A Conformational Factorisation Approach for Estimating the Binding Free Energies of Macromolecules

Kenji Mochizuki,*[a] Chris S. Whittleston,[b] Sandeep Somani,[b] Halim Kusumaatmaja,[b] and David J. Wales*[b]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

We present a conformational factorization approach. The theory is based on a superposition partition function, where the partition function is written as a sum over contributions from local minima. The factorisation greatly reduces the number of minima that need to be considered, by employing the same local configurations for groups that are sufficiently distant from the binding site. The theory formalises the conditions required to analyse how our definition of the binding site region affects the free energy difference between the apo and holo states. We employ basin-hopping parallel tempering to sample minima that contribute significantly to the partition function, and calculate the binding free energies within the harmonic normal mode approximation. A further significant gain in efficiency is achieved using a recently developed local rigid body framework in both the sampling and the normal mode analysis, which reduces the number of degrees of freedom. We benchmark this approach for human aldose reductase (PDB code 2INE). When varying the size of the rigid region, the free energy difference converges for factorisation of groups at a distance of 14 Å from the binding site, which corresponds to 80% of the protein being locally rigidified. This approach is likely to be useful for estimating the binding free energy of protein-ligand complexes.

1 Introduction

Predicting binding affinity between two non-covalently bound molecules is a challenging problem in molecular science. Calculating binding affinities using atomistic simulations can provide detailed molecular level insights into molecular recognition, and help inform fields such as structure-based drug design[1–3] and self-assembly.[4,5]. For instance, an accurate and efficient method for predicting protein-ligand binding free energy can help screen a library of candidate compounds against a protein target, or assist in lead optimisation, by predicting the impact of chemical modifications. Hence this is an active field for the computational drug design community.[1,6,7]

A broad class of methods for computing protein-ligand binding docks the ligand into the binding pocket and uses a scoring function to estimate the binding affinity[6,7]. The scoring functions have explicit terms to model various contributions to the binding free energy, such as the hydrophobic effect, hydrogen-bonding, and further entropic contributions, which are usually fitted to experimental binding data. This docking and scoring approach is fast, but may not be accurate, due to training set bias and an approximate treatment of conformational entropy.

An alternative class of methods employs atomistic force fields to model the interatomic and intermolecular interactions. To describe protein-ligand binding the energy function is typically taken to be an empirical form, either with explicit water molecules, or an implicit solvent model. The AMBER forcefield[8] is employed in the present study.

A range of simulation methods have previously been developed to compute binding free energies using force field energy models and molecular dynamics (MD) or Monte Carlo (MC) simulations. Alchemical methods, where atoms of one ligand are transformed to those for another ligand, are used to compute relative binding affinity. Thermodynamic integration[9] and free energy perturbation[10–12] have been employed for alchemical free energy simulations. In another approach, the absolute free energy of binding is computed by equilibrium[13,14] or non-equilibrium simulations[15,16] along a physical pathway between the free and the bound ligand. These methods are formally rigorous, but are computation-

† Electronic Supplementary Information (ESI) available: The maximum rotation amplitudes for each amino acid chain group (Table S1), the average CPU time for diagonalisation of the Hessian matrix (Fig. S1), the potential energy correlation between in vacuum and implicit solvent (Fig. S2), and Hessian expression in local rigid body coordinates are available as Supporting Information. See DOI: 10.1039/b000000x/

[a] *School of Physical Sciences, The Graduate University for Advanced Studies (SOKENDAI), Myodaiji, Okazaki 444-8585, Japan. Tel: +81 564 55 7394; E-mail: kmochi@ims.ac.jp*
[b] *University Chemical Laboratories, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK.*

‡ Additional footnotes to the title and authors can be included e.g. 'Present address:' or 'These authors contributed equally to this work' as above using the symbols: ‡, §, and ¶. Please place the appropriate symbol next to the author's name and include a `\footnotetext` entry in the the correct place in the list.

ally expensive due to sampling limitations in MD or MC simulations of proteins. Another class of methods, including Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA)[17–20], Linear Interaction Energy (LIE),[21–24] and others,[25] rely on MD simulation of only the free and bound states. These endpoint methods are relatively less expensive than pathway methods, since intermediate states are not considered, but still require adequate MD sampling of the end states, which can be challenging for protein sized systems. On the whole, current physics-based methods are computationally much more expensive than docking and scoring based methods and therefore have limited in applications such as virtual screening.

The force field-based methods are potentially more accurate than docking and scoring approaches, as they have been developed to account for explicit intermolecular interactions and are typically fitted using statistical mechanical theories. However, these methods can be computationally expensive, since MD or MC simulations are easily trapped in local minima of the potential or free energy surface for relatively long time scales.

The superposition approach provides an alternative formulation for global thermodynamics within the energy landscape framework.[26–28] Here, the partition function is written as the sum of contributions from the catchment basins[29] of local potential energy minima.[30–32] The contribution of each minimum can be estimated using the harmonic approximation, possibly with anharmonic[33] or quantum[34] corrections. To apply this procedure to calculate a binding free energy we can evaluate the free energy of the complex and the free molecules separately from databases of local minima for each species. This approach to binding free energy calculations is employed in the mining minima algorithm,[35] which has been successfully applied to various biomacromolecular systems, especially small host-guest systems. Benchmark superposition calculations for atomic and molecular clusters show that the energy landscape approach can be much faster than MD or MC based methods, especially for cases of broken ergodicity,[36,37,37–39] since the superposition partition function is explicitly ergodic.

To apply the superposition method for large systems requires appropriate sampling, because the number of local minima increases exponentially with system size.[40,41] A new method has recently been described to implement such sampling systematically, and was applied successfully to atomic clusters.[42] Alternatively, the mining minima method has been extended to larger protein-ligand systems[43] by focusing the calculation on regions around the binding pocket. For example, in Ref.[43], protein atoms were partitioned into three layers of different thickness with respect to the distance from ligand atoms. Atoms in the 7 Å layer closest to the ligand were free to move, while those in the middle layer of thickness 5 Å were fixed. Atoms in the outermost layer were deleted.

In the present contribution, we present a method conceptually similar to mining minima, but with key differences in the implementation, which aim to improve the accuracy and sampling efficiency. We again partition the protein atoms into three layers according to distance from the bound ligand. Atoms in the 'inner' region, adjacent to the ligand, are unconstrained, while those in the 'intermediate' region are treated using the local rigid body framework.[44] All atoms in the 'outer' layer were grouped as one rigid body, but their contributions to the potential energy of the system are retained. The local rigid body framework is used to reduce the number of degrees of freedom, both in sampling minima and in the calculation of normal mode frequencies. All ligand atoms are fully flexible.

To benchmark the procedure we systematically increase the radius defining the innermost unconstrained region until the binding free energy converges. The key idea is that contributions from minima corresponding to alternative conformations of groups that are sufficiently distant from the binding site are expected to cancel between the free protein and the complex. Hence, we only need to sample consistent conformations for these degrees of freedom. The theory, described in §2.1, therefore corresponds to a factorisation of the partition functions for the protein, ligand, and complex. We therefore refer to the method as a *factorised superposition approach* (FSA).

We apply the FSA procedure to compute the binding free energy for human aldose reductase (5113 atoms) and one of its inhibitors, phenyl acetic acid (PAC). Human aldose reductase is an NADPH-dependent oxidoreductase, which catalyses the reduction of a variety of aldehydes and carbonyls, including monosaccharides. It is primarily known for catalysing the reduction of glucose to sorbitol, the first step in the polyol pathway of glucose metabolism.[45]

The next section describes the theory underlying the factorisation procedure, the calculation of approximate free energies, local rigidification, and the sampling of local minima. We then describe the system setup for aldose reductase in §3, and discuss the conditions for convergence. We find that a flexible region of 14 Å, corresponding here to rigidification of about 80% of the protein, is required to obtain a converged binding free energy.

2 Methodology

2.1 Factorised Superposition Approach

We wish to estimate the binding free energy or the free energy change, ΔF, involved in forming a complex AB from noncovalent association of two molecules A and B. The standard free energy difference of this reaction is given by[46–48]

$$e^{-\beta\Delta F^\circ} = \frac{C^\circ}{8\pi^2}\frac{Z_{AB}}{Z_A Z_B}, \qquad (1)$$

where Z_X is the configurational part of the single-molecule partition function of species $X \in \{A, B, AB\}$, C° is the standard concentration, and $\beta = 1/kT$, with k the Boltzmann constant and T the temperature. The above expression is derived using classical thermodynamics so that the momentum factors in the partition function of the free molecules and the bound complex cancel. The translational and rotational degrees of freedom have been integrated out from the configurational integral; see references[47,48] for a detailed derivation of the above expression.

We compute the partition functions [henceforth referring to the configurational integrals in Eq. (1)] using the superposition approach,[26,28,30] where each Z_X is written as a sum of contributions from local minima of the potential energy surface. In this section, we describe the FSA framework, an extension of the superposition approach, which facilitates calculation of an approximate binding free energy from a subset of local minima.

