JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

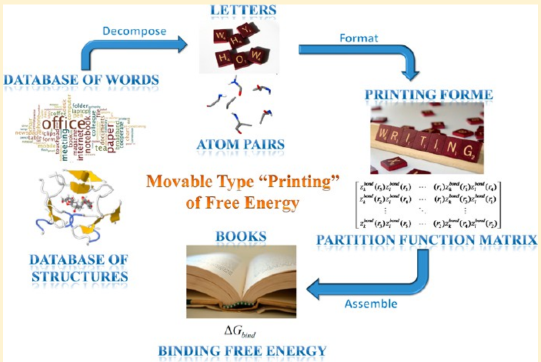
The Movable Type Method Applied to Protein−Ligand Binding

Zheng Zheng, Melek N. Ucisik, and Kenneth M. Merz*

Department of Chemistry and the Quantum Theory Project, University of Florida, 2328 New Physics Building, P.O. Box 118435, Gainesville, Florida 32611-8435, United States

S Supporting Information

ABSTRACT: Accurately computing the free energy for biological processes like protein folding or protein−ligand association remains a challenging problem. Both describing the complex intermolecular forces involved and sampling the requisite configuration space make understanding these processes innately difficult. Herein, we address the sampling problem using a novel methodology we term "movable type" (MT). Conceptually it can be understood by analogy with the evolution of printing and, hence, the name movable type. For example, a common approach to the study of protein−ligand complexation involves taking a database of intact drug-like molecules and exhaustively docking them into a binding pocket. This is reminiscent of early woodblock printing where each page had to be laboriously created prior to printing a book. However, printing evolved to an approach where a database of symbols (letters, numerals, etc.) was created and then assembled using a MT system, which allowed for the creation of all possible combinations of symbols on a given page, thereby, revolutionizing the dissemination of knowledge. Our MT method involves identifying all of the atom pairs seen in protein−ligand complexes and then creating two databases: one with their associated pairwise distant dependent energies and another associated with the probability of how these pairs can combine in terms of bonds, angles, dihedrals, and nonbonded interactions. Combining these two databases coupled with the principles of statistical mechanics allows us to accurately estimate binding free energies as well as the pose of a ligand in a receptor. This method, by its mathematical construction, samples all of the configuration space of a selected region (the protein active site here) in one shot without resorting to brute force sampling schemes involving Monte Carlo, genetic algorithms, or molecular dynamics simulations making the methodology extremely efficient. Importantly, this method explores the free energy surface eliminating the need to estimate the enthalpy and entropy components individually. Finally, low free energy structures can be obtained via a free energy minimization procedure yielding all low free energy poses on a given free energy surface. Besides revolutionizing the protein−ligand docking and scoring problem, this approach can be utilized in a wide range of applications in computational biology which involve the computation of free energies for systems with extensive phase spaces including protein folding, protein−protein docking, and protein design.



■ INTRODUCTION

Sampling the configuration space of complex biomolecules is a major hurdle impeding our ability to advance the understanding of a diverse range of processes including protein folding and the accurate prediction of ligand binding to a biological receptor.[1−10] Major advances have been made in computer hardware, which has allowed molecular dynamics (MD) simulations to reach the millisecond barrier, but this method is brute force in nature and requires highly sophisticated hardware and software.[2,10−14] Moreover, a major hurdle in the modeling of biological systems is associated with how the inter and intramolecular energies are modeled. Modern force fields are highly evolved but still need to be further refined to reach chemical accuracy in many applications.[9,14−17]

Predicting how a ligand (drug) binds its receptor and predicting its associated binding affinity is a highly challenging problem, which if solved, would singularly advance modern structure-based drug design.[8,15,17−31] This approach has largely employed so-called end-point methods that dock (place) a candidate molecule into a receptor and compute the binding free energy using a range of physics-based or empirical "scoring" functions. From an analysis of the error propagation properties in the statistical mechanics based prediction of protein−ligand binding affinities it was shown that the end-point class of approaches maximizes energy function uncertainties.[32−34] This can be alleviated through the use of sampling approaches including MD methods or methods that exhaustively sample the configuration space associated with protein−ligand binding.[1−3,5−7,9,11−14] These methods have shown that they can be successful but are brute force in nature, which lead us to consider ways in which we can use ideas more akin to the end-point methods but incorporate sampling at the same time. The concept being that this approach would give us the best of both worlds, while mitigating the effects of energy function deficiencies.

Using MD or exhaustive sampling procedures to evaluate protein−ligand binding is conceptually similar to woodblock

Received: July 10, 2013
Published: October 28, 2013

printing technology where all the words (molecules) are carefully placed on a board (receptor site) and the whole book can be printed (binding free energy determined). While a more advanced printing technology, movable type (MT) printing, (which was invented in China in the 11th century and introduced by Gutenberg into the alphabetic language system) uses a "database" of letters that is preconstructed and then the printing of any word involves a database search followed by the appropriate combination from the MT system. Using a typical pairwise potential the molecular energy of a system can be decomposed into atom pairwise interaction energies including bond, angle, torsion, and long-range noncovalent forces (van der Waals and electrostatic forces), which by analogy to the MT systems is our database of "letters". Each interaction has a different intensity and probability of occurrence along an atom pairwise coordinate axis. Herein, we develop the mathematics necessary to bring end-point methods up to the "MT printing level", via building a database of energy intensities and probabilities for all atom type pair interactions found in protein–ligand complexes. Using this information we then demonstrate that the MT approach enhances our ability to predict protein–ligand binding free energies and also allows us to extract the associated low-energy poses all at a fraction of the cost associated with "brute" force sampling methods. Moreover, the docking and scoring problem is an example of a broad class of problems in computational biology that involve both the computation of the free energy and the structure of a biological system, which includes challenges like the prediction of protein folds, protein–protein interactions, and protein design all of which the MT method can address.

METHODOLOGY

The MT Method Applied to Protein–Ligand Binding. A thermodynamic cycle modeling the binding free energy ΔG_b^s in solution (shown in Figure 1) is typically employed in end-point methods:

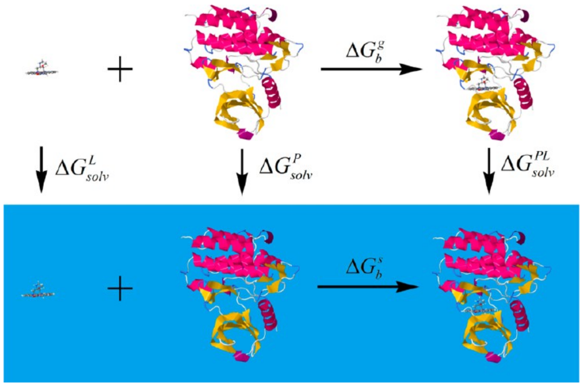


Figure 1. The thermodynamic cycle used to formulate the free energy of protein–ligand binding in solution.

$$\Delta G_b^s = \Delta G_b^g + \Delta G_{solv}^{PL} - \Delta G_{solv}^{L} - \Delta G_{solv}^{P} \tag{1}$$

where P and L indicate the protein and ligand, s and g represent the behavior in solution and the gas phase, respectively, ΔG_{solv} is the solvation free energy, and ΔG_b is the binding free energy in gas (g) and solution (s), respectively.

Using $\Delta\Delta G_{solv} = \Delta G_{solv}^{PL} - \Delta G_{solv}^{L} - \Delta G_{solv}^{P}$, eq 1 becomes

$$\Delta G_b^s = \Delta G_b^g - \Delta\Delta G_{solv} \tag{2}$$

The binding free energy in solution is now separated into two terms: The binding free energy in the gas phase and the change in the solvation free energy during the complexation process. At this point we introduce the MT algorithm to model both terms each with its own design.

The binding free energy in the gas phase is the most important term to evaluate in order to predict the protein–ligand binding affinity because it contains all interactions between the protein and ligand. When approximated as the Helmholtz free energy (NVT), the Gibbs (we use the canonical ensemble throughout, but will predominantly use the ΔG notation) binding free energy in the gas phase can be generated using the ratio of the partition functions describing the protein–ligand complex, the protein, and the ligand.

$$\Delta G_b^g \approx \Delta A_b^g = -RT\ln\left[\frac{Z_{PL}}{Z_P Z_L}\right] = -RT\ln\left[\frac{\int e^{-\beta E_{PL}(r)}dr}{\int e^{-\beta E_P(r)}dr \int e^{-\beta E_L(r)}dr}\right] \tag{3}$$

where Z represents the canonical ensemble partition function, and β is the reciprocal of the thermodynamic temperature $k_B T$. Partition functions are integrals over all possible coordinates of the protein–ligand complex, the protein, and the ligand. Equation 3 can be manipulated into the following form:

$$\Delta G_b^g = -RT\ln\left[\frac{F_{PL}\langle e^{-\beta E_{PL}(r)}\rangle}{F_P\langle e^{-\beta E_P(r)}\rangle F_L\langle e^{-\beta E_L(r)}\rangle}\right] \tag{4}$$

where the partition functions are expressed as the Boltzmann-weighted average of the pose energies (shown in brackets) multiplied the volume of configuration space available to each state, shown as F in eq 4. F_{PL} is approximated as the product of the external degrees of freedom (DoFs) of the bound protein and ligand (including the rotational and translational DoFs), and the internal DoFs of the bound protein and ligand (including the relative-positional and vibrational DoFs), given as

$$F_{PL} = F_{boundP}^{external}F_{boundL}^{external}F_{boundP}^{internal}F_{boundL}^{internal} \tag{5}$$

