GeauxDock: A Novel Approach for Mixed-Resolution Ligand Docking Using a Descriptor-Based Force Field

Yun Ding,[a] Ye Fang,[b,c] Wei P. Feinstein,[d] Jagannathan Ramanujam,[b,c] David M. Koppelman,[b] Juana Moreno,[a,c] Michal Brylinski,*[c,d] and Mark Jarrell*[a,c]

Molecular docking is an important component of computer-aided drug discovery. In this communication, we describe GeauxDock, a new docking approach that builds on the ideas of ligand homology modeling. GeauxDock features a descriptor-based scoring function integrating evolutionary constraints with physics-based energy terms, a mixed-resolution molecular representation of protein-ligand complexes, and an efficient Monte Carlo sampling protocol. To drive docking simulations toward experimental conformations, the scoring function was carefully optimized to produce a correlation between the total pseudoenergy and the native-likeness of binding poses. Indeed, benchmarking calculations demonstrate that GeauxDock has a strong capacity to identify near-native conformations across docking trajectories with the area under receiver operating characteristics of 0.85. By excluding closely related templates, we show that GeauxDock maintains its accuracy at lower levels of homology through the increased contribution from physics-based energy terms compensating for weak evolutionary constraints. GeauxDock is available at http://www.institute.loni.org/lasigma/package/dock/. © 2015 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24031

Introduction

Computational identification of potential leads against a specific protein target is of paramount importance to modern drug design. As of April 2015, the ZINC database of commercially available small molecule entities for drug discovery contains 17,900,742 drug-like compounds collected from the catalogs of 236 vendors.[1] At the outset of drug development, this vast number of compounds must be downsized to typically hundreds to thousands of the most promising candidate molecules. High-throughput screening (HTS) often adopted by the pharmaceutical industry is a conventional approach for lead identification, however, it suffers from high costs and low hit rates. Conversely, computational methods such as virtual screening (VS) provide faster and cheaper alternatives to HTS with many successful examples described in the literature.[2–4] Current VS techniques fall into two main categories: ligand-based similarity searching and structure-based molecular docking.[5] Although the experimentally solved structures of target proteins are not required in the ligand-based approach, an initial set of already developed compounds must be known. However, this information is often unavailable, particularly for novel molecular targets. In contrast, the advances in X-ray crystallography and nuclear magnetic resonance result in the accumulation of atomic-level structures of biological molecules fostering docking-based drug discovery projects.[6,7]

A typical molecular docking program incorporates two important components, the prediction of the binding mode of a drug candidate within the target pocket and the estimation of binding affinity from molecular interactions. Most currently available docking approaches implement effective algorithms to predict near-native binding modes,[8–11] however, noticeable differences still exist when compared with the experimental data. For instance, a recent study evaluated seven popular docking programs on a dataset of 1300 complexes showing a wide range of the average root-mean-square-deviation (RMSD) values from 2.7 Å up to 4.5 Å.[12] In addition to binding mode prediction, a scoring function is another pivotal component of molecular docking that guides the exploration of the conformational space and estimates the binding affinity for putative binding modes. Many scoring functions developed to date[13–18] can be broadly categorized into three classes, force field-based, empirical, and knowledge-based.[19–21] Recently, Liu and Wang proposed a new type of scoring function called descriptor-based or machine learning-based to capture the new trend in this field.[22] Methods using descriptor-based scoring functions encode the properties of ligands and

[a] Y. Ding, J. Moreno, M. Jarrell
Department of Physics and Astronomy, Louisiana State University, Baton Rouge, Louisiana 70803
[b] Y. Fang, J. Ramanujam, D. M. Koppelman
School of Electrical Engineering and Computer Science, Louisiana State University, Baton Rouge, Louisiana 70803
[c] Y. Fang, J. Ramanujam, J. Moreno, M. Brylinski, M. Jarrell
Center for Computation & Technology, Louisiana State University, Baton Rouge, Louisiana 70803
E-mail: jarrellphysics@gmail.com; E-mail: michal@brylinski.org
[d] W. P. Feinstein, M. Brylinski
Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803
Contract grant sponsor: National Science Foundation Under the NSF EPSCoR Cooperative Agreement; Contract grant number: EPS-1003897; Contract grant sponsor: Louisiana Board of Regents Through the Board of Regents Support Fund; Contract grant number: LEQSF(2012-15)-RD-A-05
© 2015 Wiley Periodicals, Inc.

proteins as well as protein-ligand interactions into sets of descriptors followed by applying machine learning to compute protein-ligand binding scores.[22] Notwithstanding the progress in the development of scoring functions for ligand docking, several comparative studies reported that no single algorithm systematically outperforms other methods across all protein targets.[8,23,24]

In general, high-resolution protein structures are required for satisfactory results from molecular docking regardless of which scoring function is used.[25] Additionally, the prediction success rate drops from the ligand-bound to ligand-free conformational state of a target protein.[26] This is due to the fact that many proteins undergo structural changes in functionally relevant regions on ligand binding.[27] It has been demonstrated that even minor changes affect the docking accuracy; for example, the mean protein rearrangement greater than 1.5 Å may cause a loss of 90% of the initial docking accuracy.[28] Although high-resolution structures are usually preferred in docking simulations, these may not be available in the near future for many pharmacologically important drug targets such as membrane spanning G-protein coupled receptors and ion channels.[29] Conversely, Skolnick et al. pointed out that high-resolution structures may actually conceal the inherent structural plasticity of ligand binding regions.[30] For instance, the structural variation of a highly conserved ATP-binding site is about 2.4 Å, as measured over a subset of inhibitor-bound crystal structures of protein kinases.[31] To address this issue, a recently developed ligand homology modeling (LHM) approach[32] integrates structural information extracted from evolutionarily related proteins into the modeling of protein-ligand interactions to improve the tolerance to distortions in target binding sites. LHM was one of the first approaches to successfully incorporate evolutionary information in ligand docking and VS.[30,33] Q-Dock^{LHM}[34] further exploited the ideas of LHM by implementing a descriptor-based scoring function. Nevertheless, an open question is how evolutionary descriptors supplement physics-based components in a force field that combines these two classes of scoring terms.

In this study, we describe the development and benchmarking of GeauxDock, a new approach for ligand molecular docking. GeauxDock uses a descriptor-based scoring function integrating evolutionary constraints with statistical potentials and physics-based energy terms. Moreover, it features a mixed-resolution molecular representation of protein-ligand complex structures at the level of ligand heavy atoms and protein effective points. A Monte Carlo protocol is used to efficiently sample the conformational space with the flexibility of ligand and receptor molecules modeled using an ensemble-based approach. The scoring function in GeauxDock was parameterized on a large dataset of protein-ligand complexes and further optimized to produce a correlation between the total pseudoenergy and the native-likeness of binding poses. Finally, we carry out an analysis of the contribution of various scoring terms to the identification of final docking conformations. We demonstrate that although evolutionary constraints generally improve the docking accuracy, the scarcity of this information can be effectively compensated by increasing the contribution from physics-based energy components.

Materials and Methods

Datasets

Two datasets of protein-ligand complexes are used in this study. The first set was compiled from the *e*FindSite library[35] by clustering template proteins at 40% sequence identity using PISCES[36] and then selecting representative chains that noncovalently bind small organic molecules at distinct locations. This procedure resulted in 14,059 nonredundant structures of protein-ligand complexes, referred to as the *e*FindSite/Protein Data Bank (PDB)[37] dataset, that are used to derive potentials and parameters for the docking force field. The second dataset comprises 201 high-quality crystal structures taken from the Astex/CCDC collection of pharmacologically relevant drug targets complexed with ligand molecules.[38] As our force field includes potentials calculated from evolutionarily related binding pockets, we selected those proteins for which *e*FindSite predicted the binding site within a distance of 8 Å from the geometric center of a ligand in the experimental complex structure. *e*FindSite is a threading/structure-based method that detects conserved binding sites across sets of homologous proteins.[35] For each target, we ran *e*FindSite at two different thresholds for the maximum target-template sequence identity, 80 and 40%. The first protocol uses both close and remote homologs to detect functional sites, whereas the second uses only those templates that are evolutionarily weakly related to the target. The Astex/CCDC dataset is used for the force field optimization and benchmarking.

Molecular representation of complex structures

Docking systems are described using a mixed-resolution molecular representation; heavy atoms are used for ligands, whereas proteins are represented at the coarse-grained subresidual level.[39] The following SYBYL chemical types[40] are used for ligand atoms: carbon (C.1, C.2, C.3, C.ar, and C.cat), nitrogen (N.1, N.2, N.3, N.4, N.am, N.ar, and N.pl3), oxygen (O.2, O.3, and O.co2), phosphorous (P.3), sulfur (S.2, S.3, S.O, and S.O2), and halogens (Br, Cl, F, I). For proteins, two effective backbone points per residue are placed at the position of its Cα atom and the geometrical center of the peptide plane (PP). Small side chains of Ala, Asn, Asp, Cys, Ile, Leu, Pro, Ser, Thr, and Val are reduced to one pseudo atom located at the geometric center (e.g., Ala-1, Asn-1, etc.), whereas longer side chains of Arg, Gln, Glu, His, Lys, Met, Phe, Trp, and Tyr are described by two effective points corresponding to the middle of a virtual bond between Cβ and Cγ atoms (e.g., Arg-1, Gln-1, etc.), and the geometric center of the remaining side-chain atoms (e.g., Arg-2, Gln-2, etc.). Such a low-resolution description of proteins has been shown to improve the tolerance to deformations in the target structures, while maintaining reasonable details of the physicochemical features of amino acids.[34]

Protein-ligand contacts

Intermolecular contacts between ligand atoms and protein effective points are calculated using type-dependent distance thresholds, D_{lp}^{cnt}, that accurately reproduce high-resolution interatomic contacts.[41] In addition to several contact-based components of the docking force field, mixed-resolution protein-ligand contacts are used to quantify the similarity between binding modes. Specifically, we use a Contact Mode Score (CMS) that calculates Matthew's correlation coefficient between two sets of intermolecular contacts derived from a pair of ligand binding modes for a given protein-ligand system:

$$\mathrm{CMS}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

where TP is the number of true positives, that is, interatomic contacts that are correctly predicted, and TN is the number of true negatives, that is, those pairs of ligand atoms and protein effective points that are correctly predicted not to be in contact. FP and FN are the numbers of false positives and false negatives, respectively, that is, those contacts that are overpredicted and underpredicted. Theoretically, CMS ranges from -1 to 1 with higher values indicating a better overlap between the two sets of contacts. In practice, because ligand conformations are confined to the vicinity of a protein pocket, CMS varies from about 0 up to 1. It also has certain advantages over other similarity measures. In contrast to the RMSD, CMS is fairly independent of the ligand size, therefore, it provides a more intuitive metric for the analysis of datasets comprising different compounds and target proteins. Finally, compared with the fraction of overlapping contacts, CMS penalizes those contacts that are overpredicted and underpredicted in docking models.

