possible contact types of the remaining atom types will be sufficiently represented in the database, an increasing differentiation will raise another problem with respect to the reference state as defined in eq 13. If an atom type p_1 is split into two new types p_{11} and p_{12}, the possible contact types $p_1\_l_x$ are considered twice as $p_{11}\_l_x$ and $p_{12}\_l_x$. This is desired in case that all contact density functions $p_{11}\_l_x$ are different from the corresponding functions $p_{12}\_l_x$, but as shown in Figure 2 it is also possible that two contact types are essentially equal. In that case ($p_1\_l_x = p_{11}\_l_x = p_{12}\_l_x$), the only effect of splitting up p_1 is to double the weight of $p_1\_l_x$ in the reference state. Theoretically, one could split p_1 into a large number of subtypes p_{1y}, but if these subtypes do not differ from p_1, the result would be a reference state that is equal to the average of the $p_1\_l_x$ and the information of other contact types $p_y\_l_x$ would be significantly downscaled.

Our strategy to reduce this problem is the clustering of the density functions by means of an appropriate similarity measure. In contrast to the merging of atom types, here it is possible to cluster two density functions $p_{11}\_l_1$ and $p_{12}\_l_1$ but to keep the differentiation between $p_{11}\_l_2$ and $p_{12}\_l_2$. The definition of the density functions for *DSX* results as

$$\rho^{DSX}(c,r) = \frac{\sum\limits_{p\_l \,\in\, c} N(p,l,r)}{\Delta V(r) \sum\limits_{p\_l \,\in\, c} \sum\limits_{r'} N(p,l,r')/\Delta V(r')} \qquad (14)$$

$$\rho_{\text{ref}}^{DSX} = \rho_{\text{ref}}^{DSX}(r) = \frac{\sum\limits_{c'} \rho(c',r)}{n_c} \qquad (15)$$

where c denotes an individual cluster of contact types and n_c is the number of clusters.

From a probabilistic point of view, eq 14 is an estimator for the conditional probability to find a contact at distance r, given the contact type c. The reference is an estimator for the averaged probability to find an arbitrary contact at distance r and the resulting potential is a log-likelihood function.

$$\rho^{DSX}(c,r) = P(r|c) \tag{16}$$

$$\rho^{DSX}_{\mathrm{ref}} = \frac{\sum_{c'} P(r|c')}{n_{\mathrm{c}}} = \bar{P}(r) \tag{17}$$

$$\mathrm{score}^{DSX}_{\mathrm{pair}} = -\ln\left(\frac{P(r|c)}{\bar{P}(r)}\right) \tag{18}$$

In principle, arbitrary likelihoods could serve as appropriate scoring measures, as long as the calculated density functions are good estimates for the corresponding probability functions. For example, equivalently to eq 18, one could define

$$\mathrm{score}_{\mathrm{pair}} = -\ln\left(\frac{P(c|r)}{\bar{P}(c)}\right) \tag{19}$$

$$\rho(c,r) = P(c|r) = \frac{N(c,r)}{F(c)\sum_{c'} N(c',r)/F(c')} \tag{20}$$

$$\rho_{\mathrm{ref}} = \rho_{\mathrm{ref}}(c) = \bar{P}(c) = \frac{\sum_{r'} \rho(c,r')}{n_{\mathrm{r}}} \tag{21}$$

where $\rho(c,r)$ is the conditional probability to find a specific contact type c, given the contact distance r, and n_{r} is the number of used distance bins. With this definition, a normalization with respect to the (corrected) theoretically available volume would not be necessary. Instead, a normalization with respect to the occurrence frequencies of contact types would be mandatory, because otherwise the highly populated types would dominate $\rho(c,r)$. One possible normalization factor could be $F(c) = N(c) = \sum_{r'} N(c,r')$, where $N(c)$ is the total number of contacts of type c found in the knowledge base.

DSX Torsion Potentials. To allow for a local relaxation of docking poses and to deal with unlikely torsion angles produced by docking programs, we developed knowledge-based torsion angle-dependent potentials for *DSX*. On the basis of eq 5, we defined the state i of a torsion as a function of the four atom types a, b, c, and d being part of the torsion, a qualifier e, and the actual torsion angle ϕ.

$$\begin{aligned} \mathrm{score}^{DSX}_{\mathrm{tors}}(t,\phi) &= -\ln\left(\frac{\rho(t,\phi)}{\rho_{\mathrm{ref}}}\right) = -\ln\left(\frac{P(\phi|t)}{\bar{P}(\phi)}\right) \\ \rho(t,\phi) &= \frac{N(t,\phi)}{\sum_{\phi'} N(t,\phi')} \\ \rho_{\mathrm{ref}} &= \rho_{\mathrm{ref}}(\phi) \\ &= \frac{\sum_{t'} \rho(t',\phi)}{n_{\mathrm{t}}} \\ t &= t(a,b,c,d,e) \end{aligned} \tag{22}$$

We are using four different values for e: (i) $e = 1$, neither b nor c is part of a ring system; (ii) $e = 2$, b or c (exclusive) is part of a ring system; (iii) $e = 3$, b and c are part of different (not fused) ring systems; and (iv) $e = 4$, b and c are part of the same ring system.

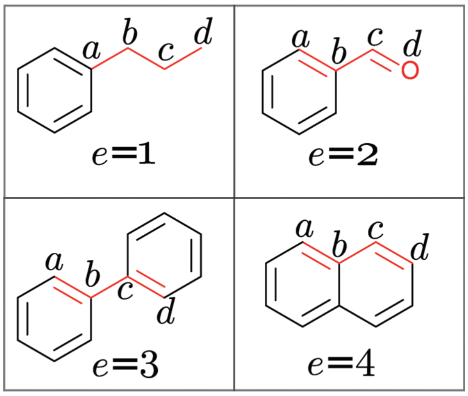


Figure 3. Examples to illustrate different values for the qualifier e.

An example for all four types is given in Figure 3.

The primary intention for the torsion score is to penalize unlikely torsion angles rather than a good correlation with correct torsional energies. As a particular bond can be part of more than one torsion, the score for each bond is calculated as the mean of all torsions it participates in. A clustering of torsion types t is not necessary, as only a set of rather general atom types is considered (see Methods).

DrugScore SAS- and DSX SR-Potentials. To account for desolvation effects, Gohlke et al.[35] introduced a statistical potential in DrugScore that is based on the solvent-accessible surface (SAS). Individual SAS potentials for either protein atom types p and ligand atom types l can be derived from a database, but we will only use l in the following equations, as the formalism is identical.

$$\begin{aligned} \mathrm{score}^{\mathrm{DS}}_{\mathrm{SAS}}(l,\mathrm{SAS}) &= -\ln\left(\frac{\rho(l_{\mathrm{complexed}},\mathrm{SAS})}{\rho_{\mathrm{ref}}(l_{\mathrm{uncomplexed}},\mathrm{SAS})}\right) \\ \rho(l,\mathrm{SAS}) &= \frac{N(l,\mathrm{SAS})}{\sum_{\mathrm{SAS}'} N(l,\mathrm{SAS}')} \end{aligned} \tag{23}$$

For an isolated atom a_l, its SAS corresponds to the surface of a sphere with radius $r(l) = r_{\mathrm{vdW}}(l) + 1.4$ Å, because 1.4 Å is the approximate radius of a sphere occupied by a water molecule. The SAS for an atom in the complexed state is calculated as the part of its surface that is not in contact with any other protein or ligand atom. The SAS for a protein atom in the uncomplexed state does not consider ligand atoms, and the SAS for a ligand atom in the uncomplexed state does not consider protein atoms, respectively. Gohlke et al.[35] denote the SAS in the uncomplexed state as SAS_0, and we will use both terms synonymously. Parts of SAS_0 that correspond to polar atoms are not excluded from SAS if the contacting atom in the complex is also polar, because hydrophilic groups transferred from the solvent to a polar protein environment should exhibit roughly balanced desolvation contributions. In eqs 7 and 8 of the original paper, Gohlke et al.[35] used $\Delta W_i(\mathrm{SAS}, \mathrm{SAS}_0)$, which requires a more detailed specification. As illustrated in Figure 6 of the original paper, the potentials for a given atom type only depend on the SAS of atom i. In detail, the definition would be $\Delta W_i(\mathrm{SAS}, \mathrm{SAS}_0 = \mathrm{SAS})$, which becomes more obvious considering the probabilistic definition given in eq 24:

$$\mathrm{score}^{\mathrm{DS}}_{\mathrm{SAS}}(l,\mathrm{SAS}) = -\ln\left(\frac{P(\mathrm{SAS}|l_{\mathrm{complexed}})}{P(\mathrm{SAS}|l_{\mathrm{uncomplexed}})}\right) \tag{24}$$

Given an atom type p or l, the score is the preference to find a complexed atom with a particular SAS compared to the same SAS in the uncomplexed state.