Similarly, the DoFs of the free protein and ligand molecules are also separated into the external and internal components. Internal DoFs are identical for bound and free protein/ligand structures, and the bound and free proteins are also assumed to share the same internal and external DoFs. Only the external DoFs of the ligand are differentiated between the bound and free systems. The rotational DoF of a free ligand is $8\pi^2$ on a normalized unit sphere. However, because of the inaccessible volume present in protein–ligand systems, the rotational DoFs of bound ligands are designated as $a\pi^2$ with a to-be-determined average volume factor a less than 8. The translational DoFs are treated as a constant C, which is assumed to be identical for all free ligands, while the translational DoF for bound ligands is the volume of the binding pocket V_{pocket} in which the ligands' center of mass can translate. Thereby, in the protein–ligand binding process, changes in the DoFs can be modeled as a constant with respect to the different volumes of the binding pockets. Applying these approximations we obtain:

$$\frac{F_{PL}}{F_P F_L} = \frac{F_{boundP}^{external}F_{boundL}^{external}F_{boundP}^{internal}F_{boundL}^{internal}}{F_{freeP}^{external}F_{freeP}^{internal} \times F_{freeL}^{external}F_{freeL}^{internal}} = \frac{F_{boundL}^{external}}{F_{freeL}^{external}}$$
$$= \frac{a\pi^2 V_{pocket}}{8\pi^2 C} = \frac{aV_{pocket}}{8C} \tag{6}$$

The gas-phase protein–ligand binding free energy can then be further manipulated into the following form:

$$\Delta G_{\mathrm{b}}^{\mathrm{g}} = -RT\ln\left[\frac{aV_{\mathrm{pocket}}\langle \mathrm{e}^{-\beta E_{\mathrm{PL}}(r)}\rangle}{8C\langle \mathrm{e}^{-\beta E_{\mathrm{P}}(r)}\rangle\langle \mathrm{e}^{-\beta E_{\mathrm{L}}(r)}\rangle}\right] \tag{7}$$

Again using the Helmholtz free energy approximation (eq 3), the solvation free energy can be correlated to the partition function of the solute (protein, ligand, or protein–ligand complex) and solute–solvent bulk interactions. In this way, the solvation free energy, using $\Delta G_{\mathrm{solv}}^{\mathrm{L}}$ as an example, is modeled as in eq 8, and the DoFs are approximated as being the same for the solute and the solute–solvent bulk terms.

$$\begin{aligned}\Delta G_{\mathrm{solv}}^{\mathrm{L}} \approx \Delta A_{\mathrm{solv}}^{\mathrm{L}} &= -RT\ln\left[\frac{Z_{\mathrm{LS}}}{Z_{\mathrm{L}}}\right] = -RT\ln\left[\frac{\int \mathrm{e}^{-\beta E_{\mathrm{LS}}(r)}dr}{\int \mathrm{e}^{-\beta E_{\mathrm{L}}(r)}dr}\right] \\ &= -RT\ln\left[\frac{\langle \mathrm{e}^{-\beta E_{\mathrm{LS}}(r)}\rangle}{\langle \mathrm{e}^{-\beta E_{\mathrm{L}}(r)}\rangle}\right]\end{aligned} \tag{8}$$

Finally, the remaining solvation terms given in eq 1 ($\Delta G_{\mathrm{solv}}^{\mathrm{P}}$ and $\Delta G_{\mathrm{solv}}^{\mathrm{PL}}$) can be modeled in an analogous manner yielding the change in the solvation free energy as ligand binding occurs which then can be used to evaluate the overall free energy of ligand binding in aqueous solution.

Construction of the MT System: Atom Pairwise Interaction Energy and Probability Databases. With pose energies sampled over all possible DoFs for the bound and free protein/ligand system, the gas-phase protein–ligand binding free energy can be generated using molecular dynamics, Monte Carlo, genetic algorithms, etc. by sampling over a large number of poses of the protein, ligand, and protein–ligand complex. Using the canonical ensemble the Helmholtz free energy can be obtained as the arithmetic mean (sum of the energies of all ligand poses divided by the total number of all poses along with an estimate of integration volume) of Boltzmann factors:

$$\begin{aligned}G \approx A &= -RT\ln[Z] = -RT\ln[\langle \mathrm{e}^{-\beta E}\rangle] \\ &= -RT\ln\left[\frac{\sum_i \mathrm{e}^{-\beta E_i}}{N}\right]\end{aligned} \tag{9}$$

However, the problem of pose-based energy sampling lies in the fact that pose selection and sample size significantly affect the final result, not to mention that calculating many unique poses is very time-consuming. Different ligand poses have different energy preferences for the binding site, which leads to a range of binding probabilities. When calculating the averaged partition functions in eq 7, one can assign probabilities (Q) as weights to different Boltzmann factors in order to differentiate the binding pocket preferences against ligand poses, rather than just simply using an arithmetic mean of all Boltzmann factors.

$$Q_i = \frac{\mathrm{e}^{-\beta E_i}}{\sum_i \mathrm{e}^{-\beta E_i}} \tag{10}$$

$$G \approx A = -RT\ln[Z] = -RT\ln[\langle \mathrm{e}^{-\beta E}\rangle] = -RT\ln[\sum_i Q_i \mathrm{e}^{-\beta E_i}] \tag{11}$$

The challenge in deriving the canonical partition function (as the denominator in eq 10) for a protein–ligand system is that it is difficult to include all relevant ligand pose energies within the binding pocket using brute force sampling schemes. However, the task becomes much easier when a protein–ligand system is reduced to the atom-pair level. In this way the "pose" sampling problem can then can be cast as a 1-D rather than a 3-D problem by deriving the canonical partition function as a sum of the Boltzmann factor products of all atom pairwise energies included in the system over all atom pairwise separation distance ranges.

$$\begin{aligned}Z &= \sum_{i}^{\text{all poses}} Q_i \mathrm{e}^{-\beta E_i} \\ &= \sum_{p}^{\text{all combinations}} \prod_{q}^{\text{all atom pairs}} Q_{pq}\mathrm{e}^{-\beta E_{pq}} \\ &= \sum_{\alpha}^{\text{bond distance range}} \prod_{a}^{\text{no. of bonds}} Q_{a\alpha}\mathrm{e}^{-\beta E_{a\alpha}} \\ &\times \sum_{\beta}^{\text{angle distance range}} \prod_{b}^{\text{no. of angles}} Q_{b\beta}\mathrm{e}^{-\beta E_{b\beta}} \\ &\times \sum_{\gamma}^{\text{torsion distance range}} \prod_{c}^{\text{no. of torsions}} Q_{c\gamma}\mathrm{e}^{-\beta E_{c\gamma}} \\ &\times \sum_{\delta}^{\text{vdW}-\text{elec distance range}} \prod_{d}^{\text{no. of vdW}-\text{elec interactions}} Q_{d\delta}\mathrm{e}^{-\beta E_{d\delta}}\end{aligned} \tag{12}$$

The canonical partition function can be derived following eq 12, where the index "i" refers to each ligand pose (microstate) in a "traditional" brute force sampling scheme. When the protein–ligand system is broken down to the atom-pair level, "q" indicates all atom pairs in the molecular system, and "p" indicates each possible combination of all atom pairs each of which is at a prechosen distance. a, b, c and d refer to each atom pair as a bond, angle, torsion or long-range (van der Waals or electrostatic) interaction in the canonical system, respectively, and α, β, γ and δ refers to each sampled separation distance between the corresponding atom pair. Probabilities of all the atom pairwise distributions on the right-hand side of eq 12 are normalized as $\sum_i Q_i = \sum_i(\mathrm{e}^{-\beta E_i}/(\sum_i \mathrm{e}^{-\beta E_i})) = 1$:

$$\begin{aligned}&\sum_{\alpha}^{\text{bond distance range}} \prod_{a}^{\text{no. of bonds}} Q_{a\alpha} \times \sum_{\beta}^{\text{angle distance range}} \prod_{b}^{\text{no. of angles}} Q_{b\beta} \\ &\times \sum_{\gamma}^{\text{torsion distance range}} \prod_{c}^{\text{no. of torsions}} Q_{c\gamma} \\ &\times \sum_{\delta}^{\text{vdW}-\text{elec distance range}} \prod_{d}^{\text{no. of vdW}-\text{elec interactions}} Q_{d\delta} = 1\end{aligned} \tag{13}$$

Hence our MT method is designed to decompose the molecular energy into atom pairwise energies, which then simplifies the energy sampling problem to the atom-pair level. The advantage of this idea lies in that atom pairs can be categorized based on atom and interaction types, e.g. bond, angle, torsion, and long-range noncovalent interactions and that calculation of atom pairwise energies is extremely cheap. Thereby, it is easy to build an atomic pairwise interaction matrix of energy vs distance for each interaction type and atom pair type i, j. Hence, the energy calculation for each molecule is no more than a combination of elements from different energy matrices. In addition, the MT method is a template by which any pairwise decomposable energy function can be used. In the current work, the energy for each interaction type between a certain atom type pair i, j is calculated using the knowledge-based and empirical combined scoring algorithm (KECSA) potential function.[35] In KECSA, the protein−ligand statistical potential is modified and equated to an atom pairwise energy in order to generate force field parameters for bond stretching, angle bending, dihedral torsion angles, and long-range noncovalent interactions. Please see the detailed rationale and justification for KECSA and its parametrization in the Supporting Information and the relevant literature.[35]

Along with the distance-based energy, each atom pair type also has a distance preference encoded in its distribution, resulting in different probabilities associated with Boltzmann factors for each sampled atom pairwise distance. Atom-pair radial distributions were collected from a protein−ligand structure training set (i.e., the PDBbind v2011 data set with 6019 protein−ligand structures)[36,37] and utilized in the current model. The atom pairwise radial distribution function is modeled as

$$g_{ij}(r) = \frac{n_{ij}(r)}{n^*_{ij}(r)} = \frac{n_{ij}(r)}{\frac{N_{ij}}{V} 4\pi r^a \Delta r} \tag{14}$$

where $n_{ij}(r)$ is the number of protein−ligand pairwise interactions between a certain atom pair type i and j in the bin $(r, r + \Delta r)$, with the volume $4\pi r^a \Delta r$ collected from the training set, and $n^*_{ij}(r)$ in the denominator mimics the number of protein−ligand atom type pairs i and j in the same distance bin in an ideal gas state. This removes the "non-interacting" background distribution from the protein−ligand system; Δr is defined as 0.005 Å, and N_{ij} is the total number of atom pairs of type i and j. The average volume V of the protein−ligand binding sites is given as $4/(a + 1)\pi R^{a+1}$, with the same to-be-determined parameter a as described above (eqs 7 and 14). A cutoff distance R is assigned to each atom type pair defining the distance at which the atom pairwise interaction energy can be regarded as zero. Both a and R can be derived using a previously introduced method.[35] The radial distribution frequency is then normalized by dividing the sum of radial distributions of all the atom pairs in the system (eq 15).

$$q_{ij}(r) = \frac{g_{ij}(r)}{\sum_i \prod_j g_{ij}(r)} = \frac{\frac{R^{a+1} n_{ij}(r)}{(a+1) N_{ij} r^a \Delta r}}{\sum_i \prod_j \frac{R^{a+1} n_{ij}(r)}{(a+1) N_{ij} r^a \Delta r}} \tag{15}$$

In this way, the energy and distribution frequency vs distance is calculated for any interaction type and atom pair type, thereby, forming our MT database for later use.