The FSA framework was developed in order to provide a route to protein-ligand binding energies, where the number of relevant local minima becomes problematic for the standard superposition approach. To limit the number of minima, we assume that the contributions of analogous alternative conformations of functional groups that are sufficiently distant from the binding region cancel out. This scheme can be formalised by thinking in terms of the possible local conformations of distinct parts of the protein, such as backbone and side chain geometries. As a further simplification, we consider molecules that are not rotating or translating, and focus on the vibrational partition function for each local minimum.

To index the local minima we consider the possible conformations for each part of the molecule, and assume that we can identify them independently of the conformations adopted by the rest of the system (a factorisation). Each minimum can then be represented by a vector, $\mathbf{x} = (x_1, x_2, ...)$, where each component x_i for $i = 1, 2, 3, \ldots$ identifies the local conformation of a region i. Some conformations of one region will preclude conformations of other regions, so the permitted combinations of x_i are restricted, which prevents further factorisation in general. We now identify local minima corresponding to AB as $\mathbf{x}_{AB} = (\mathbf{x}_A, \mathbf{x}_B)$. If certain local conformations are only possible in the complex, then the corresponding geometries in the separate A and B molecules are presumably high in energy, but can still be included in the possible conformations enumerated by $\mathbf{x}_A$ and $\mathbf{x}_B$. To analyse the most plausible cancellation of contributions from alternative conformations that lie outside the binding region (the factorisation) we now assume that $\mathbf{x}_{AB}$ can be partitioned into two sets as $\mathbf{x}_{AB} = (\mathbf{u}_{AB}, \mathbf{v}_{AB})$, as shown in Fig. 1. This formalism is designed to reflect our intuition that some local conformations are common to each molecule, while others associated with the binding region are not. The conformations collected in the $\mathbf{v}_{AB}$ set therefore correspond to

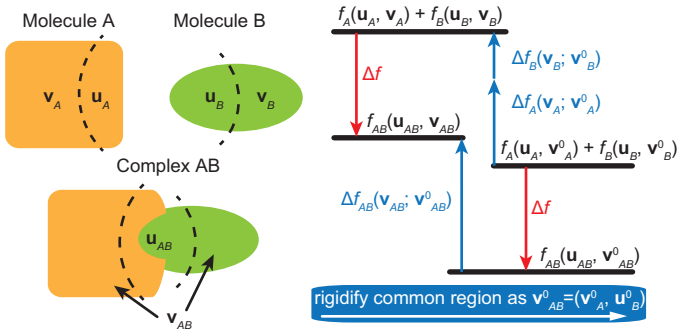

Fig. 1 (Left) Schematic representation of the conformational indexing vectors for molecules A and B, and for the complex AB. (Right) Schematic representation for the free energies of one local minimum of A, B and AB, representing $f_A(\mathbf{u}_A, \mathbf{v}_A)$, $f_B(\mathbf{v}_B, \mathbf{u}_B)$ and $f_{AB}(\mathbf{v}_{AB}, \mathbf{u}_{AB})$. The difference, Δf, does not change if the shift corresponding to different $\mathbf{v}_{AB}$, $\Delta f_{AB}(\mathbf{v}_{AB}; \mathbf{v}^0_{AB})$, is independent of $\mathbf{u}_{AB}$ [Eq. (6)] and is additive for $\Delta f_A(\mathbf{v}_A; \mathbf{v}^0_A)$ and $\Delta f_B(\mathbf{v}_B; \mathbf{v}^0_B)$ [Eq. (7)].

local structure that is identifiable in each of A, B and AB for all the conformations specified by the vector $\mathbf{u}_{AB}$. The corresponding regions in the separate A and B molecules are written as $\mathbf{u}_A$, $\mathbf{v}_A$, $\mathbf{u}_B$, and $\mathbf{v}_B$, and we assume that all possible conformations specified by $\mathbf{v}_A$ and $\mathbf{v}_B$ are also available in $\mathbf{v}_{AB}$ for any $\mathbf{u}_{AB}$.

A significant simplification is possible if we need only consider a consistent reference conformation, $\mathbf{v}^0_A$ and $\mathbf{v}^0_B$, respectively, for each group collected in $\mathbf{v}_A$ and $\mathbf{v}_B$. In fact, this choice produces a combinatorial reduction in the number of minima that may need to be sampled. The analysis that follows defines quantitative conditions under which this simplification will be valid. Furthermore, our local rigidification procedure[44] provides an ideal framework for implementing this approach, and enables us to determine a minimal set of states for estimating free energies of binding.

The partition function for separate A and B molecules factorises and we therefore consider

$$\begin{aligned} Z_A &= \sum_{\mathbf{u}_A}\sum_{\mathbf{v}_A} z_A(\mathbf{u}_A, \mathbf{v}_A) e^{-\beta V_A(\mathbf{u}_A, \mathbf{v}_A)} \\ &= \sum_{\mathbf{u}_A} z_A(\mathbf{u}_A, \mathbf{v}^0_A) e^{-\beta V_A(\mathbf{u}_A, \mathbf{v}^0_A)} \\ &\times \sum_{\mathbf{v}_A} \frac{z_A(\mathbf{u}_A, \mathbf{v}_A)}{z_A(\mathbf{u}_A, \mathbf{v}^0_A)} e^{-\beta[V_A(\mathbf{u}_A, \mathbf{v}_A) - V_A(\mathbf{u}_A, \mathbf{v}^0_A)]}, \end{aligned} \quad (2)$$

where $z_A(\mathbf{u}_A, \mathbf{v}_A)$ and $V_A(\mathbf{u}_A, \mathbf{v}_A)$ are the vibrational partition function and potential energy for minimum $(\mathbf{u}_A, \mathbf{v}_A)$. The sum is over all local minima of A, identified via their $\mathbf{u}_A$ and $\mathbf{v}_A$ conformational assignment. Next, we define a free energy shift, $\Delta f_X(\mathbf{u}_X, \mathbf{v}_X; \mathbf{v}^0_X)$, as the free energy difference between a given minimum $(\mathbf{u}_X, \mathbf{v}_X)$ and the corresponding reference

$(\mathbf{u}_X, \mathbf{v}_X^0)$,

$$e^{-\beta \Delta f_X(\mathbf{u}_X,\mathbf{v}_X;\mathbf{v}_X^0)} \equiv \frac{z_X(\mathbf{u}_X,\mathbf{v}_X)}{z_X(\mathbf{u}_X,\mathbf{v}_X^0)} e^{-\beta[V_X(\mathbf{u}_X,\mathbf{v}_X)-V_X(\mathbf{u}_X,\mathbf{v}_X^0)]}. \quad (3)$$

The partition function for each molecule can then be written as

$$Z_X = \sum_{\mathbf{u}_X} e^{-\beta f_X(\mathbf{u}_X,\mathbf{v}_X^0)} \sum_{\mathbf{v}_X} e^{-\beta \Delta f_X(\mathbf{u}_X,\mathbf{v}_X;\mathbf{v}_X^0)}, \quad (4)$$

$$\text{where} \qquad e^{-\beta f_X(\mathbf{u}_X,\mathbf{v}_X^0)} \equiv z_X(\mathbf{u}_X,\mathbf{v}_X^0) e^{-\beta V_X(\mathbf{u}_X,\mathbf{v}_X^0)}.$$

The ratio of partition functions in Eq. (1) becomes

$$\begin{aligned} \frac{Z_{AB}}{Z_A Z_B} &\equiv e^{-\beta \Delta F} \\ &= \sum_{\mathbf{u}_{AB}} e^{-\beta f_{AB}(\mathbf{u}_{AB},\mathbf{v}_{AB}^0)} \sum_{\mathbf{v}_{AB}} e^{-\beta \Delta f_{AB}(\mathbf{u}_{AB},\mathbf{v}_{AB};\mathbf{v}_{AB}^0)} \\ &/ \sum_{\mathbf{u}_A} e^{-\beta f_A(\mathbf{u}_A,\mathbf{v}_A^0)} \sum_{\mathbf{v}_A} e^{-\beta \Delta f_A(\mathbf{u}_A,\mathbf{v}_A;\mathbf{v}_A^0)} \\ &/ \sum_{\mathbf{u}_B} e^{-\beta f_B(\mathbf{u}_B,\mathbf{v}_B^0)} \sum_{\mathbf{v}_B} e^{-\beta \Delta f_B(\mathbf{u}_B,\mathbf{v}_B;\mathbf{v}_B^0)}. \end{aligned} \quad (5)$$

As noted above, we require the $\mathbf{v}_{AB} = (\mathbf{v}_A, \mathbf{v}_B)$ conformations to appear in both the separate molecules and in the complex, and they must be identifiable for each minimum specified by different conformations in $\mathbf{u}_{AB} = (\mathbf{u}_A, \mathbf{u}_B)$. Next we introduce two assumptions, schematically described in Fig. 1, to simplify Eq. (5). First, we assume that the shifts with respect to the reference conformation in the free energies, Δf_X, are independent of $\mathbf{u}_X$, if the partition of the complex is chosen appropriately. That is,

$$\Delta f_X(\mathbf{u}_X,\mathbf{v}_X;\mathbf{v}_X^0) \approx \Delta f_X(\mathbf{v}_X;\mathbf{v}_X^0) \quad \forall\, \mathbf{u}_X. \quad (6)$$