The use of an SAS-dependent term as defined in eq 24 to score desolvation effects can be questioned, because only changes in the solvent accessible surface ΔSAS can contribute to binding energy, but the SAS as calculated here only contains averaged information about this difference. Therefore, it does not allow one to deduce ΔSAS for a specific complex. In other words, the original DrugScore SAS potentials define a score based on the probability to find a specific atom type with a defined degree of burial in protein−ligand complexes, but they do not measure effects depending on the actual ΔSAS. In analogy to eq 18, one could define a ΔSAS-dependent potential as

$$\text{score}(l, \Delta\text{SAS}) = -\ln\left(\frac{P(\Delta\text{SAS}|l)}{\bar{P}(\Delta\text{SAS})}\right) \tag{25}$$

but instead, we decided to use a potential that holds information about both, the preference for a distinct SAS (as in DrugScore) and the amount of ΔSAS. Therefore, for each atom a_l we calculate a ratio

$$\text{SR}(a_l) = \frac{\text{SAS}(l_{\text{complexed}})}{\text{SAS}(l_{\text{uncomplexed}})} = 1 - \frac{\Delta\text{SAS}}{\text{SAS}_0} \tag{26}$$

and defined the *DSX*-SR-potential as shown in eq 27,

$$\begin{aligned}
\text{score}_{\text{SR}}^{DSX}(c, \text{SR}) &= -\ln\left(\frac{P(\text{SR}|c)}{\bar{P}(\text{SR})}\right) = -\ln\left(\frac{\rho(c, \text{SR})}{\rho_{\text{ref}}}\right) \\
\rho(c, \text{SR}) &= \frac{\sum_{l \in c} N(l, \text{SR})}{\sum_{l \in c} \sum_{\text{SR}'} N(l, \text{SR}')} \\
\rho_{\text{ref}} &= \rho_{\text{ref}}(\text{SR}) = \frac{\sum_{c} \rho(c, \text{SR})}{n_c}
\end{aligned} \tag{27}$$

where c denotes a cluster of ligand-atom types and n_c is the number of clusters. We like to point out again that in case of eq 23, the SAS in an uncomplexed state is an averaged value for the entire database, whereas in eq 26 it is a specific value for each individual atom in a specific protein−ligand complex. The SR potentials for protein atoms are calculated analogously.

METHODS

For all purposes of atom-type perception, ring perception, or generally parsing of input files, the fconv libraries[52] were used. It is important to use identical atom type assignments in both processes, scoring and derivation of the potentials, to reduce the bias of systematic errors in the atom-type perception. Furthermore, reassigning atom types in rescoring makes the program independent from differences in the docking solutions with respect to their atom types (which may differ among different docking programs). For higher consistency, we also decided to generally ignore any predefined hydrogens and set standard protonation states (see definition files in the Supporting Information).

We derived distance-dependent pair potentials, torsion angle potentials, and SR potentials. Whereas torsion angle potentials are derived from the CSD only, the SR potentials originate from the PDB only. The total *DSX*-score for a given protein−ligand complex is given by eq 28,

$$\begin{aligned}
\text{score}_{\text{total}} &= w_{\text{p}}\text{score}_{\text{pair}} + w_{\text{t}}\text{score}_{\text{tors}} + w_{\text{s}}\text{score}_{\text{SR}} \\
\text{score}_{\text{pair}} &= \sum_{a_i \in P} \sum_{a_j \in L} \text{score}_{\text{pair}}^{DSX}(c(a_i, a_j), r(a_i, a_j)) \\
\text{score}_{\text{tors}} &= \sum_{b} \sum_{T \in b} \frac{\text{score}_{\text{tors}}^{DSX}(t(T), \phi(t))}{n_{\text{T}}} \\
\text{score}_{\text{SR}} &= \sum_{a \in P} \text{score}_{\text{SR}}^{DSX}(c(a), \text{SR}(a)) \\
&\quad + \sum_{a \in L} \text{score}_{\text{SR}}^{DSX}(c(a), \text{SR}(a))
\end{aligned} \tag{28}$$

where a is an atom from either set of protein atoms P or the set of ligand atoms L, c is a cluster type, b is a central bond of a torsion T, t is a torsion type, n_{T} is the number of torsions for a given bond, SR is the SAS-ratio for a protein or ligand atom, and the $w_{\text{p/t/s}}$ are the weighting factors used.

To enable an unbiased comparison (not trained for a particular test set), we did not adjust the weightings of the individual potentials but only toggled them on or off with a weighting of 1.0 or 0.0. For validation, we used eight different schemes, where DSX^{CSD} denotes derivation of pair potentials from the CSD and DSX^{PDB} denotes derivation from the PDB:

DSX^{CSD} :: Pair	: $w_{\text{p}} = 1.0,$	$w_{\text{t}} = 0.0$	$w_{\text{s}} = 0.0$
DSX^{CSD} :: PairSR	: $w_{\text{p}} = 1.0$	$w_{\text{t}} = 0.0$	$w_{\text{s}} = 1.0$
DSX^{CSD} :: Tors	: $w_{\text{p}} = 0.0$	$w_{\text{t}} = 1.0$	$w_{\text{s}} = 0.0$
DSX^{CSD} :: PairTors	: $w_{\text{p}} = 1.0$	$w_{\text{t}} = 1.0$	$w_{\text{s}} = 0.0$
DSX^{CSD} :: All	: $w_{\text{p}} = 1.0$	$w_{\text{t}} = 1.0$	$w_{\text{s}} = 1.0$
DSX^{PDB} :: Pair	: $w_{\text{p}} = 1.0$	$w_{\text{t}} = 0.0$	$w_{\text{s}} = 0.0$
DSX^{PDB} :: SR	: $w_{\text{p}} = 0.0$	$w_{\text{t}} = 0.0$	$w_{\text{s}} = 1.0$
DSX^{PDB} :: PairSR	: $w_{\text{p}} = 1.0$	$w_{\text{t}} = 0.0$	$w_{\text{s}} = 1.0$

Pair Potentials. Similar to DrugScore, two different knowledge bases were used to derive the potentials for *DSX*. The first is the Protein Data Bank[55] (PDB) and the second is the Cambridge Structural Database (CSD).[54]

In the PDB-case, an initial list of 37 067 X-ray-structures with a resolution up to 2.4 Å and containing at least one ligand was used (see the Supporting Information). Only contacts between atoms with *B*-factors ≤ 40 Å^2 and occupancies ≥0.5 were considered. We also derived a set of potentials after exclusion of all structures being part of the primary test set (see Test Sets and Validation), but we could not observe a difference in the validation as presented in the Results and Discussion section. This is not surprising, as the test set represents only 0.5% of the knowledge base. All HETATM molecules (including cofactors) with more than five non-hydrogen atoms and also water molecules were considered as ligand. When processing one of these ligands, the remaining part of the HETATMs was considered as part of the protein.

In the CSD-case we used ConQuest[56] to query the database for all structures with an *R*-factor ≤ 0.075, at least one carbon, no error flag set, and completeness of all coordinates. After removal of duplicates (some structures have two entries, with and without hydrogens, respectively), for the resulting 345 726 structures (see the Supporting Information) the crystal packings were generated using fconv.[52] To evaluate the contact data, the central molecule of each packing was treated as a ligand and the surrounding molecules as a protein. Therefore, the packings were generated with a size that guarantees to consider all atoms within a range of 6 Å around the central molecule. In the case of different molecules in the unit cell, each of them was treated as the "ligand" once.

All contact data were derived symmetrically; hence, contact type *A_B* is equal to *B_A*. Although this appears obvious only in the CSD-case, we also achieved better results in the PDB-case when not using asymmetric data.

To account for the inherent limitation of low occurrence frequencies for some of the desired atom types we did not use the full set of fconv atom types but we merged some of them initially. As a result, there are 17 types for carbon, 24 for nitrogen, 10 for oxygen, 4 for sulfur, 2 for phosphorus, F, Cl, Br, I, and 7 different metal ion types. The corresponding fconv definition file is available as Supporting Information and gives details about the atom types used. It also holds information about the standard protonation rules applied in the atom type perception.

Only contact types with more than 1000 contacts (within 6 Å) in the database were considered for further processing. Furthermore, we neglected all types where not at least one of the two atoms was either a carbon, nitrogen, oxygen, sulfur, or phosphorus. After applying these filters, we obtained 930 contact types in the PDB-case and 1561 in the CSD-case. The lower number of different contact types in the PDB case is not just due to the smaller database but due to the fact that many atom types are never part of a protein molecule. If a contact type which remained unconsidered due to too low occurrence frequency has to be handled in rescoring, it is mapped to the most similar contact type with sufficient occurrence frequency. The criterion for similarity is the same as the one used for clustering (see below). In case a completely unknown contact type appears, it will not be considered. However, this situation will be rather rare and not of significant influence as many additional contacts for this distinct atom will be regarded.