Binding Free Energies from the MT Method. Based on eq 4, the binding free energy is defined as a ratio of partition functions of the different molecules involved in the binding process, i.e., the protein, ligand, and the protein−ligand complex. Instead of sampling over poses of one molecule, the MT method simplifies the partition function of each system into a collection of partition functions (c) over each observed atom pair, which are equal to the normalized distribution probability of the atom type pair along the distance (q), multiplied by the corresponding atom pairwise partition function (z):

$$c = q \cdot z \tag{16}$$

By combining the partition functions c over all atom pairs in one molecule the partition function of one molecule averaged over all possible conformations is derived (eq 17).

$$\langle e^{-\beta E(r)} \rangle = \sum_j^M \prod_i^N c_{ij}(r) \tag{17}$$

where the averaged molecular partition function is given as a sum of atom pairwise partition functions c sampled over distance intervals (M) of all combination of N atom pairs at all possible distances.

Starting from the protein−ligand complex database, we constructed the partition function matrices for the MT algorithm. When converted into a partition function matrix, the atom pairwise energy multiplier sampled as a function of distance is the basic element needed to assemble the total energy, as shown in eq 18, using the protein bond energy as an example.

$$\mathbb{z}_k^{\text{bond}} = \begin{bmatrix} z_k^{\text{bond}}(r_1) \\ z_k^{\text{bond}}(r_2) \\ \vdots \\ z_k^{\text{bond}}(r_a) \\ \vdots \\ z_k^{\text{bond}}(r_n) \end{bmatrix} = \begin{bmatrix} e^{-\beta E_k^{\text{bond}}(r_1)} \\ e^{-\beta E_k^{\text{bond}}(r_2)} \\ \vdots \\ e^{-\beta E_k^{\text{bond}}(r_a)} \\ \vdots \\ e^{-\beta E_k^{\text{bond}}(r_n)} \end{bmatrix} \tag{18}$$

where subscript k indicates a bonded atom pair i and j, and each distance increment between any r_a and r_{a+1} is 0.005 Å. Using this scheme the distance sampling size is given by: $n = (r_n - r_1)/(0.005\text{Å})$, where r_1 and r_n are the lower and upper bounds for distance sampling, which varies depending on the each atom pair and interaction type. The product over all bond-linked atom pairs derives the total bond partition function in the protein:

$$\begin{aligned} Z_P^{\text{bond}} &= \mathbb{z}_1^{\text{bond}} \otimes \mathbb{z}_2^{\text{bond}} \otimes \mathbb{z}_3^{\text{bond}} \otimes \cdots \otimes \mathbb{z}_m^{\text{bond}} \\ &= \mathbb{z}_1^{\text{bond}} \cdot (\mathbb{z}_2^{\text{bond}})^{\mathrm{T}} \cdot (\mathbb{z}_3^{\text{bond}})^{\mathrm{T}} \cdot \cdots \cdot (\mathbb{z}_m^{\text{bond}})^{\mathrm{T}} \end{aligned} \tag{19}$$

$$Z_{\mathrm{P}}^{\text{bond}} = \begin{bmatrix} z_1^{\text{bond}}(r_1) z_2^{\text{bond}}(r_1) \cdots z_m^{\text{bond}}(r_1) & z_1^{\text{bond}}(r_1) z_2^{\text{bond}}(r_1) \cdots z_m^{\text{bond}}(r_2) & \cdots & z_1^{\text{bond}}(r_1) z_2^{\text{bond}}(r_n) \cdots z_m^{\text{bond}}(r_n) \\ z_1^{\text{bond}}(r_2) z_2^{\text{bond}}(r_1) \cdots z_m^{\text{bond}}(r_1) & z_1^{\text{bond}}(r_2) z_2^{\text{bond}}(r_1) \cdots z_m^{\text{bond}}(r_2) & \cdots & z_1^{\text{bond}}(r_2) z_2^{\text{bond}}(r_n) \cdots z_m^{\text{bond}}(r_n) \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{\text{bond}}(r_n) z_2^{\text{bond}}(r_1) \cdots z_m^{\text{bond}}(r_1) & z_1^{\text{bond}}(r_n) z_2^{\text{bond}}(r_1) \cdots z_m^{\text{bond}}(r_2) & \cdots & z_1^{\text{bond}}(r_n) z_2^{\text{bond}}(r_n) \cdots z_m^{\text{bond}}(r_n) \end{bmatrix} \tag{20}$$

In eqs 19 and 20, m indicates the total number of atom pairs that need to have their bond stretch term computed (i.e., number of covalent bonds), and n is the distance sampling size. T indicates the transpose. Thus, the matrix $Z_{\mathrm{P}}^{\text{bond}}$ has a total of n^m elements, and includes all combinations of the sampled atom pairwise distances and atom pairs (see eq 20). Energy matrices for other kinds of atom pairwise interactions are assembled in the same way (i.e., bond, angle, torsion, and long-range interactions). A simple example is given in Supporting Information (butane−methane interaction), which illustrates the method in more detail. Products over these matrices generate the entire protein partition function matrix (eq 21), representing all possible combinations of the protein internal energies with different atom pairwise distances.

$$\mathbb{Z}_{\mathrm{P}} = \mathrm{Z}_{\mathrm{P}}^{\mathrm{bond}} \otimes \mathrm{Z}_{\mathrm{P}}^{\mathrm{angle}} \otimes \mathrm{Z}_{\mathrm{P}}^{\mathrm{torsion}} \otimes \mathrm{Z}_{\mathrm{P}}^{\mathrm{long\text{-}range}} \tag{21}$$

where

$$\mathrm{Z}_{\mathrm{P}}^{\mathrm{long\text{-}range}} = \mathrm{Z}_{\mathrm{P}}^{\mathrm{vdW-elec}} \otimes \mathrm{Z}_{\mathrm{P}}^{\mathrm{H\text{-}bond}} \tag{22}$$

The KECSA van der Waals–electrostatic interaction models and hydrogen bond models[35] are applied to the protein, ligand and protein–ligand complex systems. Similarly, the ligand energy (eq 23) and protein–ligand interaction energy matrices (eq 24) can be obtained.

$$\mathbb{Z}_{\mathrm{L}} = \mathrm{Z}_{\mathrm{L}}^{\mathrm{bond}} \otimes \mathrm{Z}_{\mathrm{L}}^{\mathrm{angle}} \otimes \mathrm{Z}_{\mathrm{L}}^{\mathrm{torsion}} \otimes \mathrm{Z}_{\mathrm{L}}^{\mathrm{long\text{-}range}} \tag{23}$$

$$\begin{aligned}\mathbb{Z}_{\mathrm{PL}} = {} & \mathrm{Z}_{\mathrm{P}}^{\mathrm{bond}} \otimes \mathrm{Z}_{\mathrm{P}}^{\mathrm{angle}} \otimes \mathrm{Z}_{\mathrm{P}}^{\mathrm{torsion}} \otimes \mathrm{Z}_{\mathrm{P}}^{\mathrm{long\text{-}range}} \otimes \mathrm{Z}_{\mathrm{L}}^{\mathrm{bond}} \otimes \mathrm{Z}_{\mathrm{L}}^{\mathrm{angle}} \\ & \otimes \mathrm{Z}_{\mathrm{L}}^{\mathrm{torsion}} \otimes \mathrm{Z}_{\mathrm{L}}^{\mathrm{long\text{-}range}} \otimes \mathrm{Z}_{\mathrm{PL}}^{\mathrm{long\text{-}range}}\end{aligned} \tag{24}$$

The distribution frequency matrix is built in the same way, with the $q_{ij}(r)$ derived from eq 15 as elements in each multiplier (also using the protein bond term as an example):

$$\mathbb{q}_{k}^{bond} = \begin{bmatrix} q_k^{\mathrm{bond}}(r_1) \\ q_k^{\mathrm{bond}}(r_2) \\ q_k^{\mathrm{bond}}(r_3) \\ \vdots \\ q_k^{\mathrm{bond}}(r_n) \end{bmatrix} \tag{25}$$

$$\begin{aligned}\mathrm{Q}_{\mathrm{P}}^{\mathrm{bond}} = {} & \mathbb{q}_{1}^{\mathrm{bond}} \otimes \mathbb{q}_{2}^{\mathrm{bond}} \otimes \mathbb{q}_{3}^{\mathrm{bond}} \otimes \cdots \otimes \mathbb{q}_{k}^{\mathrm{bond}} \otimes \cdots \\ & \otimes \mathbb{q}_{m}^{\mathrm{bond}}\end{aligned} \tag{26}$$

$$\mathbb{Q}_{\mathrm{P}} = \mathrm{Q}_{\mathrm{P}}^{\mathrm{bond}} \otimes \mathrm{Q}_{\mathrm{P}}^{\mathrm{angle}} \otimes \mathrm{Q}_{\mathrm{P}}^{\mathrm{torsion}} \otimes \mathrm{Q}_{\mathrm{P}}^{\mathrm{long\text{-}range}} \tag{27}$$

where

$$\mathrm{Q}_{\mathrm{P}}^{\mathrm{long\text{-}range}} = \mathrm{Q}_{\mathrm{P}}^{\mathrm{vdW-elec}} \otimes \mathrm{Q}_{\mathrm{P}}^{\mathrm{H\text{-}bond}} \tag{28}$$