Second, we assume that, for a given minimum, the shifts in energy and vibrational frequencies relative to the reference conformation, $\mathbf{v}_{AB}^0 = (\mathbf{v}_A^0, \mathbf{v}_B^0)$, are the same in the complex and the separated molecules for all $\mathbf{u}_{AB}$, that is,

$$\Delta f_{AB}(\mathbf{v}_{AB};\mathbf{v}_{AB}^0) \approx \Delta f_A(\mathbf{v}_A;\mathbf{v}_A^0) + \Delta f_B(\mathbf{v}_B;\mathbf{v}_B^0) \quad \forall\, \mathbf{u}_{AB}. \quad (7)$$

Note that by construction, every $\mathbf{v}_{AB}$ conformation in the numerator of Eq. (5) can be associated with a product of terms from the $\mathbf{v}_A$ and $\mathbf{v}_B$ sums in the denominator. Therefore, using Eq. (7), the factors with summations over the common region in Eq. (5) cancel, giving the final result

$$e^{-\beta \Delta F} \approx \frac{\sum_{\mathbf{u}_{AB}} e^{-\beta f_{AB}(\mathbf{u}_{AB},\mathbf{v}_{AB}^0)}}{\sum_{\mathbf{u}_A} e^{-\beta f_A(\mathbf{u}_A,\mathbf{v}_A^0)} \times \sum_{\mathbf{u}_B} e^{-\beta f_B(\mathbf{u}_B,\mathbf{v}_B^0)}}, \quad (8)$$

defines the free energy of a specific minimum. We must therefore sum over members of the $\mathbf{u}_{AB}$ minima and over local minima corresponding to all conformations of A and B in the same regions, with a Boltzmann weighting. A consistent set of local reference conformations $\mathbf{v}_{AB}^0$ must be used for the other regions corresponding to $\mathbf{v}_{AB}$.

In the present work, the common regions consisted of only the protein atoms (molecule A) while the ligand (molecule B) was treated as fully flexible, reducing Eq. (8) to

$$e^{-\beta \Delta F} \approx \frac{e^{-\beta F_{AB}(\mathbf{v}_{AB}^0)}}{e^{-\beta F_A(\mathbf{v}_A^0)} \times e^{-\beta F_B}}, \quad (9)$$

where

$$e^{-\beta F_X(\mathbf{v}_X^0)} \equiv \sum_{\mathbf{u}_X} e^{-\beta f_X(\mathbf{u}_X,\mathbf{v}_X^0)}, \; X \in A, AB,$$

and

$$e^{-\beta F_B} \equiv \sum_{\mathbf{u}_B} e^{-\beta f_B(\mathbf{u}_B)}$$

are the free energies for the free molecules and the complex. Note that the free energy of the protein and complex depend on the reference configuration of the common regions. Eq. (9) is the working equation for the applications considered below. Since we are primarily interested in the convergence of the binding free energy with respect to the FSA framework, we do not include the $8\pi^2/C^\circ$ prefactor from Eq. (1), and we treat all molecules in vacuum for this initial benchmarking.

Eq. (9) is quite intuitive, with consistent reference conformations selected for regions of the protein that interact only weakly with the binding site. The derivation defines the validity of this approximation. In particular, it is clear that sampling over a small number of local minima where the conformations in the weakly interacting region are not consistent would introduce systematic errors. For large systems the number of possible conformations will be combinatorial, and randomly chosen conformations are unlikely to be in correspondence.

A straightforward method for implementing Eq. (9) is to sample local minima with the common region constrained in the reference conformation. This sampling is accomplished here using the local rigidification framework[44]. Our strategy for testing Eq. (9) is to check the convergence of the binding free energy as we expand the unconstrained region specified by $\mathbf{u}$. As a cross-validation, the result should be independent of the reference conformations specified by $\mathbf{v}^0$.

2.2 Free Energy of Local Minima

In the harmonic approximation, the free energy, f, of a minimum is given by

$$e^{-\beta f} = \frac{e^{-\beta V_{\min}}}{(\beta h \bar{\nu})^\kappa}, \qquad \text{with} \qquad \bar{\nu} = \left(\prod_i \nu_i\right)^{1/\kappa}, \quad (10)$$

where h is Planck's constant, $V_{\min}$ is the potential energy of the minimum, and $\bar{\nu}$ is the geometric mean of the $\kappa = 3N-6$ vibrational normal mode frequencies. For a fully flexible molecule, $\kappa = 3N-6$ where N is the number of atoms. The number of vibrational modes is reduced if parts of the molecule are rigidified, as described in the next section. When applying the superposition formula we collect together the identical contributions for all permutation-inversion isomers of a given minimum, which corresponds to weighting $e^{-\beta f}$ by $1/o$, with o the order of the corresponding point group.[26,30,49] An additional factor that depends on the atomic composition of the system is needed to enumerate the distinct local minima precisely, but cancels from all thermodynamic quantities. Since the point group is C_1 for all the minima considered in the present work, $o = 1$. Eq. (10) also ignores overall translational and rotational contributions (Section 2.3.1 in Ref[47]), which were found to make a negligible contribution to the free energy differences of interest in the present study.

2.2.1 Normal Mode Analysis in the Local Rigid Body Framework. The cost of diagonalisation of the $3N \times 3N$ dimensional Hessian matrix required for calculating the normal mode frequencies for each minimum scales as $\mathcal{O}(N^3)$. The computational expense is reduced when we consider the Hessian corresponding to local rigidification. Since the ligand is treated as fully flexible, its normal mode frequencies are computed by diagonalising the standard all-atom Hessian.

We need to address two issues in order to perform a normal mode calculation with local rigidification. First, the Hessian matrix of second derivatives required for the normal mode analysis has dimension $3N$ for N atoms. Rigidification reduces the dimensionality, and corresponds to a projection of the degrees of freedom of the constrained atoms onto the rotational and translational degrees of freedom of the rigid bodies. Second, the moment of inertia tensor is generally not diagonal for the kinetic energy expressed in the local rigid body coordinates. Hence we need two steps to calculate the corresponding normal modes, as detailed below.

First we establish our notation, denoting the number of rigid bodies by N_{RB} and the number of unconstrained atoms by N_A. In the angle-axis representation[27,50] each rigid body I has six degrees of freedom: three representing the position of the centre of mass (translational degrees of freedom) $\mathbf{r}^I = \{r_1^I, r_2^I, r_3^I\}$, and three representing its orientation (rotational degrees of freedom) $\mathbf{p}^I = \{p_1^I, p_2^I, p_3^I\}$. For clarity, we employ capital letters for rigid bodies, and lower case for the sites in the rigid bodies. The coordinates of the sites, i, for rigid body I are denoted by $\mathbf{r}^I(i) = \{r_1^I(i), r_2^I(i), r_3^I(i)\}$, where

$$\mathbf{r}^I(i) = \mathbf{r}^I + \mathbf{S}^I \mathbf{x}^I(i); \quad i \in I. \tag{11}$$

We define $\mathbf{x}^I(i) = \{x_1^I(i), x_2^I(i), x_3^I(i)\}$ as the reference coordinates of the sites relative to the centre of mass of rigid body I, and $\mathbf{S}^I$ as the rotation matrix constructed from the rotational degrees of freedom $\{\mathbf{p}^I\}$ (in the angle-axis representation) that rotates rigid body I from its reference frame to its current orientation,

$$\mathbf{S}^I = \mathbf{I} + (1-\cos\theta^I)\widetilde{\mathbf{p}}^I\widetilde{\mathbf{p}}^I + \sin\theta^I\widetilde{\mathbf{p}}^I, \tag{12}$$

with $\mathbf{I}$ the identity matrix, $\theta^I = \left((p_1^I)^2 + (p_2^I)^2 + (p_3^I)^2\right)^{1/2}$ and $\widetilde{\mathbf{p}}^I$ the skew-symmetric matrix obtained from the rotation vector $\mathbf{p}^I$:

$$\widetilde{\mathbf{p}}^I = \frac{1}{\theta^I}\begin{pmatrix} 0 & -p_3^I & p_2^I \\ p_3^I & 0 & -p_1^I \\ -p_2^I & p_1^I & 0 \end{pmatrix}. \tag{13}$$