In contrast to DrugScore and most of the other knowledge-based scoring functions, we use a bin size of 0.01 Å instead of 0.1 Å for both, deriving the contact data and the resulting potentials. It is important to note that, due to the smoothing function subsequently applied to the data, this has no impact on the statistical significance. Whether we would use a 0.0001 Å or 0.01 Å binning is irrelevant as long as we apply the same smoothing function. For *DSX* we use a Gaussian kernel for smoothing; hence, it is the σ (parameter determining the width of the Gaussian function) that is relevant for an appropriate signal-to-noise ratio. We have chosen σ = 0.15 Å, as this is in good agreement with the triangular smoothing applied by Gohlke et al.[35] In the original DrugScore paper, it was argued that the uncertainties in crystallographically determined coordinates are the rationale for smoothing. However, these uncertainties are already considered while averaging over a large number of complexes from the database. The actual rationale for smoothing is to increase the signal-to-noise ratio. This implies that there should be an optimal σ for each individual contact type, depending on the number of such contacts in the database and the distribution of the contact distances. Generally, a higher value for σ should be used for lower occurrence frequencies, but in the case of eq 15, averaging over density functions with different σ levels (lower σ for higher occurrence frequency) would increase the impact of highly populated contact types in the reference. Moreover, it would complicate the similarity measurement for the density functions. Thus, we decided to use a constant σ for all contact types. We have chosen a narrower binning of 0.01 Å to avoid a second smoothing in the scoring process. If a distance falls close to the border between two bins, an average value of both bins should be applied. If this smoothing is neglected, a high difference in the score is possible for rather small differences in the contact distances. With the use of a small bin size of only 0.01 Å, the differences between two bins are negligible and a smoothing with neighboring bins can be avoided in the scoring process, which speeds up computing.

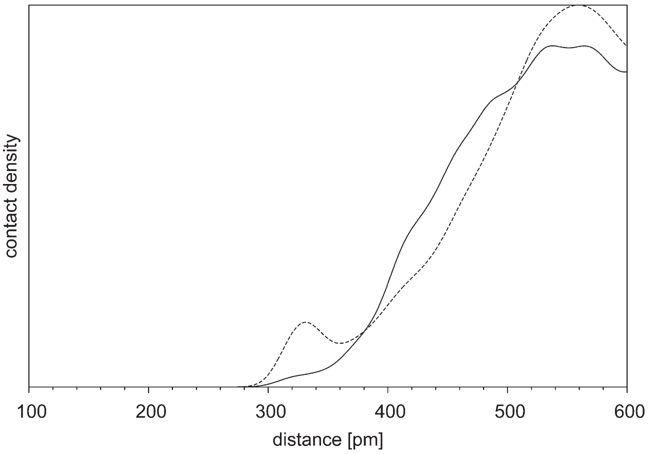


Figure 4. Density functions for two contact types processed from the CSD: Cl.0_P.o (solid line) contact between an organic chlorine and phosphorus bound to at least one oxygen; Cl.0_P.3 (dotted line) contact to other phosphorus atoms.

To cluster the contact types $c = p\_l$, we implemented a hierarchical approach with complete linkage. We evaluated different distance metrics for the $\rho(p,l,r)$ and obtained the best results using squared Euclidean distances:

$$\mathrm{dist}(\rho_a, \rho_b) = \sum_{r=1.00}^{r=5.50} (\rho_a(r) - \rho_b(r))^2 \qquad (29)$$

Distances were multiplied by a factor 10.0, if a and b were contact types that not only differed with respect to the atom types but also with respect to the element types. There is no general rule how to choose an appropriate distance threshold for clustering. After visual inspection of some density functions of different distance levels, we merged the PDB potentials down to a set of 300 contact types and the CSD potentials to 600 contact types, corresponding to maximum distances of 0.0028 and 0.0026, respectively. Figure 4 shows the unmerged density functions with the lowest distance in the CSD case. With the use of the mentioned thresholds, no contacts with different element types were merged.

For low contact distances, the reference density in eq 15 is not well-defined, simply because no structural data are available in this distance range. Usually, for simple rescoring no problem should occur, as the docking programs avoid clashes. However, to enable minimization, a repulsive term is attached in the range from 0 Å up to the first maximum that is followed by a negative potential value. The actual functional form of the repulsive term can be selected arbitrarily, but a smooth connection is desired. We have chosen to attach a function starting with the gradient of 0.025 and linearly decreasing this gradient to 0 with decreasing the distance to 0 Å.

Furthermore, we derived pair potentials that are tailored for usage in hotspot analysis. Here, we used a reduced set of pharmacophoric atom types, in detail: donor, acceptor, donor—acceptor, aromatic, hydrophobic, and metal. If an fconv atom type could not uniquely be assigned to one of these six categories, the element type was used as a dummy. The fconv definition file (with the mapping from fconv types to pharmacophore types) is available as Supporting Information. In case of the CSD knowledge base, we obtained 66 different contact types that were merged into 62 clusters. In the PDB case, we obtained 50 different contact types that were used without any clustering.

Torsion Angle Potentials. To derive torsion angle potentials, we used the same CSD data set as for the *DSX*-pair potentials. A 2°-binning was used for a range from $\phi = 0°$ to $\phi = 180°$. For smoothing, we used a Gaussian kernel with $\sigma = 5°$, and we only considered torsions with more than 50 occurrences in the database. To cover most of the torsion types that occur in ligands, the atom types are reduced according to the following scheme: "Element.Hybridization" in the case of carbon, nitrogen, and oxygen, "Hal" in the case of halogens, and "Element" for all other elements. The result are 4464 different torsion types. With respect to the primary test set (see below), only seven different torsion types were not sufficiently represented by the database. However, in only four ligands there is an unconsidered torsion type where not at least one other torsion potential for the corresponding bond is available.

SR Potentials. The structures used to derive PDB-pair potentials were also used to derive SR potentials. To approximate the SAS we use spherical grids with precomputed coordinates for each element type. All grids consist of 162 points which were calculated by 2-fold subdivision of an icosahedron and subsequent scaling to a radius of r_{vdW} + 1.42 Å (to account for the space occupied by a water molecule). For nitrogen, oxygen, and sulfur we used a vdW-radius decreased by 0.2 Å to account for putative hydrogen-bond formation. For a nitrogen for instance, this results in $r_{\mathrm{grid}} = r_{\mathrm{vdW}} + 1.42 = 2.77$ Å and a grid of a mean closest point-to-point distance of 0.81 Å with a standard deviation of 0.01 Å. The mean distance to the closest 6 points is 0.88 Å with a standard deviation of 0.08 Å. The SR for each ligand atom is then calculated with eq 30

$$\mathrm{SR}(a_{\mathrm{l}}) = \frac{\mathrm{points}_{\mathrm{complexed}}(a_{\mathrm{l}})}{\mathrm{points}_{\mathrm{uncomplexed}}(a_{\mathrm{l}})} \tag{30}$$

with a_{l} as ligand atom, $\mathrm{points}_{\mathrm{uncomplexed}}$ as the number of grid points not contacted by other ligand atoms, and $\mathrm{points}_{\mathrm{complexed}}$ is the number of grid points not occupied by other ligand- or protein atoms. If a_{l} is a nitrogen or oxygen atom, any contacting nitrogen and oxygen atoms of the receptor are not considered for $\mathrm{points}_{\mathrm{complexed}}$. The SR for protein atoms is calculated correspondingly. It is calculated for those protein atoms that are in SAS-contact to at least one ligand atom.

A bin size of 0.01 was used for the SR, and we applied a Gaussian kernel with $\sigma = 0.08$ for smoothing. The same atom type classification as used for the pair potentials has been considered, and only types with more than 50 occurrences in the database were regarded for further processing.

For clustering, squared Euclidean distances were used again:

$$\mathrm{dist}(\rho_{\mathrm{a}}, \rho_{\mathrm{b}}) = \sum_{\mathrm{SR}=0.00}^{\mathrm{SR}=1.00} (\rho_{\mathrm{a}}(\mathrm{SR}) - \rho_{\mathrm{b}}(\mathrm{SR}))^2 \tag{31}$$

In the case of receptor atoms, we did not merge any types, as they are sufficiently different. For ligand atoms we have also chosen a very low distance threshold resulting in 50 clusters (from 68 atom types).