The distribution frequency matrix for the protein is derived using eqs 26–28, and the distribution frequency matrices of the ligand and protein–ligand intermolecular interactions are analogously derived as in eqs 29 and 30.

$$\mathbb{Q}_{\mathrm{L}} = \mathrm{Q}_{\mathrm{L}}^{\mathrm{bond}} \otimes \mathrm{Q}_{\mathrm{L}}^{\mathrm{angle}} \otimes \mathrm{Q}_{\mathrm{L}}^{\mathrm{torsion}} \otimes \mathrm{Q}_{\mathrm{L}}^{\mathrm{long\text{-}range}} \tag{29}$$

$$\mathbb{Q}_{\mathrm{PL}} = \mathrm{Q}_{\mathrm{PL}}^{\mathrm{long\text{-}range}} = \mathrm{Q}_{\mathrm{PL}}^{\mathrm{vdW-elec}} \otimes \mathrm{Q}_{\mathrm{PL}}^{\mathrm{H\text{-}bond}}$$

$$\begin{aligned}\mathbb{Q}_{\mathrm{PL}} = {} & \mathrm{Q}_{\mathrm{P}}^{\mathrm{bond}} \otimes \mathrm{Q}_{\mathrm{P}}^{\mathrm{angle}} \otimes \mathrm{Q}_{\mathrm{P}}^{\mathrm{torsion}} \otimes \mathrm{Q}_{\mathrm{P}}^{\mathrm{long\text{-}range}} \otimes \mathrm{Q}_{\mathrm{L}}^{\mathrm{bond}} \\ & \otimes \mathrm{Q}_{\mathrm{L}}^{\mathrm{angle}} \otimes \mathrm{Q}_{\mathrm{L}}^{\mathrm{torsion}} \otimes \mathrm{Q}_{\mathrm{L}}^{\mathrm{long\text{-}range}} \otimes \mathrm{Q}_{\mathrm{PL}}^{\mathrm{long\text{-}range}}\end{aligned} \tag{30}$$

We chose the same range and distance increment in both the energy and distribution frequency calculations, which means that any r_x ($x = 1, 2, 3, ...$) in eq 18 is the same as corresponding r_x in eq 25. Thus, the corresponding elements in all energy and distribution frequency matrices correlate with each other. The pointwise product over all matrices ensures that the energies and distribution frequencies with the same range and distance increment are combined into one element in the final matrix of the probability-weighted partition function of the protein–ligand complex ($\mathbb{C}_{\mathrm{PL}}$ in eq 31).

$$\mathbb{C}_{\mathrm{PL}} = \mathbb{Q} \cdot \mathbb{Z} = \mathbb{Q}_{\mathrm{PL}} \cdot \mathbb{Z}_{\mathrm{PL}} \tag{31}$$

In the final matrix each element of $\mathbb{C}_{\mathrm{PL}}$ is a value of the partition function of the protein–ligand complex multiplied by its probability based on its radial distribution forming the ensemble average. Finally, the sum of all elements of the matrix $\mathbb{C}_{\mathrm{PL}}$ gives us the averaged partition function of the protein–ligand complex:

$$\begin{aligned}& \because \mathrm{Sum}(\mathbb{Q}) = 1; \\ & \therefore \mathrm{Sum}(\mathbb{C}_{\mathrm{PL}}) = \mathrm{Sum}(\mathbb{Q} \cdot \mathbb{Z}) = \langle \mathrm{e}^{-\beta E_{\mathrm{PL}}(r)} \rangle\end{aligned} \tag{32}$$

where the first equation is the normalization statement for the probabilities. In this manner, the normalized averaged partition function of the protein–ligand complex is derived in eq 32. Following the same procedure, the averaged partition functions for the protein and ligand are generated as well

$$\langle \mathrm{e}^{-\beta E_{\mathrm{P}}(r)} \rangle = \mathrm{Sum}(\mathbb{C}_{\mathrm{P}}) \tag{33}$$

$$\langle \mathrm{e}^{-\beta E_{\mathrm{L}}(r)} \rangle = \mathrm{Sum}(\mathbb{C}_{\mathrm{L}}) \tag{34}$$

Expanding the matrices, the protein–ligand binding free energy in the gas phase is defined as in eq 35, using the averaged partition functions of all three systems (protein, ligand, protein–ligand complex) derived above.

$$\begin{aligned}\Delta G_{\mathrm{b}}^{\mathrm{g}} &= -RT \ln\left[\frac{aV_{\mathrm{pocket}} \langle \mathrm{e}^{-\beta E_{\mathrm{PL}}(r)} \rangle}{8C \langle \mathrm{e}^{-\beta E_{\mathrm{P}}(r)} \rangle \langle \mathrm{e}^{-\beta E_{\mathrm{L}}(r)} \rangle}\right] \\ &= -RT\left(\ln\left[\frac{aV_{\mathrm{pocket}}}{8C}\right]\right. \\ &\quad \left. + \ln\left[\frac{\sum_{ijk}^{I+J+K} (\tilde{\mathrm{Q}}_i^{\mathrm{P}} \tilde{\mathrm{Q}}_j^{\mathrm{L}} \tilde{\mathrm{Q}}_k^{\mathrm{PL}} \exp[-\beta(E_i^{\mathrm{P}} + E_j^{\mathrm{L}} + E_k^{\mathrm{PL}})])}{(\sum_i^I \mathrm{Q}_i^{\mathrm{P}} \exp[-\beta E_i^{\mathrm{P}}])(\sum_j^J \mathrm{Q}_j^{\mathrm{L}} \exp[-\beta E_j^{\mathrm{L}}])}\right]\right)\end{aligned} \tag{35}$$

In eq 35, Q is the radial distribution frequency and E is the energy. i, j, k are the indices of the protein, ligand and protein–ligand complex, while I, J, K are the total number of protein, ligand, and protein–ligand complex samples, respectively. $\sum_j^I Q_i^{\mathrm{P}} = \sum_j^I Q_J^{\mathrm{L}} = \sum_j^K Q_j^{\mathrm{PL}} = 1$. $\tilde{Q}_i^{\mathrm{P}}$, $\tilde{Q}_j^{\mathrm{L}}$, and $\tilde{Q}_k^{\mathrm{PL}}$ are standard distribution frequency matrices normalized over all three systems, in order to satisfy $\sum_{ijk}^{I+J+K}. \tilde{Q}_i^{\mathrm{P}} \tilde{Q}_j^{\mathrm{L}} \tilde{Q}_k^{\mathrm{PL}} = 1$. In this way the protein–ligand binding free energy in the gas phase is derived using our MT algorithm.

Determination of the change in the solvation energy as a function of the binding process is computed in a similar manner. To illustrate this we describe how we obtain the solvation free energy of the ligand, which is one component of $\Delta\Delta G_{\mathrm{solv}}$ and by extension the other terms can be derived.

The ligand solvation free energy is obtained by decomposing the ligand–solvent bulk energy into the free ligand energy $E_{\mathrm{L}}(r)$, the ligand–solvent polar interaction energy $E_{\mathrm{psol}}(r)$, and the ligand–solvent nonpolar interaction energy $E_{\mathrm{npsol}}(r)$:

$$E_{\mathrm{LS}}(r) = E_{\mathrm{L}}(r) + E_{\mathrm{psol}}(r) + E_{\mathrm{npsol}}(r) \tag{36}$$

Solvent was approximated as a shell of even thickness around the ligand, in which the water molecules were evenly distributed. The solvent shell thickness was 6 Å, and the inner surface of the shell was 1.6 Å away from the ligand