Using the above notation, the Hessian corresponding to local rigid body coordinates is given by

$$\frac{\partial^2 V}{\partial r_\alpha^I \partial r_\beta^J} = \sum_{i\in I}\sum_{j\in J}\frac{\partial^2 V}{\partial r_\alpha^I(i)\partial r_\beta^J(j)},$$

$$\frac{\partial^2 V}{\partial r_\alpha^I \partial p_\beta^J} = \sum_{i\in I}\sum_{j\in J}\sum_{a=1}^{3}\frac{\partial^2 V}{\partial r_\alpha^I(i)\partial r_a^J(j)}\left[\frac{\partial \mathbf{S}^J}{\partial p_\beta^J}\mathbf{x}^J(j)\right]_a,$$

$$\frac{\partial^2 V}{\partial p_\alpha^I \partial p_\beta^J} = \sum_{i\in I}\sum_{j\in J}\sum_{a=1}^{3}\sum_{b=1}^{3}\frac{\partial^2 V}{\partial r_b^I(i)\partial r_a^J(j)}\left[\frac{\partial \mathbf{S}^I}{\partial p_\alpha^I}\mathbf{x}^I(i)\right]_b\left[\frac{\partial \mathbf{S}^J}{\partial p_\beta^J}\mathbf{x}^J(j)\right]_a, \quad \text{for } I \neq J, \tag{14}$$

$$\frac{\partial^2 V}{\partial p_\alpha^I \partial p_\beta^I} = \sum_{i_1\in I}\sum_{i_2\in I}\sum_{a=1}^{3}\sum_{b=1}^{3}\frac{\partial^2 V}{\partial r_b^I(i_1)\partial r_a^I(i_2)}\left[\frac{\partial \mathbf{S}^I}{\partial p_\alpha^I}\mathbf{x}^I(i_1)\right]_b\left[\frac{\partial \mathbf{S}^I}{\partial p_\beta^I}\mathbf{x}^I(i_2)\right]_a + \sum_{i\in I}\sum_{a=1}^{3}\frac{\partial V}{\partial r_a^I(i)}\left[\frac{\partial^2 \mathbf{S}^I}{\partial p_\alpha^I \partial p_\beta^I}\mathbf{x}^I(i)\right]_a,$$

where we have used

$$\frac{\partial r_a^I(i)}{\partial p_\alpha^I} = \left[\frac{\partial \mathbf{S}^I}{\partial p_\alpha^I}\mathbf{x}^I(i)\right]_a, \quad i \in I. \tag{15}$$

The notation $[\ldots]_a$ corresponds to the a-th component of the vector given inside the bracket. Further details of the derivations are given in the Supporting Information.

To illustrate the computation of the normal modes, we first focus on the kinetic energy terms for the rigid bodies:

$$K_{RB} = \sum_I^{N_{RB}}\frac{1}{2}M^I\left(\dot{\mathbf{r}}^I\right)^2 + \sum_I^{N_{RB}}\sum_{\alpha=1}^{3}\sum_{\beta=1}^{3}\frac{1}{2}J_{\alpha\beta}^I\dot{p}_\alpha^I\dot{p}_\beta^I, \tag{16}$$

where the mass of rigid body I is $M^I = \sum_{i \in I} m^i$ and $J^I_{\alpha\beta}$ is the corresponding moment of inertia tensor. We choose to work in the moving frame of reference, where $\mathbf{S}^I = \mathbf{I}$, as we find diagonalisation of the inertia matrix the most straightforward procedure. Here the moment of inertia has the usual definition.

We now wish to transform to coordinates where the kinetic energy is diagonal, with $\mathbf{Q}^{I,T}$ and $\mathbf{Q}^{I,R}$ for the translational and rotational degrees of freedom of rigid body I, so that

$$K_{RB} = \sum_{I=1}^{N_{RB}} \left(\frac{1}{2} \left(\dot{\mathbf{Q}}^{I,T} \right)^2 + \frac{1}{2} \left(\dot{\mathbf{Q}}^{I,R} \right)^2 \right). \tag{17}$$

For the translational degrees of freedom, the required coordinate transformation is a simple rescaling: $\mathbf{Q}^{I,T} = \sqrt{M^I}\mathbf{r}^I$. However, for the rotational degrees of freedom, we must first apply a coordinate transformation $\mathbf{w}^I = \mathbf{A}^I \mathbf{p}^I$, so that the moment of inertia becomes a diagonal matrix with diagonal elements Ω^I_α (α = 1, 2, 3).[51,52] Then we can simply rescale the orientational coordinates by the moment of inertia $Q^{I,R}_\alpha = \sqrt{\Omega^I_\alpha} w^I_\alpha$.

More generally, the total kinetic energy of the system consists of contributions from the rigid bodies and free atoms, and we can write it as

$$K = \sum_{i=1}^{\eta} \frac{1}{2} \dot{Q}_i^{\,2}, \tag{18}$$

where $\eta = 3N_A + 6N_{RB}$ is the total number degrees of freedom. For the unconstrained atoms, $Q_i = X_i \sqrt{m^i}$, where m^i is the mass of the atom corresponding to atomic Cartesian coordinate X_i.

The next step in computing the normal modes is to expand the potential energy, V, in a Taylor series around a local minimum configuration with potential energy $V_{\min}$ up to second order in the Q coordinates:

$$V = V_{\min} + \frac{1}{2} \sum_{i,j=1}^{\eta} \frac{\partial^2 V}{\partial Q_i \partial Q_j} Q_i Q_j. \tag{19}$$

Here, $\mathbf{Q}$ is understood as the deviation from the local minimum configuration, which is defined as the local origin of coordinates. The Hessian matrix $H_{ij} = \partial^2 V / \partial Q_i \partial Q_j$ can be diagonalised using a matrix $\mathbf{B}$, whose columns are the eigenvectors of $\mathbf{H}$ with associated eigenvalues $\omega_i^2 = 4\pi^2 \nu_i^2$:

$$\sum_{j=1}^{\eta} H_{ij} B_{jk} = \omega_k^2 B_{ik}; \qquad q_i = \sum_{j=1}^{\eta} B_{ij} Q_j, \tag{20}$$

where q_i are the normal mode coordinates. In this coordinate system the Hamiltonian $\mathscr{H}$ can be written as

$$\mathscr{H} = V_{\min} + \frac{1}{2} \sum_{i=1}^{\eta} (\dot{q}_i^2 + \omega_i^2 q_i^2). \tag{21}$$

Due to the overall translational and rotational symmetries, there are six zero normal mode eigenvalues. The total number of vibrational degrees of freedom in local rigid body coordinates is therefore $\kappa = \eta - 6$, which is used in Eq. (10) to define the harmonic free energy of an individual minimum.

2.3 Basin-Hopping Parallel Tempering

The minima used in Eq. (9) to compute the binding free energy were sampled using basin-hopping global optimisation.[53–55] The basin-hopping method steps between local minima of the potential energy surface, proposing moves by perturbing the current minimum, and accepting or rejecting the new structure obtained after minimisation using criteria such as the energy difference.[53–55] We used the group rotation moves[56] described in §2.3.1 for perturbing the conformation of the current minimum in both the unconstrained inner and locally rigid intermediate regions. The perturbed conformation was minimised using a modified L-BFGS algorithm[57] with a tolerance of 0.001 kcal/mol/Å on the root mean square force. The new minimum was accepted or rejected using a Metropolis criterion based on the potential energy difference with respect to the previous minimum. Since the Metropolis criterion is based on the energy difference between local minima, all downhill barriers on the potential energy surface are removed. Uphill barriers are reduced to the difference in energy of the two minima. The minimisation and reduced barriers permit large perturbations of geometry, leading to effective sampling of the low energy regions of the potential energy surface of interest.

To enhance the sampling we employed the basin-hopping parallel tempering (BHPT) approach[58]. Conventional parallel tempering involves carrying out a parallel set of canonical Monte Carlo simulations at a range of temperatures, with periodic exchange attempts between the runs.[59,60] In the BHPT approach the replicas evolving at different temperatures are all basin-hopping runs[58] and the exchanges are between the current minima in adjacent replicas.

2.3.1 Group Rotations. To propose perturbed conformations within each basin-hopping replica, generalised rotation moves were developed. This scheme allows arbitrary groups of atoms to be rotated about an axis defined by a bond vector, maintaining maximum flexibility without introducing reliance on standard topologies. Each group i has an associated user specified selection probability, $P(\text{select})_i$, and maximum rotation angle, $\theta_i^{\max}$, to allow for further fine tuning of the conformational sampling. These perturbations are referred to as group rotation moves.[56] During each basin-hopping step:

1. for each group i, a random number ρ_1 is drawn between zero and one. If $P(\text{select})_i > \rho_1$ then the group is rotated in this step,

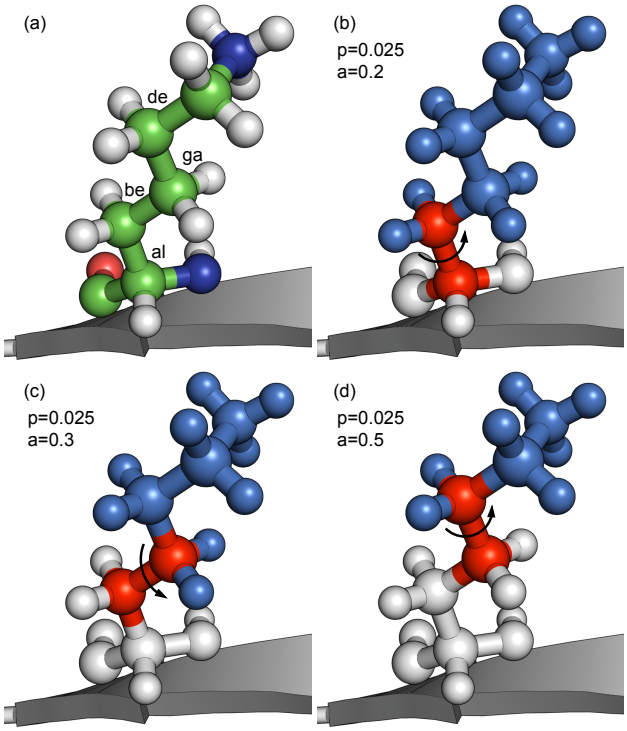

Fig. 2 The amino acid Lysine (LYS) (a) coloured by element with carbon atoms labelled. (b), (c) and (d) show the $\alpha\beta$, $\beta\gamma$ and $\gamma\delta$ groups that can be rotated during basin-hopping, with their associated selection probabilities $P(\text{select})_i$ and maximum rotation amplitudes $\theta_i^{\max}$. The axis of rotation is shown in red, while the atoms to be rotated are shown in blue. The graphical representations in Fig.2-5 were prepared by Pymol program.[61]

2. a second random number ρ_2 in the range $[-0.5, 0.5]$ is drawn and the rotation angle to be applied to the group is calculated as $\theta_i = 2\pi\rho_2\theta_i^{\max}$, where $\theta_i^{\max}$ is the maximum desired rotation angle for group i as a fraction of 2π.