Ligand Relaxation. *DSX* optionally features a local relaxation of docking poses. Therefore, we adopted Powell's method as described in the Numerical Recipes.[57] SR potentials are not used in the minimization, and for torsion- and pair-potentials the same weightings as specified for rescoring are used. Additionally, intramolecular interactions are also evaluated using the same pair potentials used for intermolecular interactions. Here, only interactions between atoms separated by at least four bonds are regarded.

Volume Correction. To calculate volume corrections for the PDB-derived contact data, we used a spherical grid-based approach similar to that described for the SR potentials. Here, we used a 5-fold icosahedron subdivision resulting in 10 242 points. The points were scaled according to the radius under investigation, whereat a 0.2 Å binning was used. For each radius bin, the corresponding points were evaluated with respect to their neighboring atoms. If no surrounding atom was closer than its vdW-radius plus 1.4 Å, the point under investigation was counted as unoccupied. The additional 1.4 Å were used to account for the volume of a putative contacting atom. Strictly speaking, one should derive individual data for each possible contacting element type using the vdW-radii of these elements. However, as we are interested in the relations between different atom types and not in absolute values, the fixed radius should be a sufficient approximation. The available volume fraction for an atom type *a* was calculated as

$$f(a,r) = \frac{\mathrm{points}_{\mathrm{unoccupied}}(a,r)}{\mathrm{points}_{\mathrm{total}}(a,r)} \tag{32}$$

Implementation Details. *DSX* is implemented in ISO C++, and binaries for Linux and MacOS are freely available (see Software Available).

Valid input formats for *DSX* are PDB or MOL2 for proteins and MOL2 or DLG for ligand files. Cofactors, metals, and water molecules can be supplied separately or together with the protein (when in MOL2 format). The user can choose between different interaction modes that specify whether cofactors, metal ions, and/or water molecules should be handled individually or as part of the protein. If for example the cofactor was kept rigid during the docking process, it should be considered as part of the protein in rescoring. However, if it was kept flexible upon docking, also the interactions between cofactor and protein should be rescored. In consequence, when choosing an interaction mode with individual cofactor, this cofactor must be supplied as an additional MOL2 file with a number of cofactor poses equal to the number of generated ligand poses.

DSX generally ignores hydrogens in the input files, thus the results are independent from any predefined protonation states. Additionally, the program always redefines the atom types using the same routines that were used for atom-type perception when deriving the potentials.

For docking solutions obtained by AutoDock, where amino acid side chains of the protein were kept flexible, the user can switch-on a flag to consider the correct side chain conformations for each solution. In that case, the original DLG file has to be supplied. For docking results using GOLD with flexible side chains, the necessary information is included in the MOL2 files, but currently the protein must be supplied in PDB format if *DSX* is supposed to consider the correct side chain conformations.

If GOLD was used to dock with explicit water molecules, there is a flag to consider the corresponding information in the GOLD result files. If the check for covalently bound ligands is turned on, *DSX* will ignore all atoms participating in a protein—ligand bond and also their neighboring atoms. The weightings used for the different types of potentials can be freely assigned by the user.

Furthermore, a PyMOL-based[58] visualization similar to the work of Block et al.[59] was implemented. Favorable and unfavorable per-atom scores are visualized by blue and red spheres, respectively, where the sphere's radius scales with the absolute value of the score. Additionally, single contacts with very high or low scores are visualized as red or blue lines, and also unfavorable torsion angles are displayed.

Test Sets and Validation. To evaluate the above-mentioned docking-, ranking-, and scoring power, we used the test set prepared by Cheng et al.[18] It consists of a primary set of 195 protein—ligand complexes and four additional sets to assess ranking- and scoring power. The authors of this test set evaluated 16 different scoring functions, making it one of the most comprehensive comparisons up until now. The primary set was compiled from the PDBbind database,[60,61] regarding quality and diversity of the structures and considering only complexes with experimentally determined binding constants. It covers 65 diverse targets, and each target is represented by three complexes, one of them with high binding affinity, one with low affinity, and one close to the mean. Up to 100 highly diverse decoy poses were generated for each complex using various docking programs followed by a subsequent cluster analysis. This primary set can be downloaded from the PDBbind, including all decoy structures and thus enabling a comparison using identical input as in the original publication. The four additional test sets consist of 112 HIV protease-, 73 trypsin-, 44 carbonic anhydrase-, and 38 thrombin complexes with known binding constants, respectively.

To assess the docking power of *DSX*, we calculated five different success rates on the primary test set and compared them to the results given in the Supporting Information (part VI) by Cheng et al.[18] In two cases, a success is defined as finding the crystal pose on rank 1 or among the first five ranks, respectively. In the other three cases, a success is defined as finding a docking pose approximating the crystal structure with an rmsd ≤ 2.0 Å on rank 1, among the first five ranks, or on rank 1 excluding the crystal pose, respectively. There are five complexes where all decoys have an rmsd ≤ 2.0 Å (1df8, 1fcx, 1fcz, 1fd0, 2f01) and seven complexes where all decoys have an rmsd > 2.0 Å (1a30, 1elb, 1nhu, 1tyr, 1u1b, 2fzc, 6rnt). Therefore, Cheng et al.[18] computed the success rate using eq 33, where *S* is the number of success cases.

$$\text{success rate} = \frac{S - 5}{195 - 5 - 7} \cdot 100\% \qquad (33)$$

This was not explicitly stated in the original paper but kindly provided by the authors on request.

To assess ranking- and scoring power, we calculated the Spearman and Pearson correlation coefficients for the four additional test sets and compared them with the results given in the Supporting Information (part VII) by Cheng et al.[18] We also calculated a success rate for ranking power based on the primary set and compared the results with the information given in Table 4 by Cheng et al.[18] Here, a success is achieved if the three complexes for one of the 65 targets are ranked in the correct order with respect to their binding constants. Furthermore, we calculated the Pearson correlation for the complete primary test set. For all tests, we used *DSX* in version 0.88. CSD and PDB potentials were used in version 05/11.

RESULTS AND DISCUSSION

From the results of Cheng et al.,[18] we only list the best performing variants of each scoring function except for DrugScore, where all results are given. The missing results are available in the Supporting Information by Cheng et al.[18] We also applied the pharmacophoric pair potentials to the test set, although they are mainly intended for hotspot analysis. They are abbreviated by DSX^{CSD}::Pharm and DSX^{PDB}::Pharm.

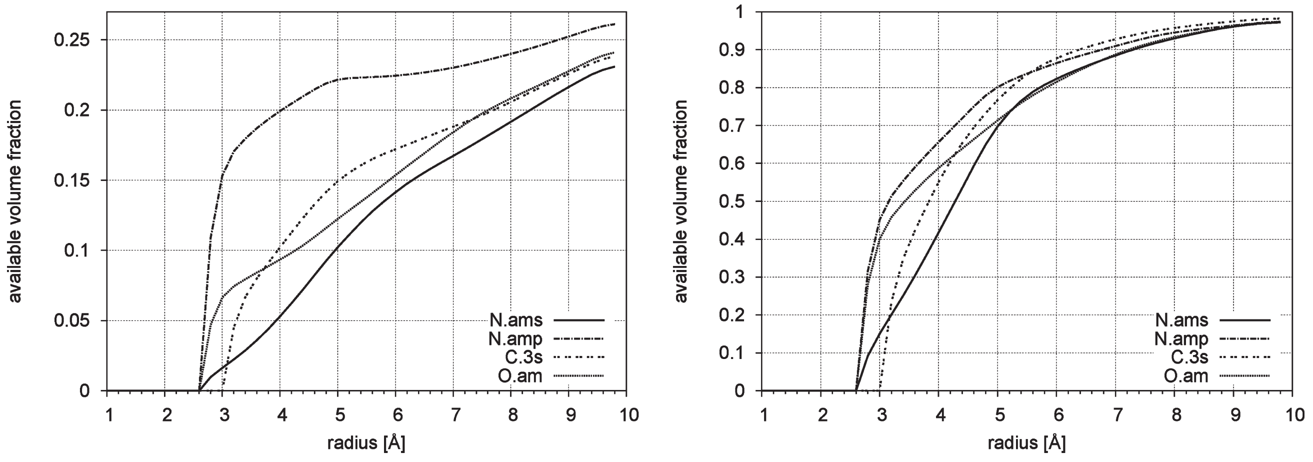
In the Theory section we mentioned that a volume correction is of higher importance for protein atoms, because there are significant differences in the available volume, especially for backbone atoms compared to side chain atoms. To assess the influence of a volume correction for PDB-based potentials, we calculated available volume fractions for protein and ligand atoms, respectively. Qualitatively, our results shown in Figure 5 are similar to what was found by Mooji and Verdonk[38] (see Figures 4 and 5 in their paper). As the fractions were derived for a distinct classification into protein- and ligand atoms, they can only be applied to asymmetric PDB data. When using a DrugScore-like reference state, Mooij and Verdonk[38] found the potential for a contact between protein backbone amides and aromatic nitrogens to never approach negative (favorable) values (Figure 6 in the original paper). In contrast, Figure 6 shows a minimum at negative values not only when we derive symmetric potentials but also when we derive asymmetric potentials from the PDB. If we apply the volume correction in the asymmetric case, we observe the expected effect of more pronounced minima and also an improvement in the docking- and ranking power in our validation. Surprisingly, the symmetric variant with less pronounced minima performs even better than both asymmetric variants. We can only speculate about the reasons. In the case of asymmetric data, certain contact types *A_B* have very low occurrence frequencies, whereas *B_A* is well populated. If such a type *A_B* has negative impact on the reference, the performance of asymmetric potentials decreases. In the symmetric case, *A_B* and *B_A* are merged to one type that is dominated by *B_A* contacts; hence, a possible bias of statistically underrepresented *A_B* is alleviated. For now, we decided to neglect the volume correction, but it could be an interesting aspect for further investigation.