Force field for molecular docking

Protein-ligand complexes are stabilized by a variety of molecular interactions. Here, we developed a new descriptor-based force field for the modeling of protein-ligand interactions that combines classical physics-based potentials with statistical and knowledge-based scoring terms. Specifically, we include the following nine energy terms: (i) electrostatic and (ii) van der Waals interactions, (iii) hydrogen bonds, (iv) hydrophobic interactions, (v) generic and (vi) pocket-specific contact potentials, (vii) a pseudopharmacophore potential, and position restraints on (viii) family conserved anchor substructures, and (ix) the binding site center.

Electrostatic and van der Waals interactions (i, ii). Because of the mixed-resolution model, we use soft electrostatic, P_{ele}^{soft}, and soft Lennard-Jones, P_{vdW}^{soft}, potentials.[42] Electrostatic interactions are described by:

$$P_{\mathrm{ele}}^{\mathrm{soft}}(l,p)=q_l q_p g(r_{lp}) \quad (2)$$

Let r_{lp} be the distance between the lth ligand atom and the pth protein effective point with the corresponding partial charges q_l and q_p. Then $g(r_{lp})=1/R_{lp}$ for $R_{lp}\geq 1$, and $g(r_{lp})=k+aR_{lp}^2+bR_{lp}^3$ for $R_{lp}<1$, where $R_{lp}=sr_{lp}$, $a=4-3k$ and $b=2k-3$. k is an adjustable parameter that controls the value of the electrostatic potential at zero separation and it is set to 2.0, and s is a scaling factor set to 0.5. Partial charges on ligand atoms are calculated using the Mulliken population analysis[43] implemented in Open Babel,[44] whereas those on protein effective points are assigned by adding partial charges from the constituent atoms according to the Assisted Model Building with Energy Refinement (AMBER) ff03ua force field.[45]

The electrostatic interaction score, E_{ele}^{soft}, is a sum of P_{ele}^{soft} values taken over $L\times P$ pairs of ligand atoms and protein effective points normalized by the total number of ligand atoms, L:

$$E_{\mathrm{ele}}^{\mathrm{soft}}=\frac{1}{L}\sum_{l}^{L}\sum_{p}^{P}P_{\mathrm{ele}}^{\mathrm{soft}}(l,p) \quad (3)$$

Van der Waals interactions are modeled using the following form of a soft Lennard-Jones potential:

$$P_{\mathrm{vdW}}^{\mathrm{soft}}(l,p)=\frac{\left(2\varepsilon_{lp}r_{lp}^{*9}/r_{lp}^{9}\right)-\left(3\varepsilon_{lp}r_{lp}^{*6}/r_{lp}^{6}\right)}{\left(2\varepsilon_{lp}r_{lp}^{*9}/r_{lp}^{9}\right)\alpha\left(1+\beta r_{lp}^{2}\right)+1} \quad (4)$$

where r_{lp}^{*} depends on both a ligand atom type and the amino acid effective point and it is defined as $r_{lp}^{*}=\kappa D_{lp}^{cnt}$. ε is the depth of the potential well, and r_{lp} is the distance between the lth ligand atom and the pth protein point. The parameter α controls the value of the function at $r_{lp}=0$, and the parameter β controls the rate at which the function approaches the maximum value at zero separation.

Type-dependent parameters ε are derived from the *e*FindSite/PDB dataset as follows:

$$\varepsilon_{lp}=\ln\left(1+\frac{n_{lp}}{n_{lp}^{0}}\right) \quad (5)$$

where n_{lp} is the observed number of contacts between a given pair of a ligand atom type and the amino acid effective point, and n_{lp}^{0} is an expected number of contacts assuming no specificity. The latter is defined as $n_{lp}^{0}=N\chi_l\chi_p$, with the total number of N protein-ligand contacts, and χ_l and χ_p corresponding to the mole fractions of ligand atoms of type l and protein points of type p, respectively.

Parameters α, β, and κ are optimized empirically on the *e*FindSite/PDB dataset by minimizing the following Z-score function:

$$Z_{\mathrm{vdW}}=\sum_{l,p}^{D}\frac{P_{\mathrm{vdW}}^{\mathrm{nat}}(l,p)-\langle P_{\mathrm{vdW}}^{\mathrm{dec}}(l,p)\rangle}{\delta} \quad (6)$$

where the summation runs over D pairs of ligand atoms and protein points that are in contact according to the mixed-resolution models of dataset complexes. P_{vdW}^{nat} is the value of the soft Lennard-Jones potential, P_{vdW}^{soft}, for a given pair of the lth ligand atom and the pth protein point. $\langle P_{vdW}^{dec}(l,p)\rangle$ is the value of P_{vdW}^{soft} averaged over a set of 10 "decoy" distances r_{lp}

randomly generated around the interaction threshold D_{lp}^{cnt}, and δ is the corresponding standard deviation. The optimal values of α=0.88, β=0.74, and κ=0.70 were found using the evolutionary search strategy.[46]

For a given protein-ligand complex, the van der Waals interaction score, E_{vdW}^{soft}, is calculated by summing P_{vdW}^{soft} values over all ligand atoms and protein effective points, and then normalizing the sum by the total number of ligand atoms L:

$$E_{\mathrm{vdW}}^{\mathrm{soft}}=\frac{1}{L}\sum_{l}^{L}\sum_{p}^{P}P_{\mathrm{vdW}}^{\mathrm{soft}}(l,p) \tag{7}$$

Hydrogen bonds (iii). The hydrogen bond potential, P_{HB}, only applies to those atom pairs that can form hydrogen bonds and it is modeled using single Gaussian restraints:

$$P_{\mathrm{HB}}(l,p)=-\exp\left\{-0.5\left(\frac{r_{lp}^{\mathrm{HB}}-\mu_{lp}^{\mathrm{HB}}}{\sigma_{lp}^{\mathrm{HB}}}\right)^2\right\}\frac{1}{\sqrt{2\pi\sigma_{lp}^{\mathrm{HB}}}} \tag{8}$$

where r_{lp}^{HB} is the distance between the lth ligand atom and the pth protein effective point, and μ_{lp}^{HB} is the average hydrogen bond length between ligand atoms of the same type as l and protein points of the same type as p across the *e*FindSite/PDB dataset, with the corresponding standard deviation σ_{lp}^{HB}.

For a given protein-ligand complex, its hydrogen bond score is calculated by summing P_{HB} over those pairs of ligand atoms and protein effective points that can form hydrogen bonds, and then averaging by the total number of ligand atoms L:

$$E_{\mathrm{HB}}=\frac{1}{L}\sum_{l}^{L}\sum_{p}^{P}\begin{cases}P_{\mathrm{HB}}(l,p), & \text{if } (l,p) \text{ can form a hydrogen bond}\\ 0, & \text{else}\end{cases} \tag{9}$$

Hydrophobic interactions (iv). Hydrophobic interactions between ligand atoms and protein effective points are modeled using a spatial hydrophobicity distribution and softened Gaussian restraints. First, we calculate an empirical hydrophobicity, $P_{HP}(l)$, at the position of a ligand atom l resulting from the surrounding P protein side chains within a distance of $r_{\max}$ using a simple sigmoid function[47]:

$$P_{\mathrm{HP}}(l)=\sum_{p}^{P}\begin{cases}\tilde{H_p}\left[1-\frac{1}{2}\left(7k_{lp}^2-9k_{lp}^4+5k_{lp}^6-k_{lp}^8\right)\right], & \text{if } r_{lp}\le r_{\max}\\ 0, & \text{else}\end{cases} \tag{10}$$

where r_{lp} is the distance between the lth ligand atom and the pth protein effective point, $r_{\max}$ has a fixed value of 9 Å,[47] and $k_{lp}=r_{lp}/r_{\max}$. $\tilde{H_p}$ is the hydrophobicity parameter for the pth protein effective point according to a scale derived for amino acids in globular proteins from crystallographic data.[48]

Next, we calculate a natural logarithm of the common Gaussian restraint with the average hydrophobicity μ_l^{HP} and the corresponding standard deviation σ_l^{HP}:

$$P_{\mathrm{HP}}^{\mathrm{rst}}(l)=\frac{1}{2}\left(\frac{P_{\mathrm{HP}}(l)-\mu_l^{\mathrm{HP}}}{\sigma_l^{\mathrm{HP}}}\right)^2-\ln\left(\frac{1}{\sigma_l^{\mathrm{HP}}\sqrt{2\pi}}\right) \tag{11}$$

Ligand type-dependent parameters μ_l^{HP} and σ_l^{HP} are derived from the *e*FindSite/PDB dataset by calculating the average empirical hydrophobicity, $P_{HP}(l)$, and the corresponding standard deviation for different ligand atom types.

The hydrophobic interaction score, E_{HP}, is taken as the average P_{HP}^{rest} calculated over all ligand atoms, L:

$$E_{\mathrm{HP}}=\frac{1}{L}\sum_{l}^{L}P_{\mathrm{HP}}^{\mathrm{rest}}(l) \tag{12}$$

Generic and pocket-specific contact potentials (v, vi). The molecular docking force field implemented in GeauxDock also includes generic and pocket-specific contact potentials. The generic potential, P_{CP}, between the lth ligand atom and the pth protein effective point is calculated as follows:

$$P_{\mathrm{CP}}(l,p)=S(r_{lp})\left(-\ln\frac{n_{lp}}{n_{lp}^0}\right) \tag{13}$$

where n_{lp} is the observed number of contacts between ligand atoms of a similar type as l and protein effective points of a similar type as p across the *e*FindSite/PDB dataset, and n_{lp}^0 is a reference number of contacts assuming no specificity [explained in eq. (5)]. $S(r_{lp})$ is a smoothing function defined as:

$$S(r_{lp})=\frac{1}{1+\exp\left[\left(6-\frac{r_{lp}}{2}\right)\left(r_{lp}-D_{lp}^{\mathrm{cnt}}\right)\right]} \tag{14}$$

where r_{lp} is the distance between l and p, and D_{lp}^{cnt} is the contact threshold that depends on the types of both l and p.