3. The bond vector that connects the group to the rest of molecule is calculated and normalised before being scaled by θ_i.

For an atom with position vector $\mathbf{r}$, the rotation matrix $\mathbf{S}$ is generated using an implementation of Rodrigues' rotation formula,[62,63]

$$\mathbf{Sr} = \left[(\mathbf{I}\cos\theta) + \hat{\mathbf{k}}_\times \sin\theta + \hat{\mathbf{k}}\hat{\mathbf{k}}^T(1-\cos\theta)\right]\mathbf{r}, \qquad (22)$$

where $\mathbf{I}$ is the identity matrix, $\hat{\mathbf{k}}$ is the rotation axis, and $\hat{\mathbf{k}}_\times$ is the 'cross-product matrix':

$$\hat{\mathbf{k}} = \begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix}, \qquad \hat{\mathbf{k}}_\times = \begin{pmatrix} 0 & -k_3 & k_2 \\ k_3 & 0 & -k_1 \\ -k_2 & k_1 & 0 \end{pmatrix}. \qquad (23)$$

This scheme was initially developed to allow for comprehensive sampling of small ligands, but in the current work it has been adapted to sample the rotameric states of protein side chains. Fig. 2 shows the rotatable groups used to explore the conformations of the LYS side chain as an illustration. We define up to three such rotatable groups for each amino acid side chain, where atoms are rotated about the C_α-C_β, C_β-C_γ and C_γ-C_δ bonds. For simplicity, we set $P(\text{select})_i = 0.025$ for all groups, giving an average of 5.5 rotations per basin-hopping step for the 220 groups present when $R = 14$ Å (see §3.4). The maximum rotation amplitude $\theta_i^{\max}$ for each group was chosen based on the group's size and spatial extent, in an effort to achieve the largest possible step size while minimising possible atom clashes following a rotation. The values used in the current work can be found in the Supporting Information (Table S1) along with associated input files.

While the conformational changes during sampling are mainly determined by the group rotation of side chains and ligand, we also included small (0.1 Å) random Cartesian perturbations for all atoms, including the backbone, at every basin-hopping step. In addition, the backbone was free to move during minimization in the free and locally rigid regions to accommodate side chain/ligand movement. Thus, the backbone moves during the sampling. To estimate the contribution of the backbone movement, we looked at eight aldose reductase crystal structures with different ligands bound, which were obtained from the Protein Data Bank. Among these complexes, the smallest ligand has 18 atoms and the largest has 49. The highest C_α-RMSD between the one we used as a starting point and any other is 0.723 Å for the whole protein and 0.609 Å for the residues within 16 Å of the ligand (Table S2). These small differences in backbone conformation reflect the fact that the backbone conformation is quite well defined for the species considered in the FSA procedure.

3 Application to Human Aldose Reductase

We employ the binding of human aldose reductase[64] with phenyl acetic acid (PAC) as a model system to test the factorisation superposition approach (see Fig. 3). Because the protein is adequately large and the ligand is quite small, which complex can be a simple case described as Eq. (9). In addition, the crystal conformation of the complex and the experimental binding free energy are known. For the purposes of this study, not looking at the catalytic activity of the enzyme, the $NADP^+$ cofactor of the enzyme is considered to be part of the protein. The details of the simulation and local rigidification are described in §3.1 and §3.2, respectively. The goals of this study are to test the following two hypotheses. First, that the binding free energy should converge if the active binding site region is sufficiently large. Second, that the binding free energy should then be independent of the configuration of the

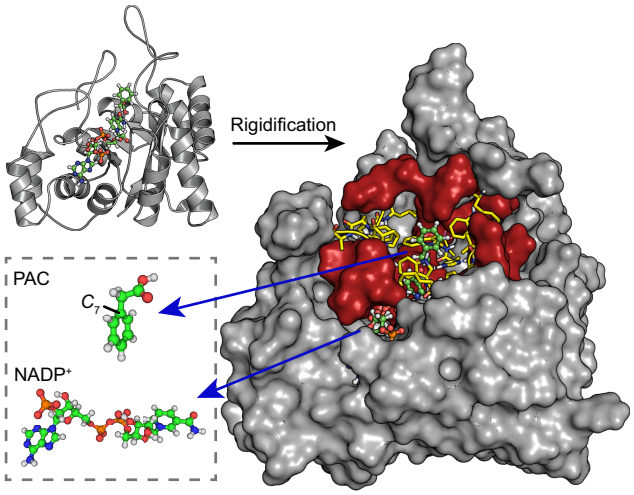

Fig. 3 Cartoon (left top) and rigidified representations (right) of St1-Comp-R12.In the rigidified structure, the yellow lines represent the unconstrained inner region, the red surface shows the locally rigidified intermediate region, and the gray part is the outer region, rigidified as one group. The ligand PAC with the atom labels used in the text and the cofactor $NADP^+$ are described in the insert.

inactive region. These hypotheses are tested by computing the binding free energy for systematic rigidification with three different reference conformations and examining the convergence to identify the maximum rigidification (factorisation) for which Eq. (9) holds.

3.1 Simulation Set Up

The simulations were performed using the AMBER ff99SB force field[8] for the protein. Parameters for $NADP^+$ were obtained from the AMBER parameter database[65]. The PAC ligand was parametrised using the General Amber Force Field[66,67] with RESP[68,69] charges generated iteratively using GAMESS-US.[70] The cutoff radius of 999.99 Å is used for non-bonded interactions. To evaluate the influence of the reference conformation, corresponding to $\mathbf{v}_A^0$ in Eq. (9), we prepared three initial conformations with different geometries for the rigid region. One conformation, named 'St-1', was obtained from the Protein Data Bank (PDB code 2INE)[64]. The other conformations, named 'St-2' and 'St-3', were prepared using a small number of basin-hopping steps starting from St-1 without any rigidification. Fig. 4(a) shows St-1 and St-2 aligned on all atoms (RMSD 1.5 Å), while Fig. 4(b) shows the alignment for St-1 and St-3 (RMSD 1.9 Å). The main differences are the partial unfolding of a helix in St-2 and St-3, respectively. Most of the calculations were performed *in vacuo* to reduce the computational cost and facilitate more thorough benchmarking. An accurate solvation model is not required for the present study since the objective is to test the factorisation approach, rather than compare directly with experiment. The calculations in aqueous solvent will be the focus of future work, discussed in §3.6. In the present contribution we have simply relaxed the key local minima using an implicit solvent model to check that the convergence criteria are robust. Example input files are provided in the Supporting Information.

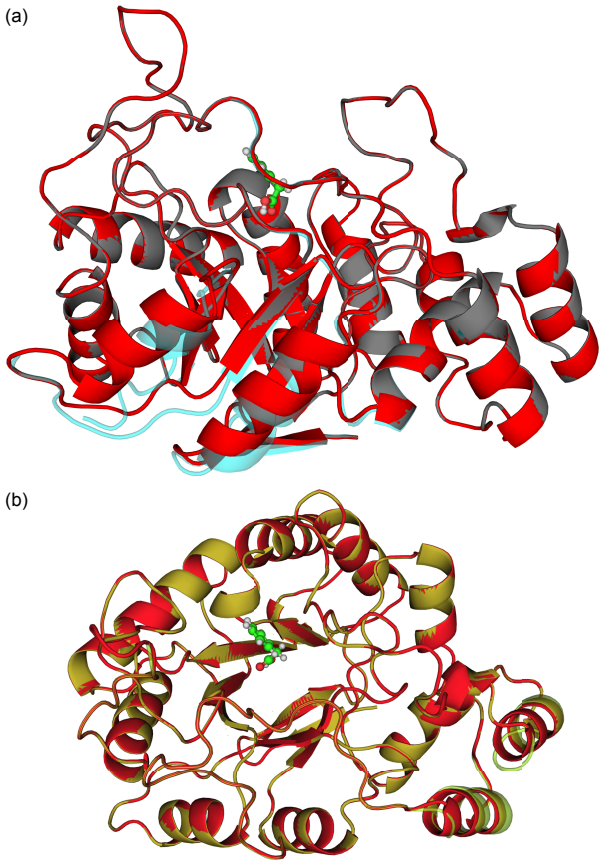

Fig. 4 (a) Cartoon descriptions of St-1 (red) and St-2 (sky blue) after alignment.The blue color of St-2 is translucent, thus the overlapped region looks gray. (b) St-1 (red) and St-3 (yellow).