Corresponding to the density functions shown in Figure 1, Figure 7 shows the resulting *DSX* pair potentials.

Figure 8 gives an example for torsion angle potentials. The solid line corresponds to a carbon chain where all atoms have hybridization sp^3, whereas the central bond is formed by two sp^2 hybridized carbons in the case of the dotted line.

An example for SR potentials is shown in Figure 9. The amount of solvent accessible surface that becomes buried upon ligand binding increases with decreasing SAS ratio given on the *x*-axis. Higher magnitudes in the case of ligand atoms indicate higher changes of the SAS upon complex formation compared to protein atoms.

Docking Power. Table 1 shows the validation similar to the results recorded in Table S6, S7 and S8 from the Supporting Information of Cheng et al.[18] The most important number can be found in the last column and corresponds to the success



a) Atom types in proteins.

b) Atom types in ligands.

Figure 5. Available volume fractions. N.ams, nitrogen in secondary amides; N.amp, nitrogen in primary amides; C.3s, secondary sp^3 carbon; O.am, amide oxygen.

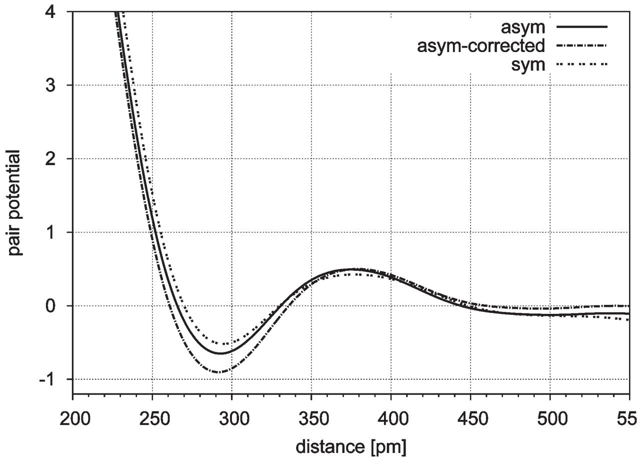


Figure 6. PDB pair potentials for contacts between N.ams (secondary amide) on the protein side and N.ar6 (aromatic nitrogen) on the ligand side (asym, derivation with differentiation between N.ams_N.ar6 and N.ar6_N.ams but without volume correction; asym-corrected, with volume correction; sym, symmetric contact types without volume correction).

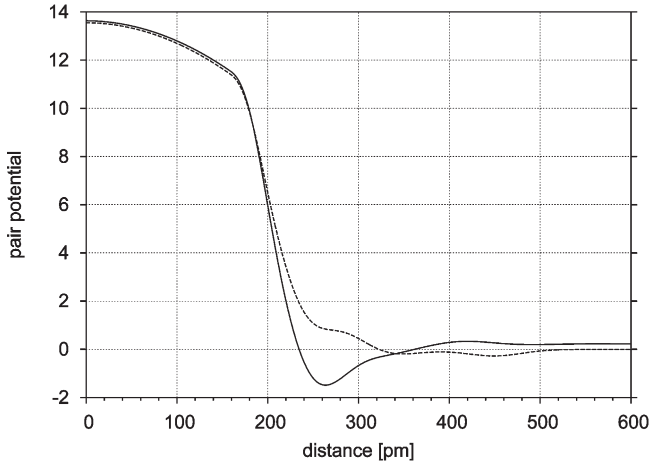


Figure 7. Pair potentials for two contact types processed from the CSD: O.3oh_O.carb (solid line), contact between hydroxyl oxygen and carbonyl oxygen; O.3et_O.carb (dotted line), contacts between oxygen in aliphatic ether and carbonyl oxygen.

rate of finding solutions with rmsd ≤ 2.0 Å on rank 1, when the native, crystallographically determined ligand geometry is excluded from the decoy set. In a virtual screening run, this solution would be the relevant pose that is compared to the top ranks of other compounds. Therefore, it must be as close as possible to a native geometry to allow for a reliable compound selection.

For both, the CSD- and the PDB-case, the combination of pair- and SR-potentials shows an improvement compared to the pair potentials alone. It has to be noted that the *DSX*$^{\text{CSD}}$::PairSR mode is a combination of information retrieved from the CSD and the PDB, whereas the *DSX*$^{\text{PDB}}$::PairSR mode only relies on PDB data. Interestingly, the differences between CSD- and PDB-derived potentials and their combinations are only marginal with respect to docking power.

A combination with torsion potentials (*DSX*$^{\text{CSD}}$::PairTors) increases the recognition of native geometries but decreases the success rate when the native pose is excluded. This indicates that they are very sensitive to deviations from ideal geometries as found in the CSD. Native geometries that are ranked on first place by *DSX*$^{\text{CSD}}$::Pair but not by *DSX*$^{\text{CSD}}$::PairTors are 2g94, 7cpa, 2bok, and 1bma. In contrast, *DSX*$^{\text{CSD}}$::PairTors ranks native geometries of 1xgj, 2azr, 1sl3, 2bz6, 2std, 1a30, and 1rnt on first place, but *DSX*$^{\text{CSD}}$::Pair does not. From the mentioned 11 structures, 8 have very large ligands with many rotatable bonds. For such structures, there is a higher chance for docking programs to fail with one of these numerous bonds; hence, it is easier to differentiate between native pose and docking solutions with respect to torsion angles. One could also speculate that, in contrast to small molecule crystal packings, higher deviations from ideal geometries are possible in protein–ligand complexes. In that case, the CSD-derived torsion angle potentials could penalize native poses too strongly.

We have to point out that the discussed improvements when applying SR and torsion potentials are only marginal. It is not sure whether these terms generally improve the results on other data sets.

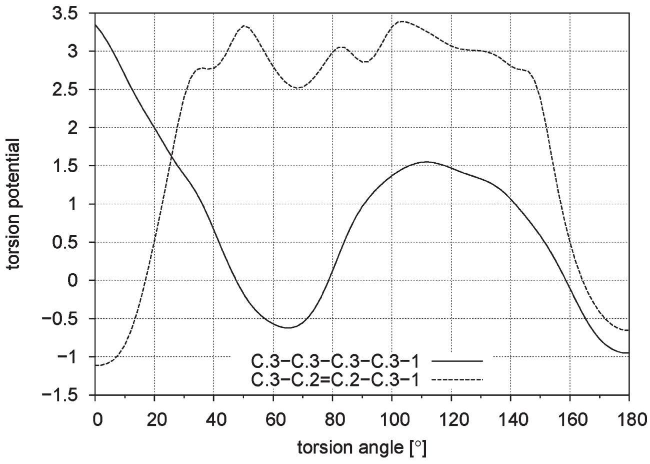


Figure 8. Torsion angle potential for an sp^3 carbon-chain (solid line) and for a carbon chain with a double bond (dotted line).

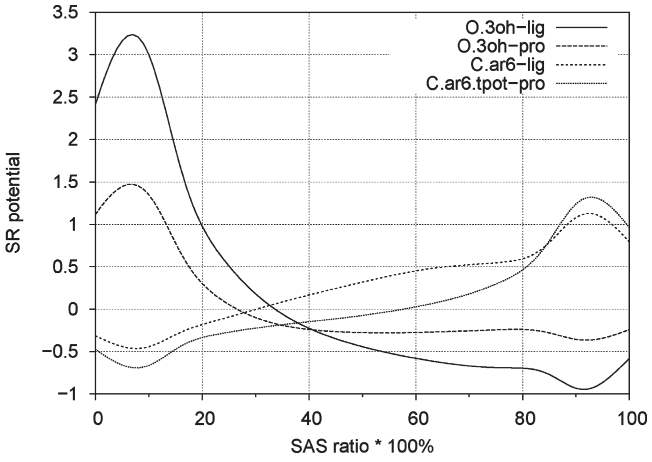


Figure 9. Example for SR potentials: O.3oh-lig, hydroxyl oxygen in ligands; O.3oh-pro, hydroxyl oxygen in proteins; C.ar6-lig, aromatic carbon in ligands; C.ar6-pro, aromatic carbon in proteins.