The generic contact score, E_{CP}, is calculated by summing P_{CP} values over all pairs of ligand atoms and protein effective points, and then averaging over the total number of ligand atoms, L:

$$E_{\mathrm{CP}}=\frac{1}{L}\sum_{l}^{L}\sum_{p}^{P}P_{\mathrm{CP}}(l,p) \tag{15}$$

In addition to the generic potential P_{CP} derived from the *e*FindSite/PDB dataset, we calculate P_{CP}^{PS}, a pocket-specific (PS) contact potential.[34] The PS version uses the same formalism as the generic potential, however, rather than derived from the *e*FindSite/PDB, the numbers of contacts n_{lp} and n_{lp}^0 are calculated using a set of evolutionarily related complex structures identified for a given target protein by *e*Thread[49] and *e*FindSite.[35] Similar to E_{CP}, the pocket-specific contact score, E_{CP}^{PS}, is calculated as:

$$E_{\mathrm{CP}}^{\mathrm{PS}}=\frac{1}{L}\sum_{l}^{L}\sum_{p}^{P}P_{\mathrm{CP}}^{\mathrm{PS}}(l,p) \tag{16}$$

Family conserved anchor substructures and pseudopharmacophore potential (vii, viii). Ligands extracted from evolutionarily related complex structures are also used to impose a series of

harmonic restraints on family conserved anchor substructures, which were shown to be highly effective in ligand docking,[50] and to construct a new pseudopharmacophore model. The former performs the chemical matching of a target ligand against all template ligands using kcombu[51] to identify the maximum common substructures (MCS). Subsequently, atomic equivalences within MCS provided by kcombu are used to calculate a weighted average for RMSD values obtained against a set of A template ligands, with weights corresponding to the target-template chemical similarity measured by the Tanimoto coefficient.[52] A position restraint, P_{MCS}, imposed on the ath anchor substructure, which is essentially an MCS detected by kcombu, is calculated as:

$$P_{\mathrm{MCS}}(a)=\sqrt{\frac{1}{E}\sum_{e}^{E}\left(r_a^e\right)^2} \tag{17}$$

where the summation runs over E pairs of equivalent atoms in the target and template ligands sharing the ath anchor substructure, and r_a^e is the atomic distance for the eth pair.

Typically, multiple templates and the corresponding anchor substructures are detected for a given protein-ligand target, therefore, the final score taking into account family conserved anchor substructures, E_{MCS}, is calculated as the natural logarithm of the weighted average of individual P_{MCS} values:

$$E_{\mathrm{MCS}}=\ln\left(\frac{1}{A}\sum_{a}^{A}TC_aP_{\mathrm{MCS}}(a)\right) \tag{18}$$

where TC_a is the Tanomoto coefficient corresponding to the chemical similarity between the ath template ligand and the target molecule, and A is the total number of templates used to extract the anchor substructures.

The second energy term in this group uses a pseudopharmacophore potential. Specifically, it applies a Kernel Density Estimation (KDE) method to the positions of heavy atoms of template ligands bound to the identified homologs to estimate a probability density function. We use a standard normal density function to describe the likelihood of an atom of the docking ligand to be at a certain position within the binding site; the following is the scaled form of the kernel, K_h:

$$K_h(l,e)=\frac{1}{(2\pi h)^{3/2}}\exp\left(-\frac{(x_l\text{-}x_e)^2+(y_l-y_e)^2+(z_l-z_e)^2}{2h^2}\right) \tag{19}$$

where h is the bandwidth with a value of 0.5, l is a target ligand atom, and e is a template ligand atom (l and e are of the same type). KDE provides a convenient way of data smoothing, where inferences about the population are made based on a finite data sample.[53,54]

The pseudopharmacophore potential on the lth ligand atom is then calculated as:

$$P_{\mathrm{PHR}}(l)=\frac{1}{E}\sum_{e}^{E}\begin{cases}K_h(l,e), & \text{if } type(e)=type(l)\\ 0, & \text{else}\end{cases} \tag{20}$$

where E is the total number of template ligands.

The pseudopharmacophore score for a given configuration of a ligand within the binding site of the target protein is calculated as the average P_{PHR} over all ligand atoms, L:

$$E_{\mathrm{PHR}}=\frac{1}{L}\sum_{l}^{L}P_{\mathrm{PHR}}(l) \tag{21}$$

Distance restraint (ix). Finally, to limit the search space to the vicinity of a binding site, the following distance constraint is imposed:

$$E_{\mathrm{DST}}=r_{\mathrm{cen}} \tag{22}$$

where r_{cen} is the distance between the ligand geometric center and the binding pocket center predicted by *e*FindSite.[35]

Ensemble docking

The flexibility of ligands and proteins in molecular docking is implemented using an ensemble-based approach. This commonly used technique first precalculates an ensemble of low-energy conformations and then performs a rigid-body docking for each conformer.[55,56] For ligands, we used a procedure described previously[50] to generate nonredundant ensembles comprising up to 50 low-energy conformations whose pairwise RMSD is >1 Å. Protein ensembles were constructed using Modeller[57] based on the experimental structure of each target (self-modeling). We used only the coordinates of Cα atoms belonging to nonbinding residues to fully explore the flexibility of ligand binding regions. For each target protein, 10 models were generated by Modeller through three rounds of optimization using a variable target function method and molecular dynamics refinement with the objective function set to 10^6.

Monte Carlo sampling

We use the Metropolis Monte Carlo (MMC) method[34,58] to sample the conformational space of protein-ligand interactions. This space consists of multiple subspaces representing unique combinations of protein and ligand conformations from the precalculated ensembles. In the current implementation, each subspace is sampled independently using the MMC method and the collected trajectories are merged at the end of simulations. In each single MMC step, the position and orientation of a ligand are randomly perturbed by translational and rotational steps about the x, y, and z-axis of up to 0.02 Å and 5 deg, respectively. We found that this protocol allows a ligand to freely explore the conformation space without compromising the precision. Furthermore, the temperature factor is chosen such that the average acceptance ratio is about 0.5. Note that in GeauxDock, both the perturbation scale and the temperature factor can be modified to achieve a better performance for particular systems. As MMC simulations search for the global minimum energy state of a system, a scoring function implemented in GeauxDock is optimized to assign low pseudoenergy values to near-native conformations. Consequently, native-like binding modes are frequently visited

during the conformational sampling providing a sufficient resolution of biologically relevant states.

Force field optimization

The total pseudoenergy score for a given configuration of a ligand within the binding site of its protein target is calculated as a linear combination of the individual energy terms:

$$E = w_1 E_{ele}^{soft} + w_2 E_{vdW}^{soft} + w_3 E_{HB} + w_4 E_{HP} + w_5 E_{CP} + w_6 E_{CP}^{PS} + w_7 E_{MCS} + w_8 E_{PHR} + w_9 E_{DST} \quad (23)$$

The energy weight factors, w_1–w_9, are optimized on a large and nonredundant set of protein-ligand conformations constructed for the Astex/CCDC dataset.[59] Specifically, for each complex, we first generated 10^5 configurations through a series of MMC simulations including only the Lennard-Jones potential (i) to prevent steric clashes and the distance constraint (ix) to confine the sampling to the vicinity of a binding pocket. Next, we calculated pairwise CMS values for all conformations to create a $10^5 \times 10^5$ CMS matrix. To remove redundancy, this matrix was subjected to clustering using CLUTO[60] resulting in 5000 groups; a cluster centroid was selected to represent each group. The final dataset comprises 102,000 nonredundant protein-ligand configurations constructed for 204 complexes.

Subsequently, we compiled two subsets for the force field optimization, a group of 36,022 native-like conformations whose CMS values calculated against the experimental complex structures are $\geq$0.6, and a set of 847,849 decoys with the CMS of $\leq$0.4. The native-like recognition capability of the scoring function was enhanced by finding the set of weights w_1–w_9 [see eq. (23)] that maximize the energy gap between native-like and decoy conformations measured by the *Z*-score:

$$Z\text{-score} = \frac{\langle E_d \rangle - \langle E_n \rangle}{\sigma_n^2 + \sigma_d^2} \quad (24)$$

where $\langle E_n \rangle$ and $\langle E_d \rangle$ are the mean energy values calculated for native-like and decoy conformations, respectively, with the corresponding standard deviations σ_n and σ_d.

We used the evolutionary search algorithm[46] emulating the principles of natural evolution to identify the optimal set of energy weight factors that maximize the *Z*-score. To avoid any bias, the optimization was performed 10 times starting from different initial random sets of weights; the final weight factors were taken as the consensus of the 10 optimization rounds.

Other scoring functions

Two state-of-the-art algorithms, DrugScoreX (DSX)[14] and Ligand-Protein Contacts (LPC),[61] were selected for comparative benchmarks of GeauxDock. DSX is a knowledge-based scoring function that features a distance-dependent pair potential, a torsion angle potential, and a novel solvent accessible surface-dependent potential.[14] LPC uses a scoring function that evaluates the geometric and chemical complementarity between a ligand and its receptor protein.[62] Both programs were used with their default set of parameters.