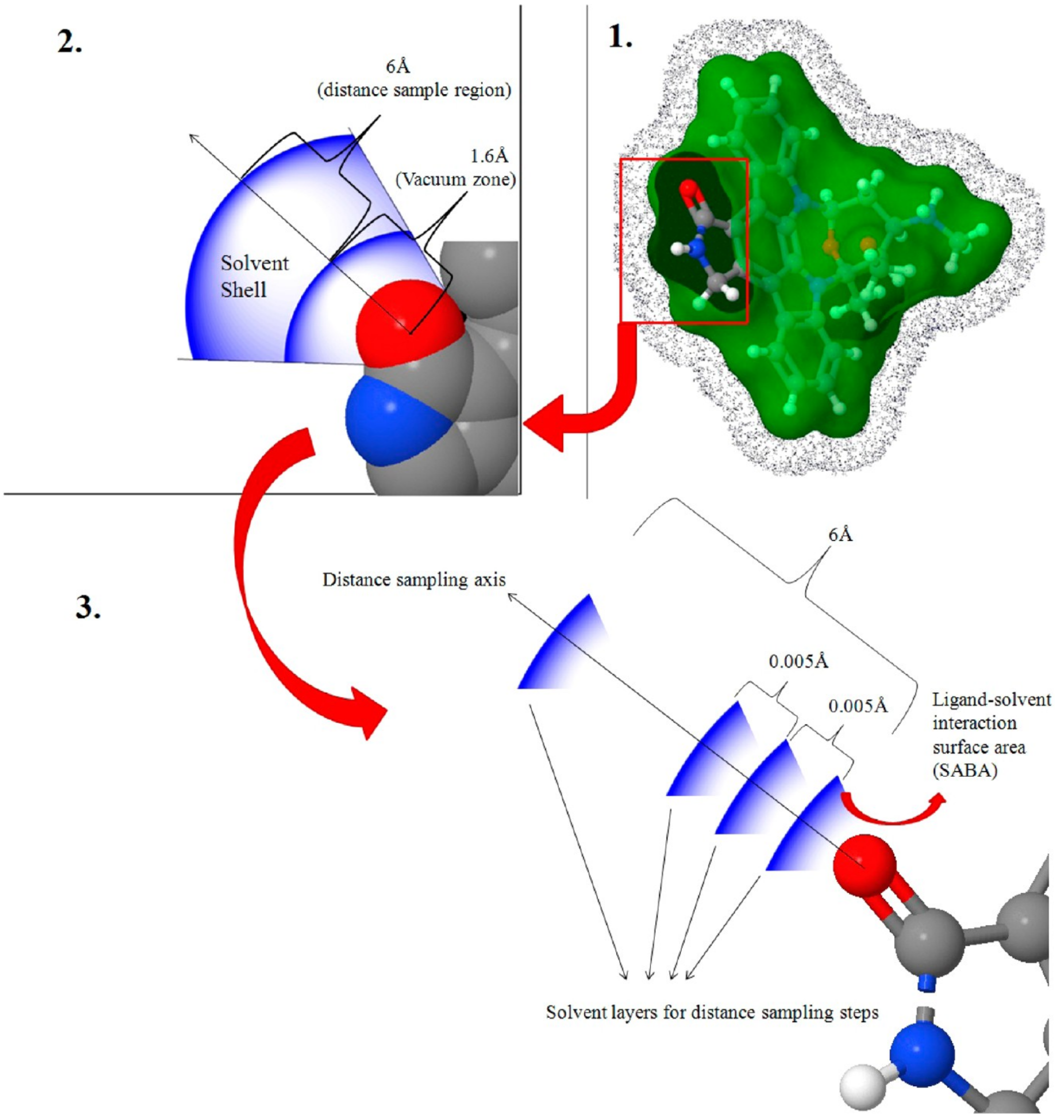


Figure 2. Modeling of ligand−solvent polar interaction using a Boltzmann factor multiplier. A carbonyl oxygen atom is used as an example here. (1) The green surface shows the solvent accessible surface of the ligand (inner layer of the solvent shell). The surface consisting of blue dots represents the outer boundary surface of the solvent shell. (2) A close-up view of a selected polar atom (carbonyl oxygen) with its solvent shell. (3) Monte Carlo sampling along carbonyl oxygen−solvent shell layer distance.

surface, which approximates the radius of a water molecule. Herein, for simplicity, the ligand−solvent polar interaction was considered as a surface (solvent accessible polar surface of the ligand)−surface (solvent bulk layer surface at a certain distance away from ligand) interaction, instead of a point−point interaction, i.e. atom pairwise interaction. Using this model the ligand polar atom−solvent interaction energy was modeled as a solvent accessible buried area (SABA) of the ligand polar atoms multiplied by the polar atom−oxygen interaction energy terms taken from KECSA,[35] to simulate the ligand−solvent surface interaction energy. All SABA-weighted interaction energies along the solvent shell thickness, with a 0.005 Å increment were collected and stored. The ligand−solvent polar interaction Boltzmann factor multiplier was modeled using eq 37, with k indicating each polar atom in the ligand, r_1 = 1.6 Å, which is the inner layer of the solvent shell and r_n = 6 + 1.6 = 7.6 Å, which is the outer boundary layer of the solvent shell.

$$\mathbb{Z}_k^{\text{psol}} = \begin{bmatrix} z_k^{\text{psol}}(r_1) \\ z_k^{\text{psol}}(r_2) \\ z_k^{\text{psol}}(r_3) \\ \vdots \\ z_k^{\text{psol}}(r_n) \end{bmatrix} = \begin{bmatrix} e^{-\beta(\text{SABA})_k E_k^{\text{psol}}(r_1)} \\ e^{-\beta(\text{SABA})_k E_k^{\text{psol}}(r_2)} \\ e^{-\beta(\text{SABA})_k E_k^{\text{psol}}(r_3)} \\ \vdots \\ e^{-\beta(\text{SABA})_k E_k^{\text{psol}}(r_n)} \end{bmatrix} \tag{37}$$

The ligand−solvent polar interaction Boltzmann factor matrix is then derived using eq 38, covering all ligand polar atoms up to m. The distribution frequency matrices were not included in ligand−solvent energy calculation because the radial distribution function is approximated as being identical along all ligand−solvent distances (i.e., a featureless continuum). Figure 2 further illustrates the modeling of the ligand−solvent polar interaction.

$$\mathbb{Z}_{psol} = \mathbb{Z}_1^{\text{psol}} \cdot (\mathbb{Z}_2^{\text{psol}})^{\text{T}} \cdot (\mathbb{Z}_3^{\text{psol}})^{\text{T}} \cdots\cdots (\mathbb{Z}_k^{\text{psol}})^{\text{T}} \cdots\cdots (\mathbb{Z}_m^{\text{psol}})^{\text{T}} \tag{38}$$

The nonpolar atom buried area (NABA) is used to simulate the interactions between the nonpolar atoms and aqueous solvent, because the interaction energy between nonpolar atoms and water molecules has a weaker response to changes in distance.

$$\mathbb{Z}_{\mathrm{npsol}} = [\mathrm{e}^{-\beta \mathrm{NABA}}] \tag{39}$$

The ligand energy is the same as was introduced in the gas-phase protein–ligand binding free energy calculation. So, the matrix for the ligand–solvent interaction energy is:

$$\mathbb{C}^{L}_{\mathrm{solv}} = \mathbb{Z}_{L}\cdot\mathbb{Z}_{\mathrm{psol}}\cdot\mathbb{Z}_{\mathrm{npsol}} \tag{40}$$

$$\mathrm{Sum}(\mathbb{C}^{L}_{\mathrm{solv}}) = \mathrm{Sum}(\mathbb{Z}_{\mathrm{L}}\cdot\mathbb{Z}_{\mathrm{psol}}\cdot\mathbb{Z}_{\mathrm{npsol}}) = \langle \mathrm{e}^{-\beta E_{\mathrm{LS}}(r)}\rangle \tag{41}$$

The solvation free energy was not fit to experimental solvation free energies and was found to have a small influence of the final binding free energies for the protein–ligand complexes. Nonetheless, future work will fit these models to small molecule solvation free energies, but for the present application the solvation model was used as formulated above.

With all necessary components constructed, the binding free energy in solution can be generated using:

$$\begin{aligned}\Delta G^{\mathrm{s}}_{\mathrm{b}} &= \Delta G^{\mathrm{g}}_{\mathrm{b}} - \Delta\Delta G_{\mathrm{solv}} = \Delta G^{\mathrm{g}}_{\mathrm{b}} + \Delta G^{\mathrm{PL}}_{\mathrm{solv}} - \Delta G^{\mathrm{P}}_{\mathrm{solv}} - \Delta G^{\mathrm{L}}_{\mathrm{solv}} = -RT\ln\left[\frac{aV_{\mathrm{pocket}}\langle \mathrm{e}^{-\beta E_{\mathrm{PL}}(r)}\rangle}{8C\langle \mathrm{e}^{-\beta E_{\mathrm{P}}(r)}\rangle\langle \mathrm{e}^{-\beta E_{\mathrm{L}}(r)}\rangle}\right] - RT\ln\left[\frac{\langle \mathrm{e}^{-\beta E_{\mathrm{PLS}}(r)}\rangle}{\langle \mathrm{e}^{-\beta E_{\mathrm{PL}}(r)}\rangle}\right] \\ &+ RT\ln\left[\frac{\langle \mathrm{e}^{-\beta E_{\mathrm{PS}}(r)}\rangle}{\langle \mathrm{e}^{-\beta E_{\mathrm{P}}(r)}\rangle}\right] + RT\ln\left[\frac{\langle \mathrm{e}^{-\beta E_{\mathrm{LS}}(r)}\rangle}{\langle \mathrm{e}^{-\beta E_{\mathrm{L}}(r)}\rangle}\right] = -RT\ln\left[\frac{aV_{\mathrm{pocket}}\langle \mathrm{e}^{-\beta E_{\mathrm{PLS}}(r)}\rangle}{8C\langle \mathrm{e}^{-\beta E_{\mathrm{PS}}(r)}\rangle\langle \mathrm{e}^{-\beta E_{\mathrm{LS}}(r)}\rangle}\right] \\ &= -RT\left(\ln\left[\frac{aV_{\mathrm{pocket}}}{8C}\right] + \ln\left[\frac{\sum_{ijk}^{I+J+K}(\tilde{Q}_i^{\mathrm{P}}\tilde{Q}_j^{\mathrm{L}}\tilde{Q}_k^{\mathrm{PL}}\exp[-\beta(E_i^{\mathrm{P}} + E_j^{\mathrm{L}} + E_k^{\mathrm{PL}} + E_s^{\mathrm{PLpsol}} + \mathrm{NABA}^{\mathrm{PL}})])}{(\sum_i^{I} Q_i^{\mathrm{P}}\exp[-\beta E_i^{\mathrm{P}} + E_s^{\mathrm{Ppsol}} + \mathrm{NABA}^{\mathrm{P}}])(\sum_{js}^{J+S} Q_j^{\mathrm{L}}\exp[-\beta(E_j^{\mathrm{L}} + E_s^{\mathrm{Lpsol}} + \mathrm{NABA}^{\mathrm{L}})])}\right]\right)\end{aligned} \tag{42}$$

Performance of MT KECSA as a Scoring Function for Protein–Ligand Binding Affinity Prediction. Using the MT method we performed binding free energy calculations with the KECSA model and its associated parameters. This validation study was performed to illustrate (1) the general performance of MT method when used to predict protein–ligand binding affinities and (2) whether sampling along atom pairwise distance improves scoring performance, as done in MT KECSA, improves the prediction over the original KECSA method.

A test set containing 795 protein–ligand complexes was chosen from the PDBbind v2011 refined data set based on the following criteria: (1) Crystal structures of all selected complexes had X-ray resolutions of <2.5 Å; (2) complexes with molecular weights (MWs) distributed from 100 to 900 were selected, to avoid ligand size-dependent prediction results; and (3) complexes with ligands who have more than 20 hydrogen donors and acceptors, more than one phosphorus atom, and complexes with metalloproteins were excluded.