With respect to docking power, *DSX* outperforms all other functions tested, except for ASP which is on a similar level. For the used test set, the best results are obtained using all three types of potentials in combination.

Ranking Power. Table 2 shows the validation similar to the results recorded in Table 4 by Cheng et al.[18] In the original paper, only the best performing version of each scoring function was evaluated. For comparison, we also present only the results for the best CSD- and PDB-based *DSX* mode, respectively.

In the case of Discovery Studio, Glide, GOLD, and Sybyl, ligand optimization was performed by the functions implemented into these programs.[18] For DrugScore and X-Score, Discovery Studio was used to minimize the ligands in the CHARMm force field.[18] For *DSX*, we used the program's own local minimization in the CSD case, while in the PDB case, no minimization is possible due to the lack of torsion angle potentials.

Table 3 shows achieved ranking correlations for the additional test sets and it corresponds to the results listed in Tables S13, S14, S15 and S16 of the Supporting Information of Cheng et al.[18] The results shown in parentheses were obtained when additionally applying intramolecular interactions with a weighting of $w_{\mathrm{p-intra}} = 1.0$. Interestingly, this improves the correlations except for the case of thrombin.

Table 1. Success Rates (%) for the Evaluation of Docking Power[a]

	Crystal structure on		≤ 2.0 Å pose on		
Scoring function	Top 1 pose	Top 5 poses	Top 1 pose	Top 5 poses	Top 1 pose no cryst.[b]
DS::Jain	1.5	15.4	44.8	79.2	44.8
DS::LigScore2	17.9	49.7	71.6	92.9	69.4
DS::LUDI2	9.7	29.2	57.4	83.6	56.8
DS::PLP1	40.5	56.4	75.4	97.3	68.3
DS::PMF	19.5	44.1	43.7	67.2	39.3
DrugScore$^{\mathrm{CSD}}$::Pair	50.3	79.5	58.5	94.0	25.7
DrugScore$^{\mathrm{CSD}}$::PairSurf	44.6	80.0	54.1	95.6	25.1
DrugScore$^{\mathrm{PDB}}$::Pair	40.0	73.8	74.3	93.4	68.9
DrugScore$^{\mathrm{PDB}}$::PairSurf	39.5	74.9	74.3	95.1	69.4
DrugScore$^{\mathrm{PDB}}$::Surf	3.6	20.0	32.8	80.3	32.2
DSX$^{\mathrm{CSD}}$::Pair	50.8	77.4	83.6	95.6	77.6
DSX$^{\mathrm{CSD}}$::PairSR	51.3	79.0	84.7	96.2	78.1
DSX$^{\mathrm{CSD}}$::PairTors	52.3	77.4	84.2	95.1	77.0
DSX$^{\mathrm{CSD}}$::All	52.8	77.9	85.2	96.2	79.2
DSX$^{\mathrm{CSD}}$::Tors	8.7	20.0	38.3	76.5	36.1
DSX$^{\mathrm{PDB}}$::Pair	50.3	78.5	84.2	95.6	75.4
DSX$^{\mathrm{PDB}}$::PairSR	51.8	77.9	84.7	95.6	78.7
DSX$^{\mathrm{PDB}}$::SR	3.6	16.9	39.3	82.5	38.3
DSX$^{\mathrm{CSD}}$::Pharm	47.2	76.4	79.8	95.6	73.2
DSX$^{\mathrm{PDB}}$::Pharm	41.5	72.3	77.6	94.0	69.4
GOLD::ASP	36.9	71.8	82.5	95.6	77.6
GOLD::ChemScore	17.9	50.8	70.5	86.9	69.4
GOLD::GoldScore	8.2	28.7	68.9	89.6	68.3
GlideScore::SP	18.5	50.3	73.2	93.4	72.7
SYBYL::F-Score	21.5	49.2	64.5	90.7	60.1
X-Score1.2	32.3	64.6	67.2	91.3	63.4
X-Score1.2::HMScore	30.3	57.9	68.3	90.7	62.3

[a] Results (excluding *DSX*) cited from Cheng et al.[18] [b] The native geometry was not part of the decoy set.

Applying the torsion angle potentials in addition to the pair potentials improves ranking in the case of HIV protease and trypsin but makes the results for carbonic anhydrase worse. Remarkably, applying only torsion angle potentials without any assessment of protein−ligand interactions produces the best ranking correlation of all scoring functions in the case of HIV protease. This emphasizes the disappointing performance of all scoring functions under assessment for this target. The ligands for HIV protease are rather large and have many rotatable bonds. A possible explanation for the (at least significant) correlation with the torsion score is that in the case of ligands with lower affinities, there are often higher deviations from ideal torsion angles.

Intramolecular- and torsion angle potentials exhibit different (positive or negative) impact on ranking power, depending on the target. This implies that there is no unique best weighting scheme for the different potentials, but instead a tailored set of weighting parameters should be identified for each target of interest.

DSX performs best in the case of thrombin, is second best after X-Score in the case of trypsin, and performs second best after PLP2 in the case of carbonic anhydrase. Astonishingly, for the latter, the pharmacophoric potentials show significantly higher correlations compared to the highly specialized pair potentials.

As in the case of DrugScore and also for *DSX*, the CSD-based potentials have a higher ranking power compared to the PDB-based analogues. In contrast to docking power, the application of SR potentials decreases ranking power in most cases.

Table 2. Success Rates (%) for the Evaluation of Ranking Power on the Primary Test Set[a]

Scoring function[b]	on original complex structures	on optimized complex structures
X-Score::HSScore	58.5	52.3
DSX$^{\mathrm{CSD}}$::All	55.4	52.3
DS::PLP2	53.8	46.2
DSX$^{\mathrm{PDB}}$::PairSR	52.3	/
DrugScore$^{\mathrm{CSD}}$::PairSurf	52.3	49.2
SYBYL::Chemscore	47.7	52.3
SYBYL::D-Score	46.2	46.2
SYBYL::G-Score	46.2	36.9
GOLD::ASP	43.1	49.2
DS::LUDI3	43.1	43.1
DS::Jain	41.5	35.4
DS::PMF	41.5	35.4
SYBYL::PMF-Score	38.5	33.8
GOLD::ChemScore	36.9	41.5
DS::LigScore2	35.4	47.4
GildeScore::XP	33.8	35.4
NHA[c]	32.3	32.3
GOLD::GoldScore	23.1	38.5

[a] Results (excluding *DSX*) cited from Cheng et al.[18] [b] Scoring functions are ranked by their success rates. [c] Ranking by the number of heavy atoms of each ligand.

Table 3. Spearman Correlations for the Four Additional Test Sets[a]

Scoring function	HIV protease	trypsin	carbonic anhydrase	thrombin
DS::Jain	0.023	0.698	0.133	0.491
DS::LigScore1	0.106	0.536	0.330	0.371
DS::LigScore2	0.167	0.418	0.143	0.424
DS::LUDI2	0.047	0.791	0.405	0.558
DS::PLP2	0.168	0.774	0.772	0.666
DS::PMF04	0.200	0.395	0.612	0.022
DS::PMF	0.200	0.693	0.389	0.275
DrugScore$^{\mathrm{CSD}}$::Pair	0.129	0.737	0.542	0.622
DrugScore$^{\mathrm{CSD}}$::PairSurf	0.147	0.768	0.535	0.617
DrugScore$^{\mathrm{PDB}}$::Pair	0.163	0.744	0.488	0.515
DrugScore$^{\mathrm{PDB}}$::PairSurf	0.170	0.743	0.468	0.535
DrugScore$^{\mathrm{PDB}}$::Surf	0.170	0.743	0.468	0.535
DSX$^{\mathrm{CSD}}$::Pair	0.199 (0.225)	0.762 (0.789)	0.559 (0.611)	0.709 (0.682)
DSX$^{\mathrm{CSD}}$::PairSR	0.184 (0.198)	0.733 (0.756)	0.496 (0.547)	0.679 (0.642)
DSX$^{\mathrm{CSD}}$::PairTors	0.300 (0.319)	0.782 (0.797)	0.413 (0.442)	0.703 (0.668)
DSX$^{\mathrm{CSD}}$::All	0.267 (0.291)	0.752 (0.776)	0.429 (0.454)	0.660 (0.636)
DSX$^{\mathrm{CSD}}$::Tors	0.423	0.744	0.089	0.226
DSX$^{\mathrm{PDB}}$::Pair	0.179 (0.199)	0.753 (0.782)	0.575 (0.580)	0.671 (0.637)
DSX$^{\mathrm{PDB}}$::PairSR	0.160 (0.174)	0.728 (0.758)	0.519 (0.537)	0.663 (0.657)
DSX$^{\mathrm{PDB}}$::SR	0.006	0.239	0.139	0.419
DSX$^{\mathrm{CSD}}$::Pharm	0.109 (0.123)	0.744 (0.762)	0.708 (0.703)	0.744 (0.647)
DSX$^{\mathrm{PDB}}$::Pharm	0.140 (0.151)	0.710 (0.708)	0.753 (0.761)	0.708 (0.588)
GOLD::ASP	0.140	0.744	0.486	0.287
GOLD::ChemScore	0.138	0.280	0.572	0.489
GOLD::GoldScore	0.232	0.052	0.079	0.603
GlideScore::SP	0.183	0.177	0.280	0.525
SYBYL::ChemScore	0.228	0.773	0.631	0.587
X-Score1.2::HSScore	0.214	0.824	0.595	0.586
X-Score1.3::HPScore	0.373	0.815	0.494	0.558
X-Score1.3::HSScore	0.291	0.809	0.555	0.593