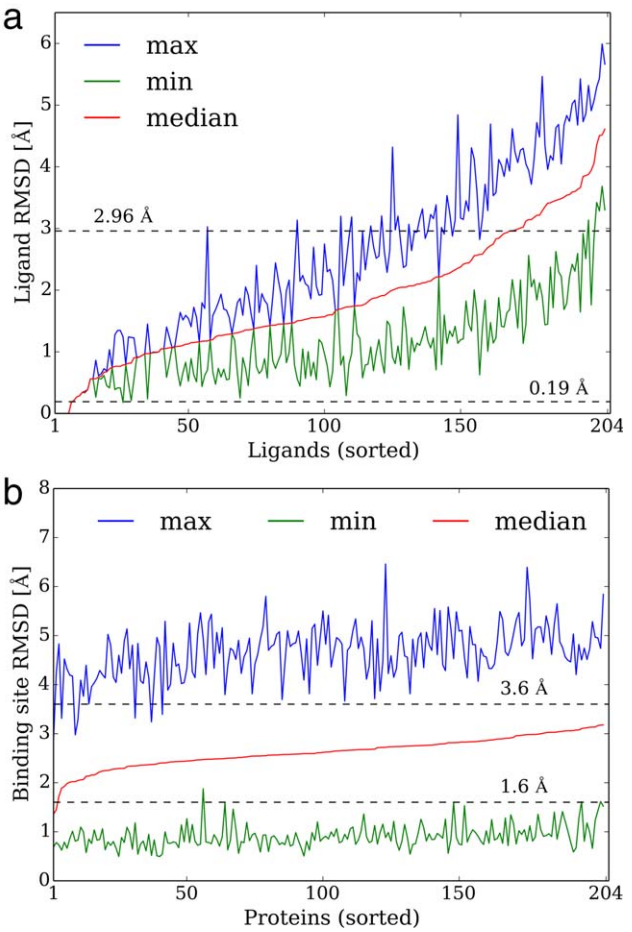

Figure 1. Structural characteristics of protein and ligand ensembles for pseudoflexible docking. All-atom RMSD values are calculated using the native conformation for a) ligands and b) protein binding sites. Dashed lines point out the estimated ranges of the molecular plasticity. Blue, green, and red lines correspond to the maximum, minimum, and median RMSD within each ensemble; molecules are sorted on the *x*-axis by their median values.

Results and Discussion

Ensembles for pseudoflexible docking

It is well known that both proteins and ligands often undergo structural changes on complex formation[27,63–65]; for instance, an analysis of 27 flexible ligands shows the RMSD variation from 0.19 to 2.96 Å[65] calculated between single-crystal and protein-bound states. A larger structural diversity is expected as the size of ligand molecules increases; for instance, the conformational range for two ubiquitous compounds, nicotinamide adenine dinucleotide and flavin adenine dinucleotide was calculated as 3.58 Å $\pm$ 0.08 and 3.49 Å $\pm$ 0.13, respectively, when bound to proteins.[64] On that account, an accurate representation of biomolecules in simulations requires sampling multiple conformational states.[66] We use an ensemble docking technique to handle this issue in a discrete manner. Specifically, conformers are selected from a precomputed pool of configurations covering a large conformational space that includes biologically relevant molecules. In that regard, we investigate the coverage of Astex/CCDC ligands by calculating

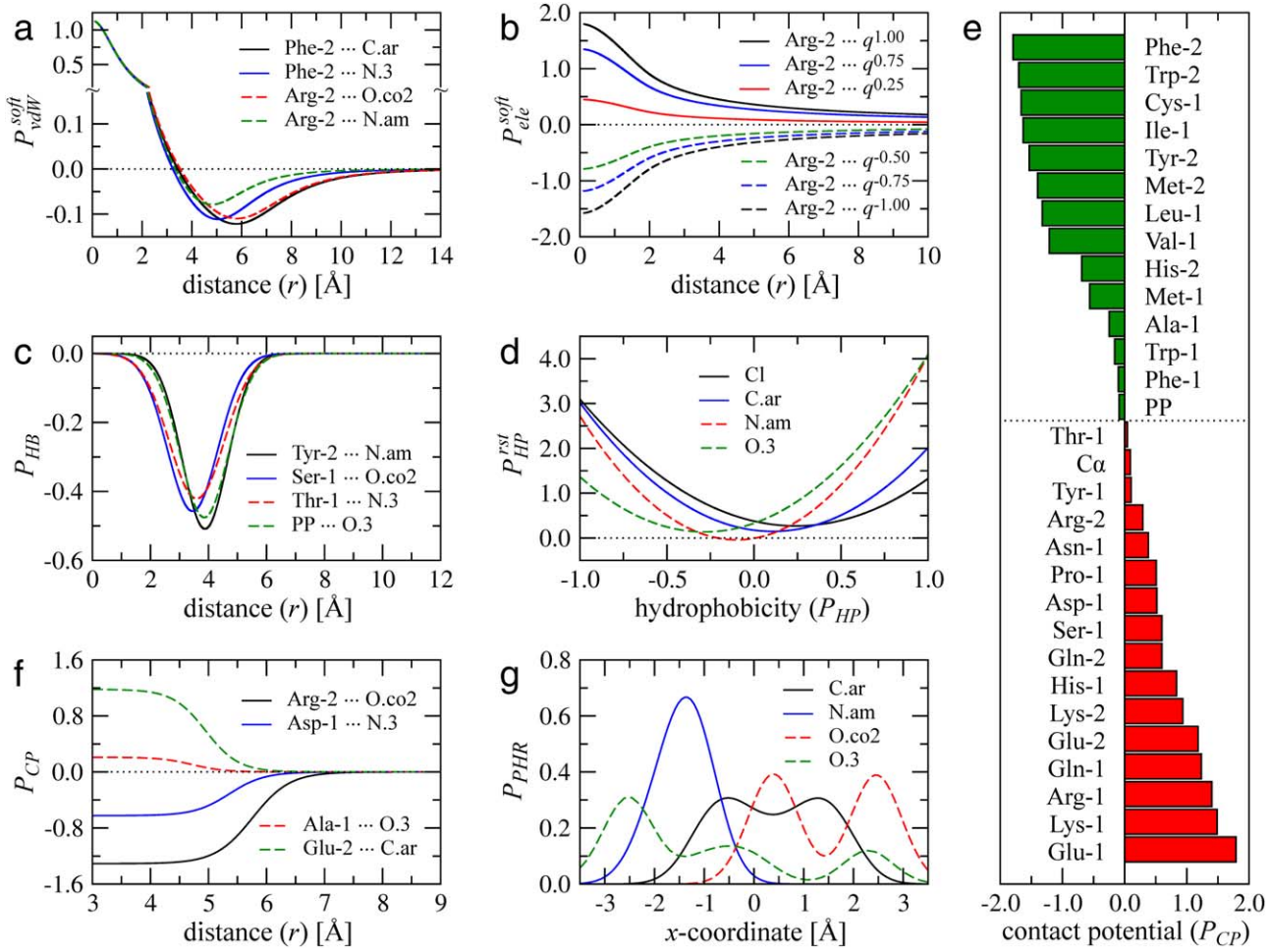

Figure 2. Examples of selected force field potentials. a) Type-dependent soft Lennard-Jones potential, b) soft electrostatic potential between protein effective points and various charges on ligand atoms *q*, c) hydrogen bond restraints, d) restraints for hydrophobic interactions between different ligand atoms as a function of local hydrophobicity, e) extreme values for the log-odds potential between aromatic carbon C.ar and protein effective points, f) generic contact potential including a smoothing function, and g) probability density for different ligand atoms estimated by KDE along the *x*-axis.

RMSD values using conformational ensembles and the corresponding experimental structures. The results in terms of maximum, minimum, and median RMSD values are presented in Figure 1. Figure 1a shows that the median RMSD for ~81% of the flexible ligands is within the reference range of 0.19 to 2.96 Å[65] suggesting that the ligand flexibility is well represented across the generated docking ensembles. Furthermore, the average plasticity of ligand-binding regions in proteins expressed as the mean RMSD was estimated as 2.6 Å with a standard deviation of 1.0 Å.[67] Protein ensembles constructed in this study fall within this range with the median binding site RMSD calculated over 204 ensembles of 2.61 Å, as shown in Figure 1b. Collectively, these results demonstrate that conformational ensembles for pseudoflexible docking provide a sufficient coverage of biologically relevant structures of both ligands and their protein targets.

Force field parameterization

Force fields for molecular modeling typically require a careful parameterization to reproduce experimental data. We derived the parameters for GeauxDock from the *e*FindSite/PDB dataset, a representative and nonredundant collection of protein-ligand complex structures. Selected force field potentials parameterized against *e*FindSite/PDB are presented in Figure 2. Figure 2a shows the soft Lennard-Jones potential used to model van der Waals interactions between effective points on Phe and Arg side chains, and selected ligand atoms. The corresponding parameters ε that define the depth of the potential well are reported in Table 1. For instance, aromatic interactions between Phe-2 and C.ar, and a salt bridge between Arg-2 and O.co2 have deeper potential wells with $\varepsilon=1.95$ and $\varepsilon=1.54$, respectively, compared to those less favorable, for example, Phe-2 and N.3 ($\varepsilon=1.07$), and Arg-2 and N.am ($\varepsilon=0.43$). Furthermore, the softened potential, which is weaker at short distances than the traditional form, is more appropriate to model interactions involving effective points representing clouds of atoms rather than the hard spheres of individual particles.

We also use a soft version of the electrostatic potential, where its values do not extend to infinity when the interaction distance *r* approaches zero. As shown in Figure 2b, the electrostatic potential creates a repulsion at close distances between

Table 1. Force field parameters for van der Waals interactions and the generic contact potential for selected ligand atom types and protein effective points.