3.2 Systematic Rigidification

For each structure, the free energy calculations were performed on multiple rigidified versions of the protein. The rigidification was applied systematically to fewer protein atoms, with the corresponding complex initially defined from identical protein and ligand coordinates. We determined the rigidified regions using the distance, R (in Angstroms), from the C_7 atom of the PAC ligand, labelled in Fig. 3. The unconstrained inner layer consisted of all atoms of amino acid residues having any atom within a radius R of the C_7 refer-

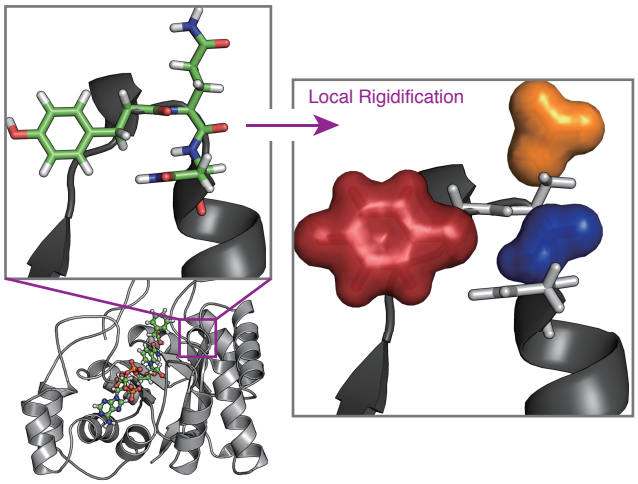

Fig. 5 Examples of the local rigidifications corresponding to the intermediate region of Fig. 3.The image at the top left shows residues 47-49 in human aldose reductase. In the local rigidification (right), the red group is an aromatic ring corresponding to the TYR47 residue, the blue group corresponds to a peptide bond between GLN48 and ASN49, and the orange group is a trigonal centre (an amide group in this case) in the side chain of GLN48.

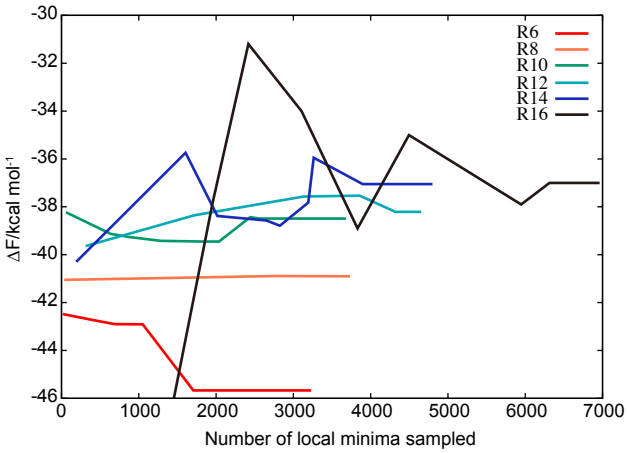

Fig. 6 Binding free energies as a function of the number of distinct local minima sampled, corresponding to the progress of the BHPT run. Results are shown for six different values of the radius R, which defines the unconstrained region.

ence atom. If any atoms of a residue lay between radii R and $R+1$ then we rigidified their peptide bonds, sp^2 centres, and aromatic rings. This set formed the intermediate layer with local rigid bodies. Atoms in the outer layer were rigidified as a single group. Fig. 3 shows the resulting rigidification scheme for the complex with a threshold value of $R = 12$ Å defined for St-1 (denoted St1-Comp-R12) and the details of the local rigidification are shown in Fig. 5. For St-1, six different rigidified versions were used, corresponding to $R = 6, 8, 10, 12, 14$ and 16, Å. For St-2, four versions ($R = 8, 10, 12, 14$, Å) were prepared, and for St-3, two versions ($R = 12, 14$, Å) were prepared. In each case the cofactor $NADP^+$ was part of the rigidified region. The number of atoms in each group is summarised in Table 1 as a function of R.

3.3 Sampling Local Minima

The BHPT method (§2.3)[58] implemented in our GMIN[71] program was used to sample local minima for the protein and complex for the different R values with both reference structures. All BHPT simulations were performed with 12 replicas and temperatures exponentially spaced between 97 K and 2435 K. Minimisation was performed using a modified version of the L-BFGS[57] algorithm with a tolerance of 0.001 kcal/mol/Å for the root mean square force. Minima with energies within 0.01 kcal/mol were considered duplicates and excluded from the set used for computing the free energy. For the BHPT run for the complex of St-1, the probability of escape from the previous minimum is 37% at the lowest temperature and 62% at the highest temperature, which corresponds to an efficient choice of parameters.

For each minimum, normal mode frequencies were computed using our OPTIM[72] program and harmonic free energies were obtained from Eq. (10) as the database of minima expanded (Fig. 6). Sampling was terminated when the change within the previous 2,400 basin-hopping steps was less than 0.01 kcal/mol. Table 2 gives the total number of basin-hopping steps and the number of distinct minima sampled for the different simulations. Among these minima, only a few hundred contribute significantly to the superposition sums in Eq. (9), and this number increases with R, as expected. As an example, we show how F_{AB} varies for St-1 in Table 3.

A single basin-hopping run of 1,000 steps was performed for the ligand using a temperature of 300 K in the accept/reject step. Only the lowest two local minima contribute significantly to the partition function of the noninteracting ligand.

3.4 Convergence of the Free energy with the Size of the Unconstrained Region

We calculated binding free energies, ΔF, using Eq. (10) for St-1, St-2 and St-3 as a function of R, as shown in Fig. 7. For St-1, ΔF increases from $R = 6$ Å to $R = 14$ Å and appears to have converged at $R = 14$ Å . The ΔF values obtained for St-2 deviate significantly from that of St-1 at $R = 8$ Å , but at $R = 10$ Å ΔF approaches the value obtained at $R = 14$ and 16 Å for St-1. Similar ΔF values are also obtained for St-3. ΔF at $R = 14$, 16 Å for St-1, $R = 12$, 14 Å for St-2 and $R = 12$, 14 Å for St-3 are within 1.1 kcal/mol, even though the number of degrees of freedom (κ_X) and absolute free energies (F_X) are quite differ-

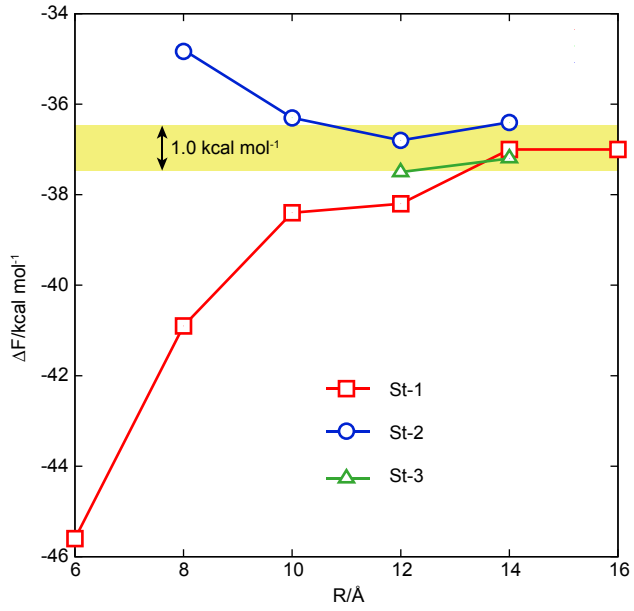

Fig. 7 Binding free energies as a function of the rigidification radius, R. Results for St-1, St-2 and St-3 are shown in red, blue and green, respectively. The shaded region represents 1.0 kcal mol^{-1} around the average converged value.

ent for each R, as detailed in Table 1 and Table 2. Thus, we conclude that the factorisation superposition approach seems to be applicable for this system with $R \geq 14$ Å, independent of the reference conformation.

3.5 Computational Cost

In the BHPT sampling using `GMIN`, each basin-hopping step for St1-Comp-R14 takes about 3.1 times longer than for St1-Comp-R6 on average, because the coordinate space is larger for St1-Comp-R14. For the normal mode analysis using `OPTIM`, the diagonalisation of the Hessian matrix for one minimum with $\kappa = 3246$ ($R = 14$) and $\kappa = 15387$ (without any rigidification) took 6 minutes and 46 minutes of cpu time on average, respectively, the computational time scales roughly as $\kappa^{1.5}$, as suggested by the data in Fig. S1 of the Supporting Information.

In spite of the speedup achieved using the rigid body framework, normal mode calculations for the protein and complex minima are still relatively expensive. It is therefore desirable to use as few minima as possible in the superposition sums. Due to the Boltzmann weight in Eq. (13), the low-energy minima dominate these sums. Table 4 shows the binding free energy computed using minima whose energies lie within a cutoff (E_{cut}) of the global minimum energy. We find that the binding free energy is determined by minima with energies within $10kT$ of the global minimum at $T = 298$ K. This cutoff corresponds to a small fraction of the total number of minima sampled for the protein and complex. For example, a cut-off of $10kT$ applied to the database of minima for the $R = 14$ Å simulations with St-1 drastically reduces the number of minima of the complex from 4,452 to 149. The number of relevant minima for smaller R is even less. A substantial reduction in the total computational cost can therefore be achieved by restricting the normal mode calculations to the low-energy minima.