[a] Results (excluding *DSX*) cited from Cheng et al.[18]

Scoring Power. Table 4 shows the obtained affinity correlations for the primary test set and corresponds to the results in Table S11 of the Supporting Information of Cheng et al.[18]

After X-Score, *DSX* is second best in this category. The minimization was applied as described for Table 2.

At this point, we want to refer back to our initial statement that scoring power has little importance for rescoring. First, we will discuss its influence on docking power: One has to keep in mind that each rescoring of docking solutions is a consensus scoring, because it is a combination of the scoring function used in docking and the function applied subsequently. To generate reasonable poses, the fitness function used in docking should regard all contributions to binding energy and must weight these contributions correctly. We will call such an ideal scoring function "complete"; hence, completeness is a measure for the amount of considered affinity contributions and the quality of the weighting of these terms. However, some aspects of binding energy can only be evaluated as rough approximations and other aspects can even be neglected. For example, covalent bond energies must not be evaluated because docking programs usually do not modify bond lengths; hence, scoring functions used for docking can be incomplete with respect to bond energies. The same holds true for functions used in rescoring. They can be incomplete with respect to some binding energy contributions that were evaluated by the docking program. Moreover, they can also assign much higher weights to contributions that discriminate between near native and decoy poses. For example, in case of hydrogen bonds, a docking function has to consider distances and angles. If it relied on distances only, it would generate unrealistic geometries. In contrast, a function for rescoring could solely rely on the H-bond angles produced by docking programs and give a much higher weight on H-bond distances (in case these distances are especially valuable to penalize decoy poses). Assigning higher weights to certain terms can decrease affinity correlations for native poses but at the same time increase docking power. Thus, we cannot generally conclude from high scoring power to high docking power. An example is GoldScore, which achieves the lowest affinity correlation for the primary test set (Table 4) but has higher docking power than X-Score1.2::HMScore on this test set (68.3 % vs 62.3 % in the most relevant category), although the latter achieves the highest affinity correlation. As a matter of fact, a function that achieves Pearson correlations of 1.0 for arbitrary data sets could still fail with respect to docking power because it could still calculate better scores for certain decoy poses compared to the native pose. Only a really complete function would be perfect in both aspects. As such an ideal function is unlikely to be developed in the near future, functions intended to exhibit high docking power should focus on terms discriminating near-native from decoy geometries and they must not be trained with respect to affinities. In contrast, scoring power (which is a measure of completeness) should be the key value while developing a fitness function for a docking engine. *DSX* is particularly suited for high docking power as it is not designed to calculate binding energies but relies on likelihoods for given geometries. We believe that this complements the functions typically used in docking, which explains the generally high docking power of knowledge-based scoring functions.

The influence of affinity correlation on ranking power is more straightforward: Of course, high scoring power implies high ranking power. However, Table 3 suggests that for different targets, different scoring functions are most suitable. This is again a consequence of the incompleteness of the scoring functions used nowadays. Functions intended to optimize ranking power should therefore be tailored toward a specific target or they should offer an option to train them for a target. The weightings for the different scoring terms in *DSX* allow at least for a moderate training with respect to a specific target.

Table 4. Pearson Correlations for the Primary Test Set[a]

Scoring function	on original complex structures	on optimized complex structures
DS::Jain	0.316	0.339
DS::LigScore2	0.464	0.479
DS::LUDI3	0.487	0.477
DS::PLP1	0.545	0.529
DS::PMF	0.445	0.294
DrugScore$^{\text{CSD}}$::Pair	0.561	0.589
DrugScore$^{\text{CSD}}$::PairSurf	0.569	0.585
DrugScore$^{\text{PDB}}$::Pair	0.524	0.543
DrugScore$^{\text{PDB}}$::PairSurf	0.531	0.536
DrugScore$^{\text{PDB}}$::Surf	0.520	0.542
DSX$^{\text{CSD}}$::Pair	0.597	0.588
DSX$^{\text{CSD}}$::PairSR	0.598	0.591
DSX$^{\text{CSD}}$::PairTors	0.607	0.599
DSX$^{\text{CSD}}$::All	0.609	0.602
DSX$^{\text{CSD}}$::Tors	0.481	0.478
DSX$^{\text{PDB}}$::Pair	0.567	/
DSX$^{\text{PDB}}$::PairSR	0.571	/
DSX$^{\text{PDB}}$::SR	0.445	/
DSX$^{\text{CSD}}$::Pharm	0.560	/
DSX$^{\text{PDB}}$::Pharm	0.547	/
GOLD::ASP	0.534	0.518
GOLD::ChemScore	0.441	0.528
GOLD::GoldScore	0.295	0.329
GlideScore::XP	0.457	0.555
SYBYL::ChemScore	0.555	0.622
X-Score1.2::HMScore	0.644	0.649

[a] Results (excluding *DSX*) cited from Cheng et al.[18]

Influence of Local Minimization. Figure 10 shows the influence of local minimization applied to the primary test set.

Both the rmsd of the best poses and the rmsd of the poses ranked on first place by *DSX* slightly improve as long as the starting geometry has an rmsd of less than 1 Å. Beyond this threshold, the results obtained after minimization get worse compared to the case without minimization. This observation reveals information about the typical size of a potential valley on the *DSX* score landscape, at least about the valley where the native pose resides. Poses with an rmsd larger than 1 Å usually give rise to minimization into different local minima.

Unfortunately, the ranking with *DSX* becomes worse when applying local minimization to the native poses (Table 2) and even the affinity correlation decreases (Table 4). A possible explanation might be the incompleteness of *DSX*. For example, proper geometry of H-bonds is only implicitly considered to some degree in the sum of pair potentials. During a minimization, the contact distances may be optimized at the price of unrealistic H-bond angles. Furthermore, the weightings of intra- and intermolecular distance-dependent potentials and torsion angle-dependent potentials are not trained on affinities.

The largest improvement upon minimization in the functions native "energy" landscapes is achieved by scoring functions of Gold, Glide, and Sybyl that are also used as target functions during the docking process (Tables 4 and 2). This is not surprising because especially the improvement in affinity correlation upon minimization is a measure of the completeness of the used scoring function, and as we suggested above, completeness is a key feature of fitness functions used for docking.

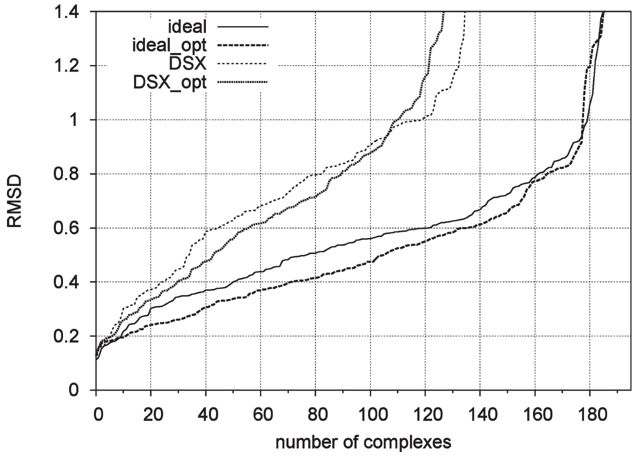


Figure 10. Primary test set docking solutions ordered by increasing rmsd values: ideal, this curve corresponds to the geometry closest to the native pose; ideal_opt, also the geometry closest to the native pose but after minimization of all docking solutions; DSX, the poses ranked on first place by *DSX*$^{\text{CSD}}$::All; DSX_opt, the poses ranked on first place by *DSX*$^{\text{CSD}}$::All after local minimization.