Protein point	Parameter[a]	Ligand atom type												
		C.3	C.ar	C.cat	N.3	N.am	N.ar	O.2	O.3	O.co2	P.3	S.3	S.O2	Cl
Cα	ε	0.59	0.65	0.99	0.53	0.44	0.86	1.08	0.86	0.95	0.05	0.62	0.82	1.03
	P_{CP}	0.21	0.09	−0.47	0.37	0.59	−0.33	−0.69	−0.26	−0.47	3.01	0.15	−0.19	−0.63
PP	ε	0.66	0.79	0.87	0.64	0.56	0.61	0.83	0.58	1.02	0.05	0.84	0.80	1.05
	P_{CP}	0.10	−0.09	−0.20	0.13	0.32	0.22	−0.23	0.14	−0.55	−0.69	−0.25	−0.14	−0.57
Phe-1	ε	0.63	0.77	1.30	0.64	0.30	0.63	0.75	0.43	0.55	0.26	0.79	0.30	1.07
	P_{CP}	0.12	−0.10	−0.96	0.10	1.22	−0.10	−0.13	0.56	0.37	4.01	−0.18	0.87	−0.67
Phe-2	ε	1.77	1.95	1.58	1.07	1.34	1.95	1.46	1.29	1.60	1.17	1.66	0.82	2.19
	P_{CP}	−1.55	−1.79	−1.33	−0.69	−1.10	−1.72	−1.21	−1.00	−1.27	−0.86	−1.48	−0.26	−2.11
Arg-1	ε	0.19	0.21	0.12	0.47	0.18	0.26	0.39	0.31	0.52	0.43	0.25	0.05	0.25
	P_{CP}	1.60	1.40	2.06	0.53	1.61	1.17	0.85	1.18	0.36	0.63	1.10	3.00	1.29
Arg-2	ε	0.63	0.55	0.27	0.70	0.43	0.62	1.00	0.82	1.54	0.00	1.08	0.85	0.56
	P_{CP}	0.09	0.29	1.07	−0.01	0.63	0.14	−0.58	−0.31	−1.31	6.52	−0.80	0.04	0.29

[a] ε is the depth of the potential well in the softened Lennard-Jones potential; P_{CP} is the value of contact potential for pairwise interactions.

those groups whose partial charges have the same sign, whereas positively and negatively charged particles attract each other. The strength of these interactions depends on the partial charges on individual groups. Table 2 lists net charges assigned to protein effective points by collapsing AMBER partial charges of the constituent atoms. A point charge on the PP has a fixed value of −0.246, which balances positively charged Cα atoms of individual amino acids. Side chains of small hydrophobic residues are slightly positively charged, for example, q_p=0.047 for Ile-1, in contrast to small polar amino acids that carry small negative charges on their side chains, for example, q_p=−0.046 for Ser-1. A small negatively charged Asp has the unit charge assigned to its side chain effective point, whereas larger charged residues have almost unit charge values; for example, the parameter q_p is −0.792, 0.901, and 0.927 for Glu-2, Arg-2, and Lys-2, respectively. Partial charges on ligand heavy atoms are calculated for individual compounds using the Mulliken population analysis,[43] which is a widely used parameterization method in molecular docking.

Table 2. Partial charges on Cα and side chain (SC) effective points of amino acids.

Amino acid	Effective point		
	Cα	SC-1	SC-2
Gly	0.246	–	–
Ala	0.215	0.031	–
Asn	0.217	0.029	–
Asp	0.246	−1.000	–
Cys	0.088	0.158	–
Ile	0.199	0.047	–
Leu	0.204	0.042	–
Pro	0.112	0.119	–
Ser	0.292	−0.046	–
Thr	0.268	−0.022	–
Val	0.201	0.045	–
Arg	0.237	0.107	0.901
Glu	0.246	−0.208	−0.792
Gln	0.210	0.010	0.026
His	0.219	0.172	−0.145
Lys	0.227	0.092	0.927
Met	0.137	0.127	−0.018
Phe	0.214	0.049	−0.017
Trp	0.248	0.066	−0.068
Tyr	0.245	0.020	−0.020

Hydrogen bonds are modeled for hydrogen donor-acceptor pairs using single Gaussian restraints. Table 3 lists force field parameters for hydrogen bonds and Figure 2c shows the parameterized potential for selected pairs. Mean values for the interaction distance, μ_{lp}^{HB}, derived from the *e*FindSite/PDB dataset, give the optimal type-dependent bond lengths, whereas σ_{lp}^{HB} parameters that describe the deviation from average interaction distances across the dataset, control the interaction strength. For instance, μ_{lp}^{HB} for Thr-1 and N.3 (3.59 Å) is slightly smaller than that for Tyr-2 and N.am (3.88 Å). Moreover, the corresponding σ_{lp}^{HB} are 0.95 and 0.78, respectively, thus, the strength of hydrogen bonded pair of Tyr-2 and N.am at the optimal distance is greater than a hydrogen bond between Thr-1 and N.3.

In our model, protein residues create a polar/hydrophobic local environment favoring certain types of ligand atoms. These hydrophobic interactions are parameterized using statistics collected for *e*FindSite/PDB protein-ligand complexes and a standard hydrophobicity scale for amino acids. The derived force field parameters reported in Table 4 are in good agreement with physicochemical properties of ligand atom types. For example, aromatic carbon atoms $\left(\mu_l^{HP}=0.11\right)$ and halogens $\left(\mu_l^{HP}=0.24\right)$ tend toward nonpolar residues, whereas amine nitrogen $\left(\mu_l^{HP}=-0.27\right)$ and carboxylate oxygen $\left(\mu_l^{HP}=-0.34\right)$

Table 3. Force field parameters for hydrogen bonds, $\mu_{lp}^{HB} \pm \sigma_{lp}^{HB}$, for selected ligand types and protein effective points.

Ligand atom type	Protein effective point				
	His-2	Ser-2	Thr-1	Tyr-2	PP
N.2	3.38 ± 0.71	3.83 ± 0.83	3.91 ± 0.99	3.64 ± 0.77	3.91 ± 0.88
N.3	3.67 ± 0.71	3.80 ± 0.88	3.59 ± 0.95	3.79 ± 0.89	3.89 ± 0.92
N.am	3.80 ± 0.75	3.82 ± 0.83	3.79 ± 0.82	3.88 ± 0.78	3.62 ± 0.82
O.2	3.58 ± 0.78	3.64 ± 0.87	3.62 ± 0.86	3.75 ± 0.92	3.69 ± 0.84
O.3	3.64 ± 0.83	3.68 ± 0.86	3.72 ± 0.85	3.74 ± 0.85	3.85 ± 0.84
O.co2	3.50 ± 0.76	3.45 ± 0.87	3.64 ± 0.92	3.46 ± 0.86	3.75 ± 0.86

Table 4. Force field parameters for hydrophobic interactions, $\mu_l^{HP} \pm \sigma_l^{HP}$, assigned to selected ligand types.

Ligand atom type	$\mu_l^{HP} \pm \sigma_l^{HP}$
C.3	-0.03 ± 0.43
C.ar	0.11 ± 0.46
C.cat	-0.26 ± 0.43
N.3	-0.27 ± 0.44
N.am	-0.10 ± 0.38
N.ar	0.03 ± 0.47
O.2	-0.21 ± 0.50
O.3	-0.28 ± 0.46
O.co2	-0.34 ± 0.46
P.3	-0.50 ± 0.41
S.3	-0.14 ± 0.45
S.O2	-0.10 ± 0.40
Cl	0.24 ± 0.52

atoms prefer a polar microenvironment. Hydrophobicity restraints P_{HP}^{rest} for selected ligand atom types are shown in Figure 2d as a function of the environment created by surrounding amino acids. The extremes of −1.0 and 1.0 describe a strongly polar and nonpolar character, respectively. The position of the function minimum determines the optimal environment for a particular atom type described by P_{HP}, thus, Cl and C.ar are on the positive side, and N.am and O.3 are on the negative side of the protein hydrophobicity range.

Statistical potentials are commonly used components of molecular docking force fields.[41,68–70] In this study, the parameters for pairwise interactions between ligand heavy atoms and protein effective points were derived from the *e*FindSite/PDB dataset. The log-odds potential expresses the likelihood of individual contacts, where the interactions averaged over the entire dataset are taken as a reference state. Figure 2e shows the extreme values for the contact potential between aromatic carbon C.ar and all types of protein effective points. Clearly, aromatic moieties on the side chain effective points of Phe-2, Trp-2, and Tyr-2, as well as the hydrophobic parts of Cys-1, Ile-1, Met-2, Leu-1, and Val-1 make contacts with C.ar more often than by a random chance. In contrast, the polar and charged groups of Glu-1, Lys-1, Arg-1, Glu-1, Glu-2, and Lys-2 are statistically unlikely to interact with ligand aromatic carbon atoms. Moreover, backbone effective points Cα and PP have no distinct preferences toward interacting with C.ar.

In the GeauxDock force field, we use a smoothing function that is less sensitive to small changes in ligand coordinates at the contact distance thresholds than the commonly used step function. This is shown in Figure 2f for selected interactions between ligand heavy atoms and protein effective points. For instance, salt bridges between Arg-2 and O.co2, and Asp-1 and N.3 contribute half of their negative interaction energy at $D_{lp}^{cnt}=5.76$ Å and $D_{lp}^{cnt}=5.36$ Å, respectively. Similarly, the positive energy contributions from destabilizing interactions between Ala-1 and O.3, and Glu-2 and C.ar reach half of their values at the corresponding contact thresholds. In addition to the generic contact potential derived from the *e*FindSite/PDB dataset, we calculate its pocket-specific variant using evolutionarily related complexes identified by sequence profile-based protein threading. These potentials are specific toward a particular family of proteins, however, they contain significantly less parameters compared with the generic potential because of much smaller sample sizes (the number of template complexes). For example, out of 720 pairwise parameters derived from the *e*FindSite/PDB dataset for P_{CP}, the average number of P_{CP}^{PS} parameters calculated across the Astex/CCDC target pockets is only 110 ± 67. Nonetheless, the latter have been demonstrated to be more accurate than the generic potential in the scoring and ranking of ligand binding modes.[34]

Different from traditional pharmacophore-based models that use known bio-actives to calculate sets of steric and physicochemical features necessary for molecular recognition,[71] the pseudopharmacophore potential in GeauxDock is derived from evolutionarily ligand-bound templates. Specifically, it estimates a probability for each ligand heavy atom type to be at a certain position within the binding site. For instance, Figure 2g shows a one-dimensional probability density for C.ar, N.am, O.co2, and O.3 along the *x*-coordinate with the pocket centered at $x=y=z=0$ Å (the full potential is the product of probabilities at *x*, *y*, and *z* coordinates). In this example, amine nitrogen and hydroxyl oxygen atoms are most likely to be found at x=-1.4 Å and x=-2.5 Å, respectively. Carboxyl oxygen atoms have a bimodal distribution typical for symmetric moieties with two equivalent peaks at $x=0.4$ Å and $x=2.5$ Å, whereas aromatic carbon atoms have a relatively broad probability of occurrence at $-0.6 < x < 2.5$ Å. Favoring ligand heavy atoms at their high probability positions predicts binding modes consistent with the conserved evolutionary profiles observed across sets of homologous proteins.