MT KECSA calculations show improvements in Pearson's r, Kendall τ and root-mean-square error (RMSE) when compared to the original KECSA model (Table 1). Importantly, judging from the slope and intercept of both calculations versus experimental data, MT KECSA (with slope of 0.85 and intercept of 0.14) better reproduces the binding affinities in the low- and high-affinity regions than the original KECSA model (with slope of 0.27 and intercept of 3.57). In the original KECSA approach, the entropy terms were empirically trained, thus its test results demonstrate training set dependence to some degree. Because complexes with medium-binding affinities are more commonly seen in the PDB database when compared to complexes with high- or low-binding affinities, they become the majority in a large training set (1982 protein–ligand complexes were used to fit the original KECSA entropy terms). This causes the trained scoring functions to overestimate the binding affinity of the low-binding complexes while underestimating that of the high-binding complexes. On the other hand, MT KECSA, using canonical partition functions to compute the binding free energies, bypasses the difficulty of empirically building the entropy term and, thereby, better reproduces the binding affinity in low- and high-binding free energy regions.

Table 1. Statistical Results for MT KECSA and Original KECSA Correlated with Experimental Binding Affinities

	Pearson's r	RMSE ($\mathrm{p}K_{\mathrm{d}}$)	Kendall τ
MT KECSA	0.72	1.88	0.53
original KECSA	0.62	2.03	0.46

Extracting Heat Maps from the MT Method. Grid based methods and their graphical representation have had a long tradition in computer-aided drug design.[4,38−41] For example, COMFA[42] creates a field describing the chemical nature of the active site pocket, and the GRID algorithm[43] uses a grid based approach to aid in molecular docking and has been adopted by other docking programs (e.g., GLIDE).[44,45]

By the very nature of the MT method we can readily generate "heat maps" describing the chemical nature of the grid points created in the MT method. These can be used to describe pairwise interactions between the grid point and the protein environment (e.g., amide hydrogen with a carbonyl oxygen) or interactions can be lumped into nonpolar or polar interactions describing the aggregate collection of polar and nonpolar pairwise interactions. Not only does this describe the nature of the grid points but also indicates regions where specific atoms should be placed to optimize binding affinity.

In contrast to energy heat maps, the MT heat maps represent the probability-weighted interaction energy on each grid point. Knowledge-based data (i.e., the probability distribution along the interacting distance) will affect the location of both unfavorable and favorable interactions depending on the nature of the system. Moreover, energy gradient maps can be generated based on heat map energy calculations, which facilitates ligand docking as described below.

Extracting Structure from the MT Method. The advantage of the MT method is that the energy and the free

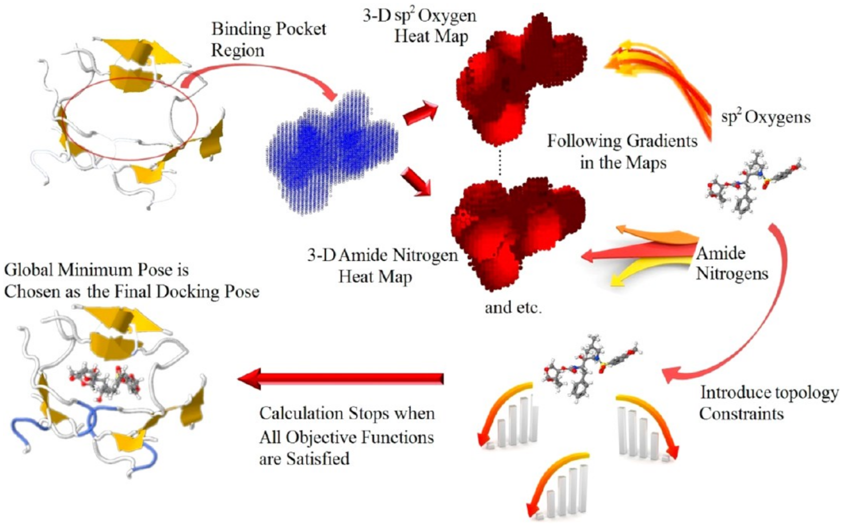


Figure 3. The MT energy maps optimization mechanism to derive the final docking pose in one protein−ligand complex.

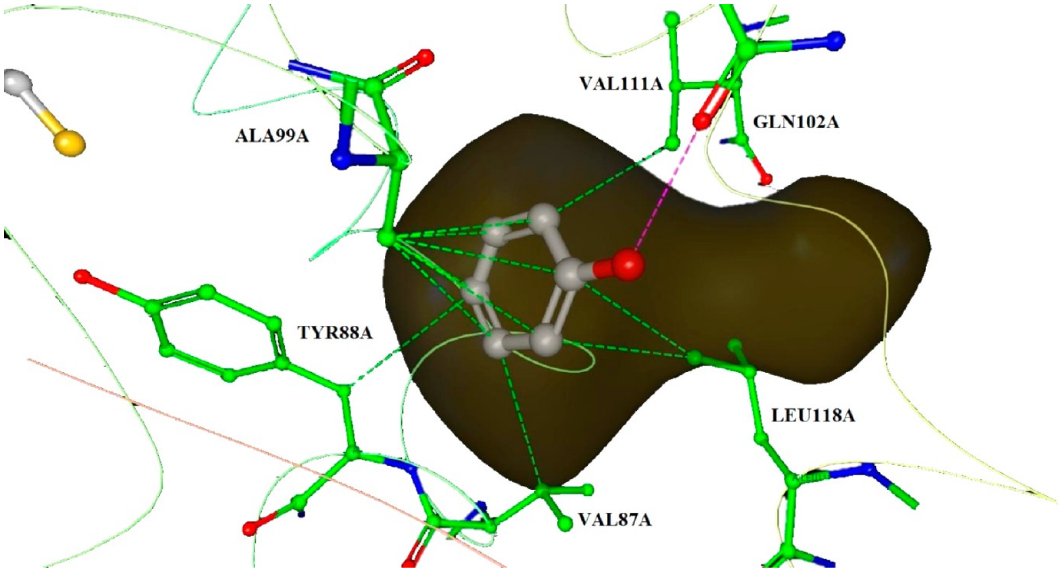


Figure 4. Contact map of the 1LI2 protein−ligand complex binding region. Hydrophobic contacts are shown as green dashed lines, and the one hydrogen bond is shown as a pink dashed line. The binding pocket cavity is defined by the dark region.

energy (when introducing the partition function) can be derived using only atomic linkage information coupled with the databases of atom pairwise distance distributions along with their corresponding energies. This offers us a new approach for protein−ligand docking without resorting to exhaustive pose sampling. Our initial efforts utilized the frozen receptor model, but the incorporation of receptor flexibility is, in principle, straightforward and will be explored in the future.

In a docking exercise, the best-docked pose for the ligand is usually obtained based on the highest binding affinity, which can be regarded as an optimization problem. With the frozen binding pocket approximation, generation of the "best" docking pose is a gradient optimization of the ligand atoms within the binding pocket, subject to the constraints of the ligand topology.

Molecular internal linkages including bond lengths and angles only slightly deviate from their optimized values, making them constraints in the ligand energy optimization within the binding pocket. These ligand atom connectivities reduce the dimensionality of the problem in that atomic collections that do not have the correct connectivity are eliminated from further consideration. On the other hand, energies of the torsions and long-range interactions between ligands and proteins vary over comparatively large distance ranges and, thereby, are regarded as the objective functions. Hence, in order to do the optimization we need to obtain the first and second derivatives of the ligand torsion and the protein−ligand long-range interaction partition functions (shown in eq 43 and 44), which can be readily seen in the gradient maps of the individual atom type pairs.

$$\frac{dc(r)}{dr} = \frac{d(q(r)\cdot z(r))}{dr} = \frac{d(q(r))}{dr}\cdot z(r) + q(r)\cdot\frac{d(z(r))}{dr} = 0 \tag{43}$$

$$\frac{d^2c(r)}{dr^2} = \frac{d^2(q(r)\cdot z(r))}{dr^2} = \frac{d^2(q(r))}{dr^2}\cdot z(r) + 2\left(\frac{d(q(r))}{dr}\cdot\frac{d(z(r))}{dr}\right) + z(r)\cdot\frac{d^2(q(r))}{dr^2} > 0 \tag{44}$$

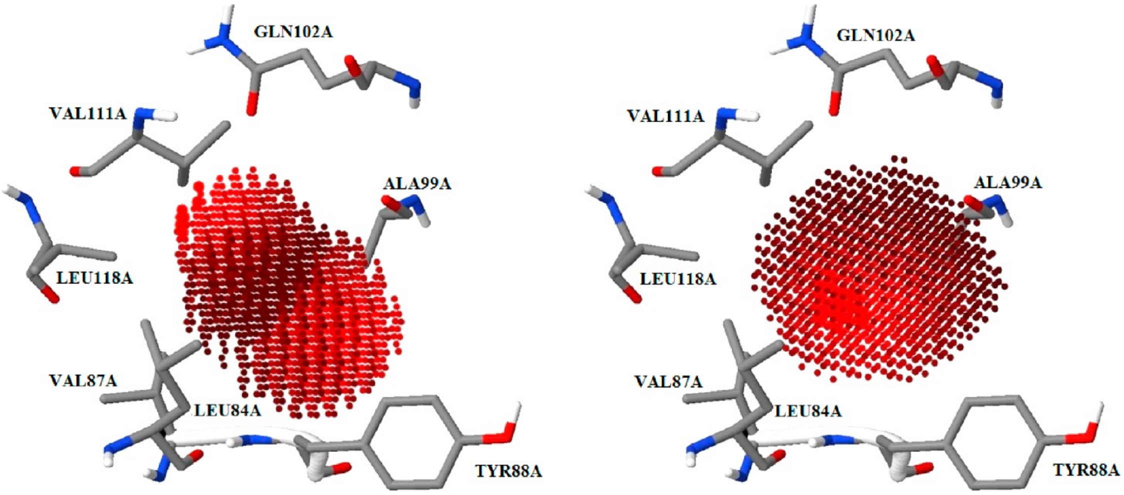


Figure 5. Heat maps for sp^3 oxygen (left) and aromatic carbon (right). Grid points with lighter color indicate energetically favorable locations for certain atom types within the binding pocket.