Table 5. Comparison of Runtime for DrugScore and *DSX* on the Primary Test Set

Scoring function	runtime in seconds
DrugScore$^{\text{CSD}}$::CalcPocket	249
DrugScore$^{\text{CSD}}$::Pair	67
DrugScore$^{\text{CSD}}$::PairSurf	3779
DSX$^{\text{CSD}}$::Pair	50
DSX$^{\text{CSD}}$::PairSR	241
DSX$^{\text{CSD}}$::Pair-Opt	3921

Runtime Performance. Table 5 gives some information about the required runtime on the primary test set for *DSX*$^{\text{CSD}}$ compared to DrugScore$^{\text{CSD}}$. To apply the DrugScore Surf potentials (SAS potentials), it is necessary to precalculate binding pockets. For this purpose, DrugScore is bundled with a program named CalcPocket. We used this program to precalculate the 7 Å pockets (with complete residues) around the ligands for each protein, respectively. Both measurements for DrugScore, Pair and PairSurf, were performed with these binding pockets, whereas for *DSX* the complete and unmodified receptor structures were used. For the calculations including a solvent accessible surface term, *DSX* is faster by a factor of 15.7 compared to DrugScore even without considering the time needed by CalcPocket. Also in the case of scoring based on pure pair potential evaluations, *DSX* is significantly faster, although DrugScore uses smaller input structures and *DSX* runtime includes full atom type perception for protein and ligand structures. The last row in the table corresponds to the *DSX* runtime with the built-in local minimization. All values were measured on an Intel Core2Duo E6600 (2.4 GHz).

■ CONCLUSIONS

Our new scoring function *DSX* combines knowledge-based potentials for atom—atom distances with similarly derived torsion angle potentials and a novel measure of the change in the solvent accessible surface. We presented a clustering method to alleviate deficiencies that arise from a combination of DrugScore-like reference states and enhanced atom type definitions. Compared to the original DrugScore formalism, *DSX* pair potentials demonstrated a significant improvement, especially with respect to the recognition of near-native ligand poses (docking power). This docking power is further improved by the application of newly defined surface-dependent SR potentials. The novel torsion angle potentials increased ranking power in the case of HIV protease and thrombin. Furthermore, the combination of all three terms produced the best success rates for docking power.

In comparison to all functions that were evaluated based on the publicly available test set by Cheng et al.,[18] *DSX* achieves the best results with respect to docking power, although not trained for this test set by any aspect. Also with respect to ranking power, *DSX* is among the best three functions. Here, we suggest the usage of pair and torsion angle potentials without usage of the SR potentials. Additionally, scoring of intramolecular interactions improved the correlation for three of the four targets used to assess ranking power.

We did not adjust the individual weightings for the different scoring terms, but we collected evidence that a specific weighting should be adjusted for each target. Therefore, users can configure those weightings on their own to produce a target-tailored scoring function.

We also developed pair potentials that are based on a generic set of pharmacophoric atom types. While intended for hotspot analyses and as a basis for pharmacophore generation, also these potentials demonstrated high ranking power for certain targets.

CSD-based potentials generally yield better results compared to PDB-based potentials. Most likely because of more comprehensive contact data for rare atom types and the general better resolution of small molecule crystal structures.

We discussed the requirements and specialization of scoring functions for docking power, ranking power, or usage as target function in the docking process. In that context, our program *DSX* claims to be most valuable for docking power, while other functions that are more target specific should be used for a proper ranking.

SOFTWARE AVAILABLE

Linux and MacOS binaries of *DSX* and all PDB-based potentials are freely available from our Web site www.agklebe.de (DSX-Online → Get DSX standalone). The software is bundled with another program, named HotspotsX, which generates contour maps based on *DSX* pair potentials. The CSD-based potentials can be obtained on request from the Cambridge Crystallographic Data Centre (CCDC). In addition, the CSD version can also be applied using the *DSX*-online Web-interface from www.agklebe.de.

ASSOCIATED CONTENT

S Supporting Information. The fconv definition files for used atom types and a list of used CSD and PDB entries. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Klebe@staff.uni-marburg.de.

REFERENCES

(1) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct., Bioinf.* **1990**, *8*, 195–202.

(2) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293–304.

(3) 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.

(4) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411–428.

(5) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: a new fast, exhaustive flexible ligand docking system. *J. Mol. Graphics Modell.* **2007**, *26*, 198–212.

(6) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.

(7) 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.

(8) 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.

(9) 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.

(10) 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.

(11) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similaritybased search engine. *J. Med. Chem.* **2003**, *46*, 499–511.

(12) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151–166.

(13) Dixon, J. S. Evaluation of the CASP2 docking section. *Proteins: Struct., Funct., Bioinf.* **1997**, *Suppl 1*, 198–204.

(14) Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein-ligand interactions. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 521–533.

(15) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225–242.

(16) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(17) Plewczynski, D.; Łázniewski, M.; Augustyniak, R.; Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **2011**, *32*, 742–755.

(18) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079–1093.

(19) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395–2417.

(20) Jorgensen, W. L. Free energy calculations: A breakthrough for modeling organic chemistry in solution. *Acc. Chem. Res.* **1989**, *22*, 184–189.

(21) Massova, I.; Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discovery Des.* **2000**, *18*, 113–135.

(22) Tang, Y. T.; Marshall, G. R. PHOENIX: A Scoring Function for Affinity Prediction Derived Using High-Resolution Crystal Structures and Calorimetry Measurements. *J. Chem. Inf. Model.* **2011**, *51*, 214–228.

(23) 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.

(24) 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.

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

(26) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol. (Cambridge, MA, U. S.)* **1995**, *2*, 317–324.

(27) 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.

(28) 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.

(29) Böhm, 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.

(30) Sotriffer, C. A.; Sanschagrin, P.; Matter, H.; Klebe, G. SFC-score: scoring functions for affinity prediction of protein-ligand complexes. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 395–419.

(31) Huang, S. Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1866–1875.

(32) Huang, S. Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876–1882.

(33) Huang, S. Y.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 262–273.

(34) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.* **2005**, *48*, 2325–2335.

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

(36) Velec, H. F.; 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.

(37) Pfeffer, P.; Gohlke, H. DrugScoreRNA—knowledge-based scoring function to predict RNAligand interactions. *J. Chem. Inf. Model.* **2007**, *47*, 1868–1876.

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

(39) Xue, M.; Zheng, M.; Xiong, B.; Li, Y.; Jiang, H.; Shen, J. Knowledge-based scoring functions in drug design. 1. Developing a target-specific method for kinase-ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 1378–1386.

(40) Shen, Q.; Xiong, B.; Zheng, M.; Luo, X.; Luo, C.; Liu, X.; Du, Y.; Li, J.; Zhu, W.; Shen, J.; Jiang, H. Knowledge-based scoring functions in drug design: 2. Can the knowledge base be enriched? *J. Chem. Inf. Model.* **2011**, *51*, 386–397.

(41) 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.

(42) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902.

(43) Xie, Z. R.; Hwang, M. J. An interaction-motif-based scoring function for protein-ligand docking. *BMC Bioinf.* **2010**, *11*, 298.

(44) Ben-Naim, A. *Statistical Thermodynamics for Chemists and Biochemists*; Plenum Press: New York, 1992; Chapters 5 and 6.

(45) Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859–883.

(46) Sippl, M. J. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 473–501.

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

(48) Jernigan, R. L.; Bahar, I. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **1996**, *6*, 195–209.

(49) Ben-Naim, A. Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys.* **1997**, *107*, 3698–3706.

(50) Koppensteiner, W. A.; Sippl, M. J. Knowledge-based potentials—back to the roots. *Biochemistry (Moscow)* **1998**, *63*, 247–252.

(51) Hamelryck, T.; Borg, M.; Paluszewski, M.; Paulsen, J.; Frellsen, J.; Andreetta, C.; Boomsma, W.; Bottaro, S.; Ferkinghoff-Borg, J. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS One* **2010**, *5*, e13714.

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

(53) Ruvinsky, A. M.; Kozintsev, A. V. The key role of atom types, reference states, and interaction cutoff radii in the knowledge-based method: new variational approach. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 845–851.

(54) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.

(55) Berman, H. M. The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899–907.

(56) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 389–397.

(57) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes: The Art of Scientific Computing*, 3rd ed.; Cambridge University Press: New York, NY, 2007; pp 507—514.

(58) *The PyMOL Molecular Graphics System*, version 1.2r2, Schrödinger, LLC. http://www.schrodinger.com.

(59) Block, P.; Weskamp, N.; Wolf, A.; Klebe, G. Strategies to search and design stabilizers of protein-protein interactions: A feasibility study. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 170–186.

(60) 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.

(61) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.