Force field optimization

Force field weights were optimized on a large dataset of protein-ligand configurations generated for Astex/CCDC complexes using the evolutionary search algorithm. The objective was to maximize the *Z*-score corresponding to the energy gap between native-like and decoy conformations. Figure 3a shows the trajectory of *Z*-score in one complete optimization process. The simulation converges within ∼400 generations, indicating an efficient update of weight factors. We performed the total of 10 simulations seeded with random initial weight factors; each calculation resulted in the same set of weight factors ($w_1=18.97$, $w_2=0.78$, $w_3=2.05$, $w_4=0.53$, $w_5=0.01$, $w_6=0.53$, $w_7=0.88$, $w_8=110.82$, and $w_9=46.4$), suggesting that the optimized values are stable and robust. Figures 3b and 3c show the distribution of energy values with different sets of weights. In Figure 3b, random weight factors do not provide a clear separation between native-like (green dots) and decoy (red dots) conformations whose median energy score is −1.67 and −1.03, respectively. In contrast, Figure 3c shows that the optimized weight factors yield better energy-based partitioning of native-like and decoy conformations; here, native-like (decoy) binding modes have a median energy of −0.16 (0.58). This analysis suggests that the total pseudoenergy calculated using the optimized set of weights has a

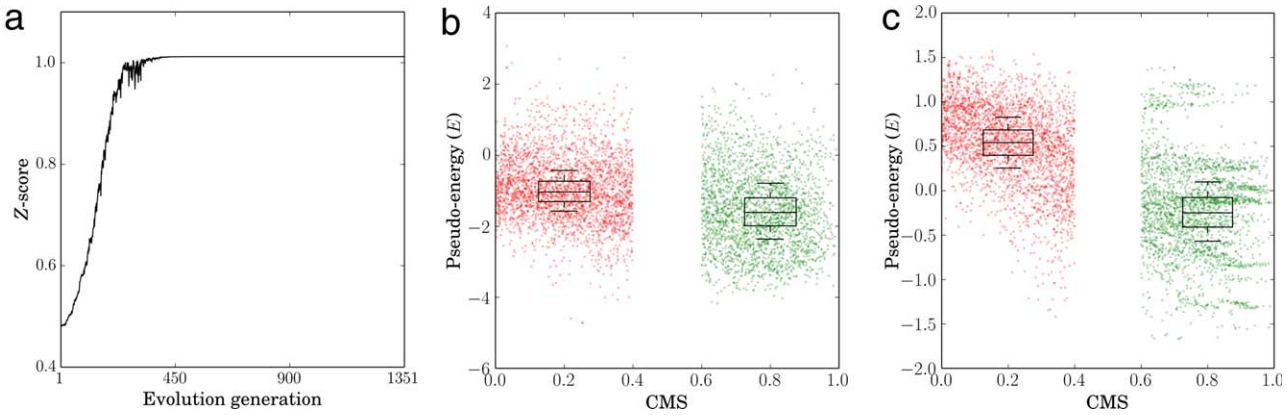

Figure 3. Force field optimization using the evolutionary algorithm. a) The trajectory of *Z*-score in the course of the optimization procedure. The distribution of pseudoenergy values for native-like (green) and decoy (red) conformations for the b) unoptimized and c) optimized force field. Boxes in b and c end at the quartiles Q_1 and Q_3, a horizontal blue line in a box is the median, and whiskers show the 1.5 interquartile range. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

great potential to effectively drive molecular docking toward correct ligand binding modes.

Recognition of native-like conformations

A strong capacity to identify native-like binding modes among a vast number of generated configurations plays a pivotal role in successful ligand docking simulations. Therefore, in Figure 4, we conduct a comparative Receiver Operating Characteristics (ROC) analysis of GeauxDock and two other scoring functions, DSX[14] and LPC.[61] Here, we use a precompiled dataset of protein-ligand configurations comprising 36,022 native-like binding poses and 847,849 decoys generated for Astex/CCDC complexes to uniformly cover the conformational space. In general, all docking algorithms are capable of identifying correct conformations across the training MMC trajectories generated for the Astex/CCDC dataset better than a purely random guess (dashed line). The area under the curve (AUC) for the unoptimized GeauxDock force field (all weight factors set to 1.0) is 0.654 in contrast to 0.851 for the optimized set of weights. For comparison, DSX_pair, DSX_pair_sas and LPC yield the AUC of 0.847, 0.858, and 0.765, respectively. Despite a slightly lower AUC, GeauxDock gives ~5% higher true positive rate than DSX_pair_sas at relatively small false positive rates of 0.1–0.2. The results for DSX consistent with the original benchmarking calculations[24] suggest that our dataset is of high quality and the CMS indeed provides an effective evaluation metric.

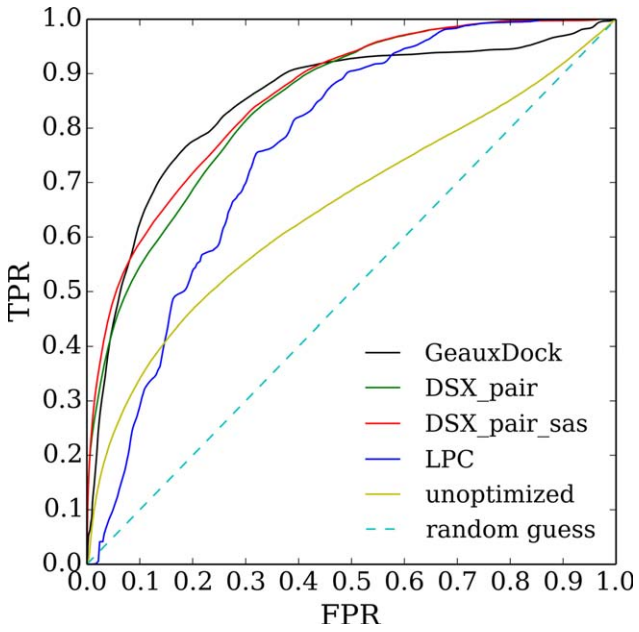

Figure 4. Recognition of native-like conformations across docking trajectories. A ROC plot for GeauxDock with an optimized force field is compared with those obtained using the unoptimized force field as well as other scoring functions, DSX and LPC. TPR – true positive rate, FPR – false positive rate. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Next, we performed full docking calculations using GeauxDock. The major difference from the previous analysis is that these validation simulations start from a random ligand conformation and use the complete, optimized force field to guide the conformational sampling. MMC trajectories generated for the Astex/CCDC dataset are analyzed in Figure 5. First, for each benchmarking complex, we calculated the *Z*-score between native-like and decoy conformations extracted from the docking trajectories. As shown in Figure 5a, ~90% of the cases have positive *Z*-score values indicating that ligand binding modes close to native are systematically assigned a lower energy than those farther away from the experimental conformation. The median *Z*-score across Astex/CCDC complexes is ~1.0, which is in accord with the training results reported in Figure 3. To further evaluate the quality of the GeauxDock force field, we calculated the Pearson correlation coefficient (PCC) between the total pseudoenergy score and CMS. Figure 5b shows that in the majority of the cases, the total pseudoenergy score and CMS are negatively correlated, that is, the energy increases with the decreasing similarity to the experimental binding mode. According to the scale provided by Salkind,[72] a very strong $(-1.0 \leq \mathrm{PCC} < -0.8)$, strong $(-0.8 \leq \mathrm{PCC} < -0.6)$, moderate $(-0.6 \leq \mathrm{PCC} < -0.4)$, weak $(-0.4 \leq \mathrm{PCC} < -0.2)$, and very weak or no correlation

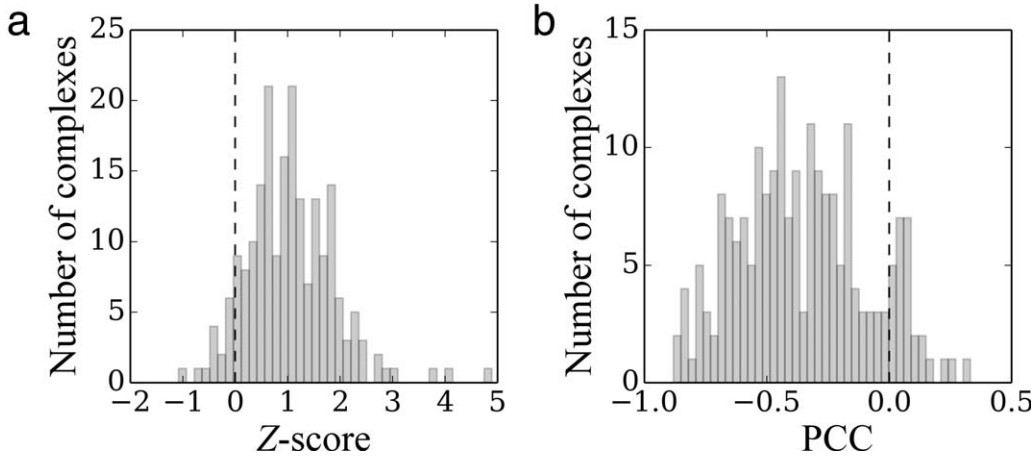

Figure 5. Quality assessment for the optimized force field implemented in GeauxDock. Histograms of a) *Z*-score and b) the PCC calculated from the Monte Carlo trajectories collected for the Astex/CCDC dataset.

$(-0.2 \leq \text{PCC} < 0.0)$ between energy and CMS was obtained for 3.43%, 15.20%, 28.43%, 25.49%, and 14.22% of the Astex/CCDC complexes, respectively; only 13.24% of the cases give the undesired positive correlation. Altogether, these results demonstrate that the scoring function in GeauxDock is correctly optimized to drive MMC simulations toward experimentally determined ligand binding modes.