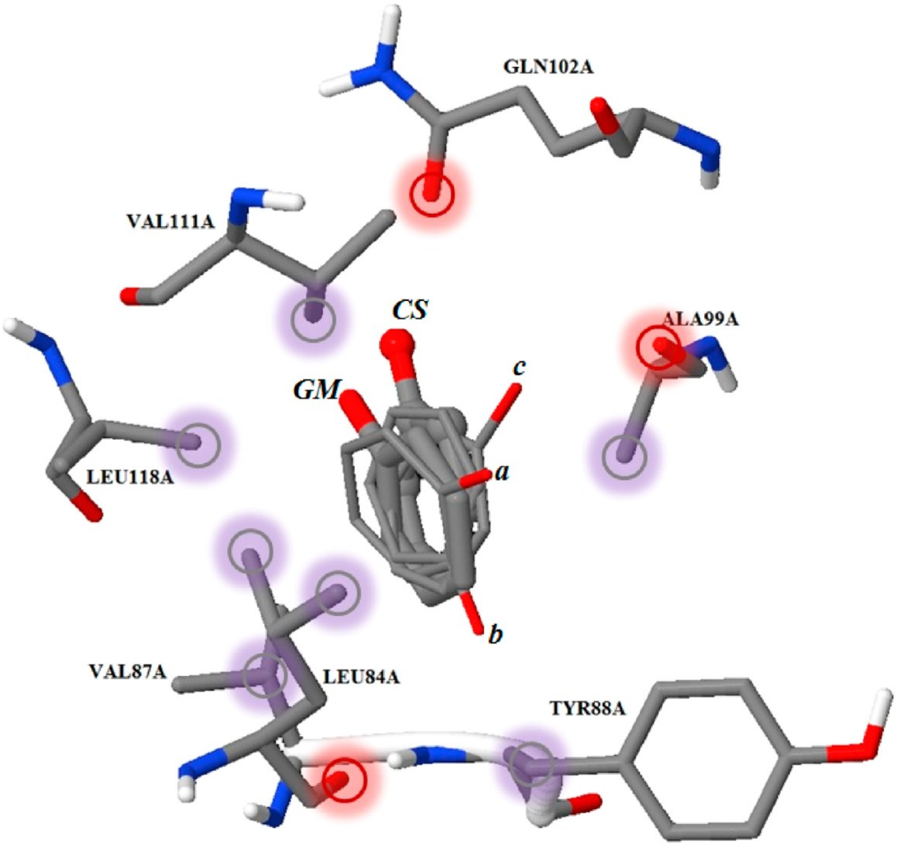


Figure 6. In the binding pocket of protein−ligand complex 1LI2, the ligand crystal structure (marked as *CS*) is shown as a ball and stick figure, the global minimum pose (marked as *GM*) is shown as a stick along with the three other identified local minimum (marked as *a*, *b*, and *c*). Red bubbles on the protein atoms indicate potential contacts with the ligand sp^3 oxygen. Gray bubbles on the protein atoms indicate potential contacts with aromatic carbons.

Optimum ligand atom locations are obtained when the calculation satisfies the minimum values for all the objective functions (ligand torsions and protein−ligand long-range interactions) and all ligand bonds and angle constraints.

In our optimization algorithm we obtain numerical derivatives of the probability distribution and analytical derivatives for the energy expression via pairwise partial derivatives of the modified Lennard-Jones potentials used in KECSA.[35] With the ligand topology and first and second derivatives we used a Newton−Raphson algorithm to optimize the ligand in the pocket. A nice feature of this method is we can identify both the lowest free energy binding mode along with all other possible local minima with higher free energies. Moreover, we can extract saddle point and higher-order transitions describing the connectivity between the local minima, but these are quite numerous and, hence, complicated and will not be discussed here in detail.

Table 2. RMSD values (Å) and binding scores (pK_d) of the global and local minima

	RMSD (Å)	binding affinity (pK_d)
global minimum	0.937	3.329
local minimum *a*	2.667	2.255
local minimum *b*	2.839	2.975
local minimum *c*	2.342	3.299

Table 3. RMSD Values (Å) and Standard Deviations of RMSD (Å) from the MT Heatmap Docking Results, Compared With Glide SP and XP Results

	RMSD (Å)	RMSD standard deviation (Å)
MT KECSA	1.97	1.27
Glide SP	2.07	2.72
Glide XP	1.87	2.01

Figure 3 introduces the process of the heatmap docking. To illustrate the method in detail we will touch on just one example whose structure is 1LI2. We have also carried out heatmap docking against the previously introduced test set of 795 protein ligand complexes, which will be summarized below.

The protein–ligand complex with PDB ID 1LI2 is used as an example to illustrate in detail the process of heatmap docking. 1LI2 is a T4 Lysozyme mutant bound to phenol with a modest binding affinity of 4.04 (pK_d). The binding pocket region is larger than the small phenol ligand structure (see Figure 4), potentially allowing several ligand poses that represent local minima. On the other hand, phenol, as the ligand, has a simple enough structure to clearly show the differences in protein–ligand contacts between low energy poses.

Judging from the crystal structure, phenol forms a hydrogen bond with GLN102A and several hydrophobic contacts with VAL87A, TYR88A, ALA99A, VAL111A, and LEU118A in the binding pocket (shown in Figure 4).

There are two atom types (sp^3 oxygen and aromatic carbon atoms) in phenol. Based on the MT KECSA calculation, heat maps for both of the atom types within the binding pocket can be generated (Figure 5).

The heatmap docking program then generated one sp^3 oxygen and six aromatic carbons to their optimized position following the gradients on their corresponding energy heatmaps while satisfying the linkage constraints of phenol. As a result, together with the energetic global minimum ligand pose (GM), three more local minimum poses (pose *a*, *b*, and *c*) were generated using the heatmap docking method. RMSD values (Å) and binding scores (pK_d) are shown in Table 2.

As can be seen in Figure 6, the GM pose slightly deviates from the crystal structure (*CS*) because of the adjustment of the hydrogen-bond distance between the phenol oxygen and the

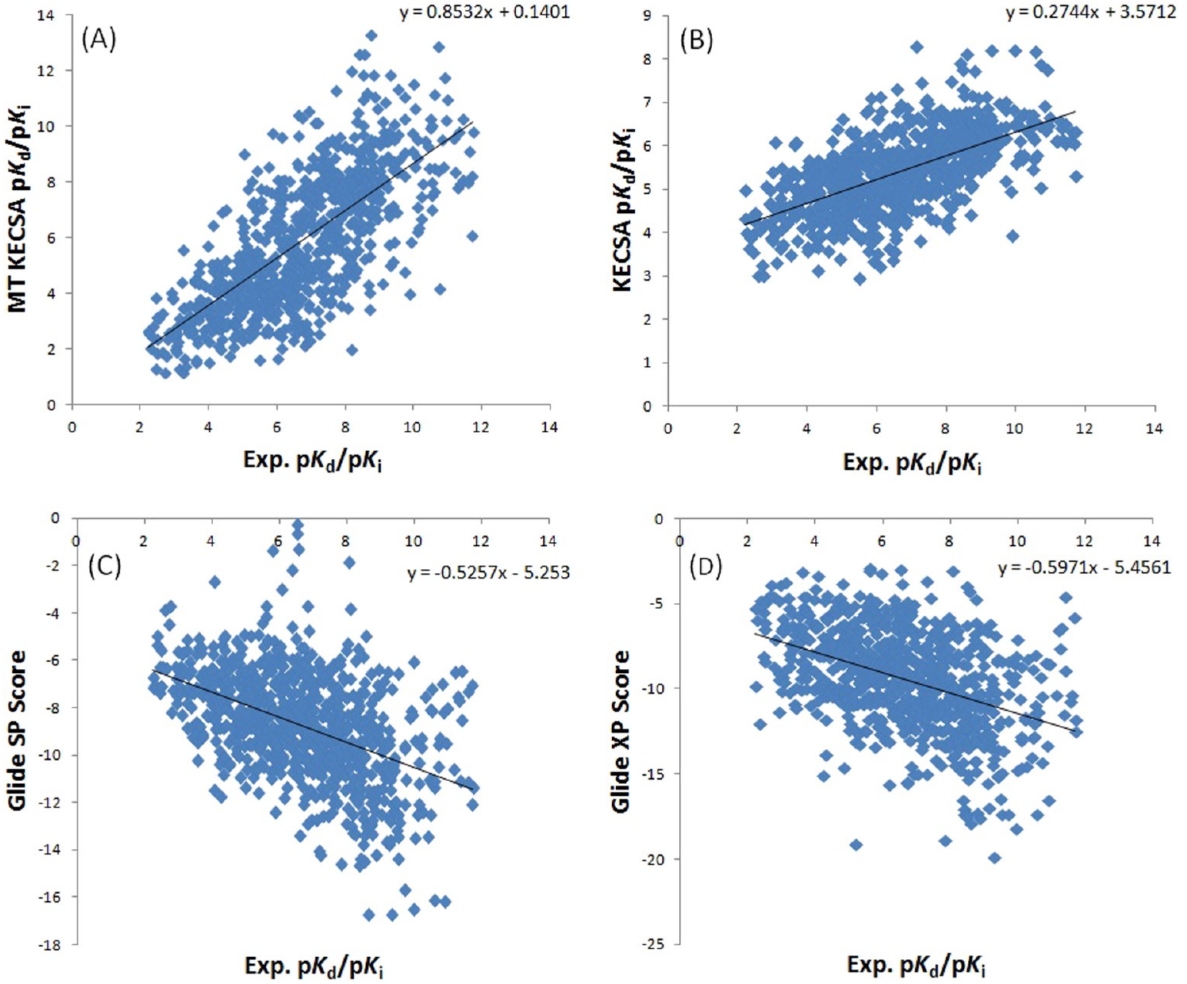


Figure 7. Plot of MT KECSA (A), the original KECSA model (B) calculated pK_d or pK_i values, Glide SP score (C) and Glide XP score (D) vs experimental pK_d or pK_i values.

sp^2 oxygen on GLN102A in the MT KECSA calculation. The phenol benzene ring balances the contacts with ALA99A and TYR88A on one side and the contacts with LEU118A, VAL87A, and LEU84A on the other. The local minimum pose *c* and *b* has close binding scores when compared to the *GM* pose. They form hydrogen bonds with different hydrogen acceptors (ALA99A backbone oxygen for pose *c* and LEU84A backbone oxygen for pose *b*) while maintaining very similar benzene ring locations. The local minimum pose *a* is trying to form a hydrogen bond with ALA99A backbone oxygen. However, the benzene ring of local minimum pose *a* is tilted toward the LEU118A, VAL87A, and LEU84A side chain carbons, weakening the hydrogen bond with the ALA99A backbone oxygen with the net result being a reduction in binding affinity.