3.6 Extension to FSA in aqueous solvent

The converged binding free energy was found to be approximately -36.8 kcal/mol, corresponding to a standard binding free energy of -29.8 kcal/mol, which is significantly lower than experimental binding affinity of -5.5kcal/mol[64]. We suspected that this discrepancy is primarily due to the absence of solvent effects. To test this hypothesis, we repeated the calculation for $R = 14$ Å with St-1, using the Generalized Born implicit solvent model, as implemented in AMBER.[73] We relaxed the lowest 500 minima identified in vacuum and recomputed the normal mode frequencies for both the protein and the complex. Note that we did not resample minima while accounting for solvent contributions, as the vacuum and the corresponding recomputed potential energies in the implicit solvent were found to be highly correlated (Fig. S2). Both the ligand minima were also relaxed using implicit solvent. The resulting binding free energy was $\Delta F^{\circ} = -8.4$ kcal/mol, much closer to the experimental value. We expect that sampling with a more accurate implicit solvent model, such as linearized Poisson-Boltzmann[74], would further improve the agreement with experiment.

We note that, in principle, the FSA framework can also be applied for explicit solvent. However, a large number of explicit water molecules would significantly increase the number of minima, and further work would be needed to sample these structures efficiently. Nevertheless, including a few water molecules might be desirable, for example, in situations where the crystal structure contains bridging water molecules between the ligand and the protein.

4 Conclusions

We have presented a new method based on potential energy landscape theory,[26] the factorisation superposition approach (FSA), for computing the binding free energy of protein-ligand complex. In this scheme the free energy of the free and bound molecules are computed using the superposition approach from a database of local potential energy minima. Due to the exponential increase in the number of minima with system size, exhaustive sampling is not feasible for a protein-sized system. The FSA approach addresses this problem by focusing the calculation on protein atoms that interact strongly

with the ligand. In §2.1 we presented the theory for factorising the conformational space of the protein and complex into two regions based on the size of the binding pocket. The factorisation facilitates estimation of the binding free energy using minima corresponding to fluctuations of the binding region, thereby reducing the number of degrees of freedom significantly. We describe the approximations under which such a factorisation is valid, employing a local rigid-body framework[44] to implement the FSA by treating atoms further from the binding site as collections of local rigid bodies. This procedure reduces the number of active degrees of freedom, but retains all the terms in the force field.

We applied the FSA method to calculate the free energy change for ligand binding with human aldose reductase protein while varying the size of the binding region. We performed the calculations for three different conformations of the rigid part of the protein and for different sizes of the binding pocket. For a given conformation of the rigidified region, we found that the binding free energy converged to within 1 kcal/mol as the size of the binding pocket was increased to about 14 Å, corresponding to an 80% reduction in the number of protein degrees of freedom. The converged binding free energy for all three conformations were found to be within 1.1 kcal/mol, suggesting weak interactions between the ligand and protein atoms beyond 14 Å.

Several further improvements in the accuracy and speed of the FSA method as presented here can be envisioned. Larger systems are likely to derive a greater benefit from the factorisation scheme, because the whole region unrelated to ligand binding can be rigidified into a single unit, with only six rigid-body degrees of freedom. A surprising result of this study is that, even though the number of minima increased rapidly with the size of the unconstrained region around the binding pocket, the number of thermally relevant minima remained small, of the order of few hundred conformations. Anharmonicity corrections[33,75,76] could improve the accuracy of the method, and the computationally intensive minima sampling and normal mode calculations should be highly amenable to distributed computing.

One key aspect of the FSA approach is the rigidification of large protein regions distant from the binding site. This approach assumes that the configurations of such regions change relatively little upon ligand binding. For proteins with significant allosteric effects,[77,78] the regions should be rigidified in smaller domains, to avoid freezing out the protein allostery. The local rigid body approach, used in the 'intermediate region', and group rotations for sampling should still be applicable for any ligand binding system.

The converged radius for the flexible region obtained in the present work, $R = 14$ Å, is not expected to be universal, and other protein/ligand combinations will require analogous convergence checks. However, for a given protein, the value of R is likely to be transferable for different ligands of comparable size.

In future work we will consider solvent effects in more detail, and present comparisons with alternative approaches for calculating the free energy difference. Our main purpose in the present work was to demonstrate the convergence of the FSA scheme. We hope that, with further benchmarking and computational optimisation, the FSA method could facilitate screening calculations associated with drug design.

5 Acknowledgement

K.M. is grateful to Prof. Iwao Ohmine and Prof. Nobuhiro Kosugi for supporting his stay in Cambridge, and thanks Dr Motoshi Kamiya for useful discussions in setting up AMBER calculations. This research was funded by EPSRC Programme grant EP/I001352/1 and ERC grant RG59508.

References

1 W. L. Jorgensen, Science, 2004, **303**, 1813–1818.
2 J. Michel, N. Foloppe and J. W. Essex, Mol. Inf., 2010, **29**, 570–578.
3 H. Guitierrez-de-Teran and J. Aqvist, Methods Mol. Biol., 2012, **819**, 305–323.
4 R. R. Johnson, A. Kohlmeyer, A. T. C. Johnson and M. L. Klein, Nano Lett., 2009, **9**, 537–541.
5 G. Ercolani, J. Am. Chem. Soc., 2003, **125**, 16097–16103.
6 T. Cheng, Q. Li, Z. Zhou, Y. Wang and S. Bryant, AAPS J., 2012, **14**, 133–141.
7 D. B. Kitchen, H. Decornez, J. R. Furr and J. Bajorath, Nature Rev. Drug Discov., 2004, **3**, 935–949.
8 D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. C. III, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, Comp. Phys. Commun., 1995, **91**, year.
9 T. Lybrand, J. A. McCammon and G. Wipff, Proc. Natl. Acad. Sci. U.S.A, 1986, **83**, 833–835.
10 P. A. Kollman, Chem. Rev., 1993, **93**, 2395–2417.
11 D. L. Beveridge and F. M. Dicapua, Annu. Rev. Biophys. Biophys. Chem., 1989, **18**, 431–492.
12 M. R. Reddy and M. D. Erion, Curr. Pharm. Des., 2005, **11**, 283–294.
13 Y. Deng and B. Roux, J. Phys. Chem. B, 2009, **113**, 2234–2246.
14 P.-C. Chen and S. Kuyucak, Biophys. J., 2011, **100**, 2466 – 2474.
15 S. Park and K. Schulten, J. Chem. Phys., 2004, **120**, 5946–5961.
16 F. M. Ytreberg, J. Chem. Phys., 2009, **130**, 164906.
17 S. P. Brown and S. W. Muchmore, J. Chem. Inf. Model., 2006, **46**, 999–1005.
18 H. Gouda, I. Kuntz, D. Case and P. Kollman, Biopolymers, 2003, **68**, 16–34.
19 D. A. Pearlman, J. Med. Chem., 2005, **48**, 7796–7807.
20 J. Srinivasan, T. Cheatham, P. Cieplak, P. Kollman and D. Case, J. Am. Chem. Soc., 1998, **120**, 9401–9409.
21 J. Aqvist, C. Medina and Samuelsson, J. Protein Eng., 1994, **7**, 385–391.
22 J. Carlsson, M. Ander, M. Nervall and J. Aqvist, J. Phys. Chem. B, 2006, **110**, 12034–12041.
23 D. Jones-Hertzog and W. Jorgensen, J. Med. Chem., 1997, **40**, 1539–1549.
24 R. Zhou, R. Friesner, A. Ghosh, R. Rizzo, W. Jorgensen and R. Levy, J. Phys. Chem. B, 2001, **105**, 10388–10397.