Case studies

We select a couple of examples to demonstrate the accuracy of GeauxDock in finding near-native ligand binding modes, cathepsin K complexed with a peptidomimetic inhibitor (PDB-ID: 1bgo, chain A),[73] and actinidin complexed with an inhibitor E-64 (PDB-ID: 1aec, chain A).[74] Both compounds were docked into the active sites of their target protein using GeauxDock starting from a random initial conformation. The results are shown in Figure 6 (panels a–c for cathepsin K and d-f for actinidin). First, we plot the values of CMS calculated against inhibitors bound in the crystal complex structures, and the total pseudoenergy, *E*, extracted from MMC trajectories. A high pseudoenergy and a low CMS for initial configurations indicate that ligands are far away from their native states (Figures 6a and 6d). During MMC simulations, a gradually decreasing energy *E* guides the conformational sampling to the vicinity of the experimental binding modes of inhibitors as indicated by high CMS values at the end of simulations. Figures 6b and 6e demonstrate that in both cases, the optimized

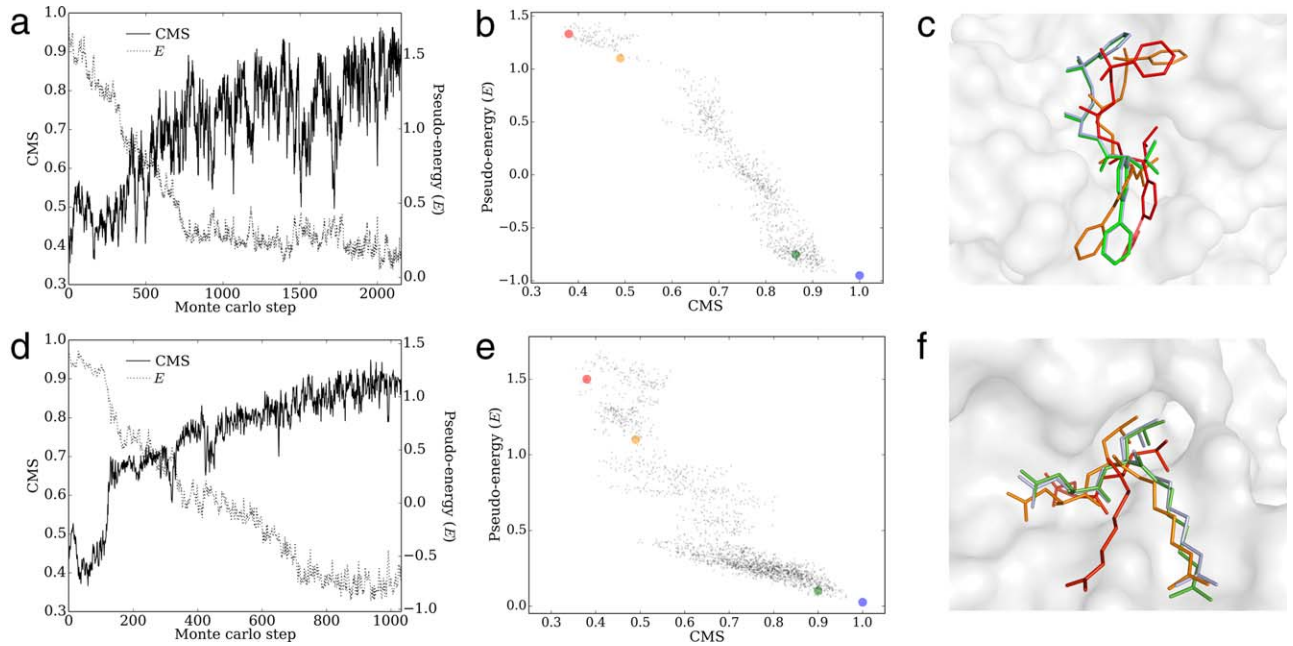

Figure 6. Docking results for a–c) cathepsin K and d–f) actinidin from GeauxDock. a, d) Monte Carlo trajectories for the Contact Mode Score (CMS) and the pseudoenergy, b, e) scatter plots of the CMS versus pseudoenergy, c, f) representative conformations taken from docking trajectories. In b, c, e, and f selected non-native, intermediate, and near-native conformations are colored in red, orange, and green, respectively, whereas the experimental binding poses are shown in ice blue. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 5. Performance of GeauxDock on the Astex/CCDC dataset assessed by the area under the curve (AUC).

	AUC	
Scoring function	40% homology	80% homology
Complete	0.831	0.848
Evolution-based	0.699	0.745
Physics-based	0.801	0.814

The force field is optimized at the homology thresholds of 40 and 80% and the performance of the complete scoring function is compared with physics- and evolution-based components.

force field yields a negative correlation between CMS and *E*, where each dot represents one MMC snapshot. Next, we select three representative conformations from those scatter plots for each inhibitor, non-native (red), intermediate (orange), and near-native (green), whose CMS are 0.38, 0.49, and 0.90 for cathepsin K, and 0.38, 0.49, and 0.86 for actinidin, respectively. The corresponding molecular representations are shown in Figures 6c and 6f using the same color scheme. In both cases, low-energy binding modes (green) significantly overlap with bound inhibitors in the experimental structures of cathepsin K and actinidin complexes (ice blue sticks), whereas non-native and intermediate conformations are characterized by notably higher pseudoenergy values.

Evolution- and physics-based components

A descriptor-based force field in GeauxDock combines evolution- and physics-based scoring terms. The former are derived from evolutionary related complex structures at two different sequence similarity thresholds, 80% to allow force field parameters to be calculated from close homologs, and 40% to use only those templates that are weakly related to their targets. Therefore, we can analyze how the level of homology affects the accuracy of molecular docking. Using the Astex/CCDC dataset, the results are reported in Table 5 as the area under the ROC curve. As expected, the AUC significantly increases when close homologs are included in force field optimization and the docking conformations are evaluated by evolution-based components alone. In contrast, the performance of physics-based scoring terms remains, to a large extent, unaffected by the maximum target-template sequence identity, because these features are calculated from physical interactions that are more universal.[75] Interestingly, the performance of GeauxDock using a complete force field at a homology threshold of 80% is only slightly better than that at 40%, suggesting that the descriptor-based scoring function is able to adapt to the supplied amount of evolutionary information to maintain its accuracy.

To further investigate this intriguing observation, we calculated the relative contribution of both classes of scoring terms to the total pseudoenergy at the two homology thresholds. Figure 7 shows that the contribution from evolution-based components to the total score is about 5% higher at 80% homology compared with 40%. Considering only a slightly better performance of GeauxDock using close homologs, this analysis suggests that the scarcity of evolutionary information can be effectively compensated by the increased contribution from physics-based scoring terms. This unique feature of GeauxDock is particularly important in its large-scale applications at the proteome level, such as in inverse VS[76,77] and rational drug repositioning,[78–80] where the availability of only weakly homologous complex structures for the majority of drug targets will not compromise the accuracy of molecular docking.

A well-balanced contribution of physics- and evolution-based energy terms to the total pseudoenergy also suggests that these two classes of scoring features effectively complement each other. Nevertheless, AUC values reported in Table 5 indicate that a linear combination of individual energy terms perhaps does not fully exploit their predictive power; for instance, adding the evolution-based component improves the AUC of physics-based terms by about 3%. This may be caused by the feature intercorrelation, which is known to limit the performance improvements despite adding more descriptors.[81] A possible solution is to combine individual energy terms using a nonlinear model, under the assumption that noncovalent interactions often depend on one another in a nonlinear manner.[82] We will explore the feasibility of a machine learning-based force field in ligand molecular docking in the near future.

Conclusions

In this study, we describe the development of GeauxDock, a molecular docking approach featuring a novel descriptor-based scoring function and a mixed-resolution description of protein-ligand complexes. The scoring function was parameterized on a large dataset of crystal structures and further optimized using sets of computer-generated native-like and decoy conformations. Encouragingly, benchmarking calculations demonstrate that GeauxDock has a strong capacity to recognize native-like binding modes with the area under ROC of 0.85. The descriptor-based scoring function implemented in GeauxDock incorporates two distinct classes of energy terms,

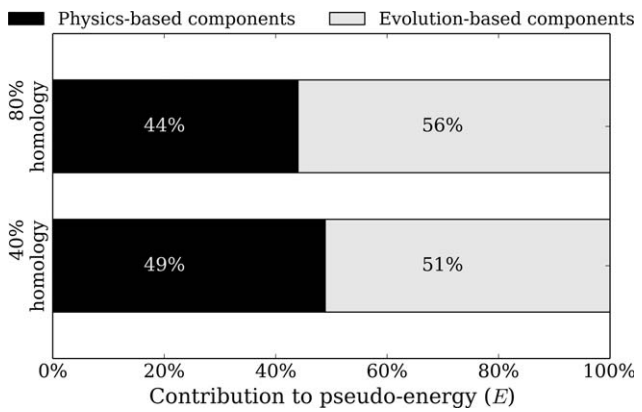

Figure 7. Balance of various energy terms in the optimized force field. The contribution from physics- and evolution-based components is calculated at the thresholds of 80 and 40% for the maximum target-template sequence identity.

physics- and evolution-based. As the latter are derived from evolutionary related complex structures, their strength depends on the level of homology between the target and template systems. Interestingly, weak evolutionary constraints are effectively compensated by the increased contribution from physics-based terms, which, in turn, help maintain the accuracy of the GeauxDock scoring function at the lower levels of protein sequence similarity. Therefore, this new ligand docking approach is well suited for proteome-scale applications taking advantage of the increasingly growing protein sequence and structural data. GeauxDock is available at http://www.institute.loni.org/lasigma/package/dock/.

Acknowledgments

The authors are grateful for discussions and comments from the members of the Technologies for Extreme Scale Computing (TESC) team formed within the Louisiana Alliance for Simulation-Guided Materials Applications (LA-SiGMA). Portions of this research were conducted with high performance computational resources provided by Louisiana State University (HPC@LSU, http://www.hpc.lsu.edu) and the Louisiana Optical Network Institute (LONI, http://www.loni.org).