Further Validation of the MT Method. Using the MT method a further docking study was carried out on the test set of 795 protein−ligand complexes. In addition, in order to better evaluate the performance of MT scoring and heatmap docking, we also carried out a Glide scoring and docking study for comparison.[44−46]

For the Glide docking and scoring study, protein structures were prepared with the Protein Preparation Wizard utility of the Schrodinger 2013-2 Suite using Epik state penalties.[44,47] Protonation states were assigned using PROPKA at pH 7.0.[48,49] Hydrogen positions were optimized with OPLS 2005.[50,51] All the crystal waters were removed. Docking runs using the Glide version 5.9 Standard Precision (SP) and Extra Precision (XP) algorithms followed the preparation step.

The MT heatmap docking generated an average RMSD of 1.97 Å with a 1.27 Å standard deviation when compared to the protein−ligand crystal structure, while Glide SP docking generated an average RMSD of 2.07 Å with a 2.72 Å standard deviation, Glide XP docking generated an average RMSD of 1.87 Å with a 2.01 Å standard deviation against the same set (Table 3). The result for each individual system studied herein is given in the Supporting Information. Based on this test result, MT heatmap docking showed a comparable performance to Glide results, all of which have ∼2.0 Å RMSD results. However, the standard deviation of the pose RMSD generated using heatmap docking is in a more narrow range than what is seen using SP and XP Glide.

We also compared the binding affinity computed by the MT method (actually pK_d/pK_i values) with the Glide SP and XP scores. We show the Glide scores vs the experimental pK_d or pK_i values across the 795 protein−ligand complexes in the test set in Figure 7, together with the calculated pK_d or pK_i values from MT KECSA and the original KECSA model vs the experimental pK_d or pK_i values of the same test set for comparison. Due to the different scales used by MT KECSA and the Glide scores, the comparison only includes Pearson's *r* and Kendall τ values. MT KECSA and the original KECSA, as discussed above, generated Pearson's *r*'s of 0.72 and 0.62, Kendall τ's of 0.53 and 0.46. Glide SP yields a Pearson's *r* of 0.46 and a Kendall τ of 0.52, while Glide XP yields a Pearson's *r* of 0.42 and a Kendall τ of 0.29. Overall, we conclude that MT KECSA shows advantages over scoring with Glide.

CONCLUSIONS

The prediction of the free energies associated with a wide range of biological problems remains a very daunting task. Balancing the sampling of the relevant degrees of freedom with accurate energy computation makes this a very difficult problem. Herein we describe a new approach that in one-shot samples, all the relevant degrees of freedom in a defined region directly affording a free energy without resorting to ad hoc modeling of the entropy associated with a given process. This is accomplished by converting ensemble assembly from a 3-D to a 1-D problem by using pairwise energies of all relevant interactions in a system coupled with their probabilities. We call this approach the MT method, and in conjunction with KECSA potential function we demonstrated the application of this approach to protein ligand pose and binding free energy prediction. The resultant MT-KECSA model out-performs the original KECSA model showing the power of this approach. Importantly, the present MT model can be applied to any pairwise decomposable potential which will allow us to attack a wide range of problems in biology including the validation of potential functions.

ASSOCIATED CONTENT

Ⓢ Supporting Information

A description of a methane−butane system as an example that illustrates, in detail, the binding free energy calculation using the MT method with the KECSA energy function. An introduction to a fast approximate algorithm for matrix multiplication in MT computation as well as the predicted pK_d or pK_i vs experimental pK_d or pK_i for our test set and the heatmap docking RMSD result against the same test set. This information is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kmerz1@gmail.com

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank the NIH (GM044974 and GM066859) for supporting the present research. Drs. Erik Deumens and John Faver are also acknowledged for numerous helpful discussions.

REFERENCES

(1) Jorgensen, W. L.; Tirado-Rives, J. Monte Carlo vs. molecular dynamics for conformational sampling. *J. Phys. Chem.* **1996**, *100*, 14508.

(2) Cancès, E.; Legoll, F.; Stoltz, G. Theoretical and numerical comparison of some sampling methods for molecular dynamics. *ESAIM: Math. Modell. Numer. Anal.* **2007**, *41*, 351.

(3) Gallicchio, E.; Levy, R. M. Advances in all atom sampling methods for modeling protein-ligand binding affinities. *Curr. Opin. Struct. Biol.* **2011**, *21*, 161.

(4) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409.

(5) Limongelli, V.; Marinelli, L.; Cosconati, S.; La Motta, C.; Sartini, S.; Mugnaini, L.; Da Settimo, F.; Novellino, E.; Parrinello, M. Sampling protein motion and solvent effect during ligand binding. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 1467.

(6) Clark, M.; Guarnieri, F.; Shkurko, I.; Wiseman, J. Grand Canonical Monte Carlo Simulation of Ligand−Protein Binding. *J. Chem. Inf. Model.* **2006**, *46*, 231.

(7) Okamoto, Y. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graphics Modell.* **2004**, *22*, 425.

(8) Hendlich, M.; Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **1990**, *216*, 167.

(9) Jones-Hertzog, D. K.; Jorgensen, W. L. Binding affinities for sulfonamide inhibitors with human thrombin using Monte Carlo simulations with a linear response method. *J. Med. Chem.* **1997**, *40*, 1539.

(10) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341.

(11) Kannan, S.; Zacharias, M. Simulated annealing coupled replica exchange molecular dynamics—An efficient conformational sampling method. *J. Struct. Biol.* **2009**, *166*, 288.

(12) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120.

(13) Schlick, T. Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules. *F1000 Biol. Rep.* **2009**, *1*, 51.

(14) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961.

(15) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21.

(16) Karney, C. F.; Ferrara, J. E.; Brunner, S. Method for computing protein binding affinity. *J. Comput. Chem.* **2005**, *26*, 243.

(17) Michel, J.; Essex, J. W. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 639.

(18) Sippl, M. J. Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859.

(19) Tuffery, P.; Derreumaux, P. Flexibility and binding affinity in protein-ligand, protein-protein and multi-component protein interactions: limitations of current computational approaches. *J. R. Soc. Interface* **2012**, *9*, 20.

(20) Fan, H.; Schneidman-Duhovny, D.; Irwin, J. J.; Dong, G.; Shoichet, B. K.; Sali, A. Statistical potential for modeling and ranking of protein-ligand interactions. *J. Chem. Inf. Model.* **2011**, *51*, 3078.

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

(22) 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.

(23) Huang, S. Y.; Zou, X. Advances and challenges in protein-ligand docking. *Int. J. Mol. Sci.* **2010**, *11*, 3016.

(24) 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.

(25) 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.

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

(27) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418.

(28) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de Novo design method based on simple, fast, and accutate free energy estimate. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733.

(29) 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.

(30) Zheng, Z.; Merz, K. M., Jr. Ligand identification scoring algorithm (LISA). *J. Chem. Inf. Model.* **2011**, *51*, 1296.

(31) Benson, M. L.; Faver, J. C.; Ucisik, M. N.; Dashti, D. S.; Zheng, Z.; Merz, K. M., Jr. Prediction of trypsin/molecular fragment binding affinities by free energy decomposition and empirical scores. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 647.

(32) Faver, J. C.; Zheng, Z.; Merz, K. M., Jr. Model for the fast estimation of basis set superposition error in biomolecular systems. *J. Chem. Phys.* **2011**, *135*, 144110.

(33) Faver, J. C.; Zheng, Z.; Merz, K. M., Jr. Statistics-based model for basis set superposition error correction in large biomolecules. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7795.

(34) Faver, J. C.; Yang, W.; Merz, K. M., Jr. The Effects of Computational Modeling Errors on the Estimation of Statistical Mechanical Variables. *J. Chem. Theory Comput.* **2012**, *8*, 3769.

(35) Zheng, Z.; Merz, K. M., Jr. Development of the Knowledge-Based and Empirical Combined Scoring Algorithm (KECSA) To Score Protein−Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53*, 1073.

(36) 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.

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

(38) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1.

(39) Leis, S.; Zacharias, M. Efficient inclusion of receptor flexibility in grid-based protein-ligand docking. *J. Comput. Chem.* **2011**, *32*, 3433.

(40) 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.

(41) Wu, G.; Robertson, D. H.; Brooks, C. L., 3rd; Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24*, 1549.

(42) Kubinyi, H. Comparative molecular field analysis (CoMFA). *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*; Wiley-VCH: Weinheim, 2008; p 1555.

(43) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849.

(44) 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. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739.

(45) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750.

(46) 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.

(47) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221.

(48) Rostkowski, M.; Olsson, M. H.; Søndergaard, C. R.; Jensen, J. H. Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. *BMC Struct. Biol.* **2011**, *11*, 6.

(49) Olsson, M. H. M.; Søndergard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Resi-

dues in Empirical pKa predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525.

(50) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

(51) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B.* **2001**, *105*, 6474.