25 C. Oostenbrink and W. F. van Gunsteren, Proteins, 2004, **54**, 237–246.
26 D. J. Wales, Energy Landscapes, Cambridge University Press, Cambridge, 2003, pp. 364–433.
27 D. J. Wales, Phil. Trans. Roy. Soc. A, 2005, **363**, 357–377.
28 B. Strodel and D. J. Wales, Chem. Phys. Lett., 2008, **466**, 105–115.
29 P. G. Mezey, Potential Energy Hypersurfaces, Elsevier, Amsterdam, 1987, pp. 198–368.
30 D. J. Wales, Mol. Phys., 1993, **78**, 151–171.
31 D. J. Wales and J. P. K. Doye, J. Chem. Phys., 1995, **103**, 3061–3070.
32 J. P. K. Doye and D. J. Wales, J. Chem. Phys., 1995, **102**, 9673–9688.
33 F. Calvo, J. P. K. Doye and D. J. Wales, J. Chem. Phys., 2001, **115**, 9627–9636.
34 F. Calvo, J. P. K. Doye and D. J. Wales, J. Chem. Phys., 2001, **114**, 7312–7329.
35 W. Chen, C. E. Chang and M. K. Gilson, Biophys. J, 2004, **87**, 3035–3049.
36 J. P. K. Doye, M. A. Miller and D. J. Wales, J. Chem. Phys., 1999, **110**, 6896–6906.
37 D. J. Wales and T. V. Bogdan, J. Phys. Chem. B, 2006, **110**, 20765–20776.
38 V. A. Sharapov, D. Meluzzi and V. A. Mandelshtam, Phys. Rev. Lett., 2007, **98**, 105701.
39 V. A. Sharapov and V. A. Mandelshtam, J. Phys. Chem. A, 2007, **111**, 10284–10291.
40 F. H. Stillinger and T. A. Weber, Science, 1984, **225**, 983–989.
41 D. J. Wales and J. P. K. Doye, J. Chem. Phys., 2003, **119**, 12409–12416.
42 T. V. Bogdan, D. J. Wales and F. Calvo, J. Chem. Phys., 2006, **124**, 044102.
43 W. Chen, M. K. Gilson, S. P. Webb and M. J. Potter, J. Chem. Theory Comput., 2010, **6**, 3540–3557.
44 H. Kusumaatmaja, C. S. Whittleston and D. J. Wales, J. Chem. Theory Comput., 2012, **8**, 5159–5165.
45 J. M. Petrash, Cell. Mol. Life, 2004, **61**, 737–749.
46 M. K. Gilson, J. A. Given, B. L. Bush and J. A. McCammon, Biophys. J., 1997, **72**, 1047–1069.
47 H. Zhou and M. K. Gilson, Chem. Rev., 2009, **109**, 4092–4107.
48 G. Emilio and M. L. Ronald, Recent theoretical and computational advances for modeling protein-ligand binding affinities, Academic Press, 2011, vol. 85, pp. 27 – 80.
49 F. G. Amar and R. S. Berry, J. Chem. Phys., 1986, **85**, 5943–5954.
50 D. Chakrabarti and D. J. Wales, Phys. Chem. Chem. Phys., 2009, **11**, 1970–1976.
51 A. Pohorille, L. R. Pratt, R. A. LaViolette, M. A. Wilson and R. D. MacElroy, J. Chem. Phys., 1987, **87**, 6070–6077.
52 D. J. Wales and I. Ohmine, J. Chem. Phys., 1993, **98**, 7257–7268.
53 Z. Q. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A., 1987, **84**, 6611–6615.
54 D. J. Wales and J. P. K. Doye, J. Phys. Chem. A, 1997, **101**, 5111–5116.
55 D. J. Wales and H. A. Scheraga, Science, 1999, **285**, 1368–1372.
56 C. Whittleston, PhD Thesis, University of Cambridge, 2011.
57 W. Press, B. Flannery, S. Teukolsky and W. Vetterling, Numerical Recipes, Cabridge University Press, Cambridge, 1986, pp. 487–555.
58 B. Strodel, J. W. L. Lee, C. S. Whittleston and D. J. Wales, J. Am. Chem. Soc., 2010, **132**, 13300–13312.
59 G. J. Geyer, Stat. Sci., 1992, **7**, 437.
60 K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn., 1996, **65**, 1604–1608.
61 The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrodinger, LLC.
62 H. Goldstein, Classical mechanics, Addison-Wesley, Reading, Massachusetts, 1980, pp. 128–187.
63 D. Chakrabarti and D. J. Wales, Phys. Chem. Chem. Phys., 2009, **11**, 1970–1976.
64 J. M. Brownlee, E. Carlson, A. C. Milne, E. Pape and D. H. T. Harrison, Bioorg. Chem., 2006, **34**, 424–444.
65 N. Holmberg, U. Ryde and L. Bulow, Prot. Engin., 1999, **12**, 851–856.
66 J. Wang, P. Cieplak and P. A. Kollman, J. Comput. Chem., 2000, **21**, 10491074.
67 J. Wang, W. Wang, P. A. Kollman and D. A. Case, J. Mol. Graph. Model., 2006, **25**, 247–260.
68 C. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, J. Phys. Chem., 1993, **97**, 10269–10280.
69 W. Cornell, P. Cieplak, C. I. Bayly and P. A. Kollman, J. Am. Chem. Soc., 1993, **115**, 9620–9631.
70 M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Kosecki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis and J. A. Montgomery, J. Comput. Chem., 1993, **14**, 1347–1363.
71 D. J. Wales, GMIN: A program for basin-hopping global optimisation, basin-sampling, and parallel tempering.
72 D. J. Wales, OPTIM: A program for optimizing geometries and calculating reaction pathways.
73 A. Onufriev, D. Bashford and D. A. Case, Proteins, 2004, **55**, 383–394.
74 G. Sigalov, A. Fenley and A. Onufriev, J. Chem. Phys., 2006, **124**, 124902.
75 M. J. P. C. E. Chang and M. K. Gilson, J. Phys. Chem. B, 2003, **107**, 1048–1055.
76 K. A. A. B. Temelso and G. C. Shields, J. Phys. Chem. A, 2011, **115**, 12034–12046.
77 R. J. Hawkins and T. C. B. McLeish, J. R. Soc. Interface, 2006, **3**, 125–138.
78 R. J. Hawkins and T. C. B. McLeish, Phys. Rev. Lett., 2004, **93**, 098104.

Table 1 The binding free energy calculations are performed with the protein (molecule A) atoms separated into three different regions. The inner region is fully flexible, the intermediate region consists of local rigid bodies (LRB), and the outer region is treated as a single rigid body. The total number of atoms in the ligand, protein (including the cofactor $NADP^+$) and complex (molecule AB) are 18, 5113 and 5131, respectively. The table gives the number of degrees of freedom for protein (κ_A) and complex (κ_{AB}). The number of degrees of freedom for the ligand is $\kappa_B = 48$

St	Radius R (Å)	% rigid (protein)	Number of rigidified atoms		κ_A	κ_{AB}
			intermediate (# LRB)	outer		
1,2	6	99	0 (0)	5091	66	120
	8	97	36 (7)	4903	564	618
	10	92	92 (17)	4640	1245	1299
	12	87	114 (25)	4338	2133	2187
	14	80	135 (29)	3972	3192	3246
1	16	78	100 (21)	3886	3507	3561
3	12	88	137 (27)	4378	1956	2010
	14	81	146 (32)	3992	3117	3171

Table 2 Total basin-hopping (BH) steps for 12 temperatures and the number of distinct local minima obtained for the complex (AB) and the protein (A). The binding free energies (ΔF) are calculated from the free energies of the complex (F_{AB}), protein (F_A) and ligand (F_B). A converged value of $F_B = 11.4$ kcal/mol is obtained from the two lowest minima characterised in a BH run of 1000 steps

St	R	Total BH steps		# minima obtained		Free energies (kcal/mol)		
		Complex	Protein	Complex	Protein	Complex	Protein	ΔF
1	6	50723	46602	3229	2298	−9920.3	−9886.1	−45.6
	8	25002	24106	3733	3595	−9464.1	−9434.6	−40.9
	10	24072	20859	3680	2861	−8921.4	−8894.4	−38.4
	12	28467	29130	4873	4606	−8221.7	−8194.9	−38.2
	14	26351	23930	4800	4056	−7348.8	−7323.2	−37.0
	16	47880	41172	6962	6200	−4052.7	−4027.1	−37.0
2	8	18303	19473	2117	2259	799.3	822.7	−34.8
	10	15610	31317	1593	2091	1337.2	1362.1	−36.3
	12	17183	16474	2400	2351	2021.2	2046.6	−36.8
	14	19235	15030	3932	2899	2876.9	2901.9	−36.4
3	12	38676	31608	4388	3394	−5104.1	−5077.9	−37.5
	14	33816	26340	3382	3405	−4398.7	−4372.9	−37.2

Table 3 Free energies of the complex (kcal/mol) for reference St-1 using the $N_{\min}$ lowest minima. The free energies changing by more than 0.001 kcal/mol from the previous value are summarised below. The final values correspond to F_{AB} in Table 2

$N_{\min}$	$R = 6$ Å	$R = 8$ Å	$R = 10$ Å	$R = 12$ Å	$R = 14$ Å	$R = 16$ Å
1	−9916.821	−9463.359	−8920.421	−8219.631	−7346.759	−4049.182
10	−9917.676	−9464.074	−8921.024	−8220.497	−7347.315	−4050.390
30	−9920.302	−9464.075	−8921.306	−8220.884	−7348.228	−4050.837
50			−8921.314	−8221.049	−7348.355	−4051.481
70			−8921.416	−8221.126	−7348.515	−4052.084
90				−8221.161	−7348.649	−4052.097
110				−8221.589	−7348.732	−4052.173
130				−8221.667	−7348.777	−4052.181
150				−8221.669	−7348.796	−4052.634
170					−7348.801	−4052.636
190					−7348.804	−4052.683

Table 4 Binding free energy (kcal/mol) for reference St-1 using protein and complex minima with energies within E_{cut} of the global minimum. The binding free energy computed using all the minima is given in Table 2. E_{cut} is in units of kT for $T = 298\,\text{K}$

E_{cut}	$R = 6$ Å	$R = 8$ Å	$R = 10$ Å	$R = 12$ Å	$R = 14$ Å	$R = 16$ Å
2	−42.8	−40.9	−40.6	−39.0	−36.7	−35.2
5	−45.6	−40.8	−38.4	−39.5	−37.7	−36.5
10	−45.6	−40.8	−38.4	−38.2	−37.0	−37.0
20	−45.6	−40.8	−38.4	−38.2	−37.0	−37.0