Keywords: molecular docking · force field development · force filed optimization · Monte Carlo simulations · mixed-resolution modelling · descriptor-based force field

How to cite this article: Y. Ding, Y. Fang, W. P. Feinstein, R. Ramanujam, D. M. Koppelman, J. Moreno, M. Brylinski, M. Jarrell. *J. Comput. Chem.* **2015**, *36*, 2013–2026. DOI: 10.1002/jcc.24031

[1] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model*. **2005**, *45*, 177.
[2] N. J. Liverton, M. K. Holloway, J. A. McCauley, M. T. Rudd, J. W. Butcher, S. S. Carroll, J. DiMuzio, C. Fandozzi, K. F. Gilbert, S.-S. Mao, *J. Am. Chem. Soc*. **2008**, *130*, 4607.
[3] D. J. Hazuda, N. J. Anthony, R. P. Gomez, S. M. Jolly, J. S. Wai, L. Zhuang, T. E. Fisher, M. Embrey, J. P. Guare, M. S. Egbertson, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 11233.
[4] W. L. Jorgensen, *Science* **2004**, *303*, 1813.
[5] A. N. Jain, *Curr. Opin. Drug Discov. Dev*. **2004**, *7*, 396.
[6] S. Ghosh, A. Nie, J. An, Z. Huang, *Curr. Opin. Chem. Biol*. **2006**, *10*, 194.
[7] B. O. Villoutreix, R. Eudes, M. A. Miteva, *Comb. Chem. High Throughput Screen*. **2009**, *12*, 1000.
[8] G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, *J. Med. Chem*. **2006**, *49*, 5912.
[9] E. Kellenberger, J. Rodrigo, P. Muller, D. Rognan, *Proteins Struct. Funct. Bioinform*. **2004**, *57*, 225.
[10] N. Paul, D. Rognan, *Proteins Struct. Funct. Bioinform*. **2002**, *47*, 521.
[11] J. S. Dixon, *Proteins Struct. Funct. Bioinform*. **1997**, *29*, 198.
[12] D. Plewczynski, M. Łaźniewski, R. Augustyniak, K. Ginalski, *J. Comput. Chem*. **2011**, *32*, 742.
[13] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem*. **1998**, *19*, 1639.
[14] G. Neudert, G. Klebe, *J. Chem. Inf. Model*. **2011**, *51*, 2731.
[15] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, *J. Med. Chem*. **2004**, *47*, 1739.
[16] I. Muegge, *J. Med. Chem*. **2006**, *49*, 5895.
[17] H. Gohlke, M. Hendlich, G. Klebe, *J. Mol. Biol*. **2000**, *295*, 337.
[18] W. Mooij, M. L. Verdonk, *Proteins Struct. Funct. Bioinform*. **2005**, *61*, 272.
[19] D. B. Kitchen, H. Decornez, J. R. Furr, J. Bajorath, *Nat. Rev. Drug Discov*. **2004**, *3*, 935.
[20] T. Cheng, Q. Li, Z. Zhou, Y. Wang, S. H. Bryant, *AAPS J*. **2012**, *14*, 133.
[21] S.-Y. Huang, S. Z. Grinter, X. Zou, *Phys. Chem. Chem. Phys*. **2010**, *12*, 12899.
[22] J. Liu, R. Wang, *J. Chem. Inf. Model*. **2015**, *55*, 475.
[23] P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, C. L. Brooks, *J. Med. Chem*. **2004**, *47*, 3032.
[24] T. Cheng, X. Li, Y. Li, Z. Liu, R. Wang, *J. Chem. Inf. Model*. **2009**, *49*, 1079.
[25] C. Bissantz, P. Bernard, M. Hibert, D. Rognan, *Proteins Struct. Funct. Bioinform*. **2003**, *50*, 5.
[26] S. L. McGovern, B. K. Shoichet, *J. Med. Chem*. **2003**, *46*, 2895.
[27] S. Karthikeyan, Q. Zhou, A. L. Osterman, H. Zhang, *Biochemistry* **2003**, *42*, 12532.
[28] J. A. Erickson, M. Jalaie, D. H. Robertson, R. A. Lewis, M. Vieth, *J. Med. Chem*. **2004**, *47*, 45.
[29] S. Renfrey, J. Featherstone, *Nat. Rev. Drug Discov*. **2002**, *1*, 175.
[30] J. Skolnick, H. Zhou, M. Gao, *Curr. Opin. Struct. Biol*. **2013**, *23*, 191.
[31] M. Brylinski, J. Skolnick, *J. Chem. Inf. Model*. **2010**, *50*, 1839.
[32] M. Brylinski, J. Skolnick, *PLoS Comput. Biol*. **2009**, *5*, e1000405.
[33] J. Skolnick, M. Brylinski, *Brief. Bioinform*. **2009**, *10*, 378.
[34] M. Brylinski, J. Skolnick, Q-Dock: *J. Comput. Chem*. **2008**, *29*, 1574.
[35] M. Brylinski, W. P. Feinstein, *J. Comput. Aided. Mol. Des*. **2013**, *27*, 551.
[36] G. Wang, R. L. Dunbrack, *Bioinformatics* **2003**, *19*, 1589.
[37] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res*. **2000**, *28*, 235.
[38] J. W. M. Nissink, C. Murray, M. Hartshorn, M. L. Verdonk, J. C. Cole, R. Taylor, *Proteins Struct. Funct. Bioinform*. **2002**, *49*, 457.
[39] M. Zacharias, *Protein Sci*. **2003**, *12*, 1271.
[40] M. Clark, R. D. Cramer, N. Van Opdenbosch, *J. Comput. Chem*. **1989**, *10*, 982.
[41] Y. Ding, Y. Fang, C. Daniel, P. W. Feinstein, R. Ramanujam, D. M. Koppelman, J. Moreno, M. Jarrell, M. Brylinski, submitted to *BMC Res Notes*.
[42] C. M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman, *J. Mol. Graph. Model*. **2003**, *21*, 289.
[43] R. S. Mulliken, *J. Chem. Phys*. **1955**, *23*, 1833.
[44] N. M. O'Boyle, R. Guha, E. L. Willighagen, S. E. Adams, J. Alvarsson, J.-C. Bradley, I. V Filippov, R. M. Hanson, M. D. Hanwell, G. R. Hutchison, *J. Cheminform*. 20011, *3*, 37.
[45] L. Yang, C. Tan, M.-J. Hsieh, J. Wang, Y. Duan, P. Cieplak, J. Caldwell, P. A. Kollman, R. Luo, *J. Phys. Chem. B* **2006**, *110*, 13166.
[46] H.-P. Schwefel, Numerical Optimization of Computer Models; John Wiley & Sons, Inc., New York, NY, **1981**.
[47] M. Levitt, *J. Mol. Biol*. **1976**, *104*, 59.
[48] S. Miyazawa, R. L. Jernigan, *Macromolecules* **1985**, *18*, 534.
[49] M. Brylinski, D. Lingam, *PLoS One* **2012**, *7*, e50200.
[50] M. Brylinski, *J. Chem. Inf. Model*. **2013**, *53*, 3097.
[51] T. Kawabata, *J. Chem. Inf. Model*. **2011**, *51*, 1775.
[52] D. J. Rogers, T. T. Tanimoto, *Science* **1960**, *132*, 1115.
[53] E. Parzen, *Ann. Math. Stat*. **1962**, *33*, 1065.
[54] M. Rosenblatt, *Ann. Math. Stat*. **1956**, *27*, 832.
[55] I. Muegge, M. Rarey, *Rev. Comput. Chem*. **2001**, *17*, 1.
[56] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, T. E. Ferrin, *J. Mol. Biol*. **1982**, *161*, 269.
[57] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali, *Curr. Protoc. Bioinform*. John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30, **2006**.
[58] Z. Zhang, O. F. Lange, *PLoS One* **2013**, *8*, e72096.
[59] M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, C. W. Murray, *J. Med. Chem*. **2007**, *50*, 726.
[60] M. Steinbach, G. Karypis, V. Kumar, *Text Mining*, Department of Computer Science and Engineering, University of Minnesota Twin City, Vol. *400*, **2000**; p. 525.
[61] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, M. Edelman, *Bioinformatics* **1999**, *15*, 327.
[62] V. Sobolev, T. M. Moallem, R. C. Wade, G. Vriend, M. Edelman, *Proteins Struct. Funct. Genet*. **1997**, *29*, 210.

[63] R. Najmanovich, J. Kuttner, V. Sobolev, M. Edelman, *Proteins Struct. Funct. Bioinform*. **2000**, *39*, 261.
[64] G. R. Stockwell, J. M. Thornton, *J. Mol. Biol*. **2006**, *356*, 928.
[65] M. C. Nicklaus, S. Wang, J. S. Driscoll, G. W. A. Milne, *Bioorg. Med. Chem*. **1995**, *3*, 411.
[66] R. Knegtel, I. D. Kuntz, C. M. Oshiro, *J. Mol. Biol*. **1997**, *266*, 424.
[67] M. Brylinski, S. Y. Lee, H. Zhou, J. Skolnick, *J. Struct. Biol*. **2011**, *173*, 558.
[68] H. Fan, D. Schneidman-Duhovny, J. J. Irwin, G. Dong, B. K. Shoichet, A. Sali, *J. Chem. Inf. Model*. **2011**, *51*, 3078.
[69] I. Muegge, Y. C. Martin, *J. Med. Chem*. **1999**, *42*, 791.
[70] H. Gohlke, G. Klebe, *Curr. Opin. Struct. Biol*. **2001**, *11*, 231.
[71] C. G. Wermuth, C. R. Ganellin, P. Lindberg, L. A. Mitscher, *Pure Appl. Chem*. **1998**, *70*, 1129.
[72] N. J. J. Salkind, Encyclopedia of Measurement and Statistics; SAGE Publications, Thousand Oaks, California, **2007**.
[73] R. L. Desjarlais, O. D. S. Yamashita, H. Oh, I. N. Uzinskas, K. F. Erhard, A. C. Allen, R. C. Haltiwanger, B. Zhao, W. W. Smith, S. S. Abdel-meguid, K. D. Alessio, C. A. Janson, M. S. Mcqueney, T. A. Tomaszek, X. M. A. Levy, D. F. Veber, K. Prussia, V. Pennsyl, R. V April, *J. Am. Chem. Soc*. **1998**, *120*, 9114.
[74] K. I. Varughese, Y. Su, D. Cromwell, S. Hasnain, N. H. Xuong, *Biochemistry* **1992**, *31*, 5172.
[75] S. Yin, L. Biedermannova, J. Vondrasek, N. V Dokholyan, *J. Chem. Inf. Model*. **2008**, *48*, 1656.
[76] Y. Z. Chen, D. G. Zhi, *Proteins Struct. Funct. Bioinform*. **2001**, *43*, 217.
[77] G. Lauro, M. Masullo, S. Piacente, R. Riccio, G. Bifulco, *Bioorg. Med. Chem*. **2012**, *20*, 3596.
[78] D.-L. Ma, D. S.-H. Chan, C.-H. Leung, *Chem. Soc. Rev*. **2013**, *42*, 2130.
[79] S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, P. E. Bourne, *PLoS Comput. Biol*. **2009**, *5*, e1000423.
[80] Y. Y. Li, J. An, S. J. M. Jones, *Genome Inform*. **2006**, *17*, 239.
[81] I. Guyon, A. Elisseeff, J. Mach. *Learn. Res*. **2003**, *3*, 1157.
[82] S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie, P. E. Bourne, *J. Chem. Inf. Model*. **2011**, *51*, 408.

Received: 10 April 2015
Revised: 7 June 2015
Accepted: 3 July 2015
Published online on 6 August 2015