

# Alchemical Transfer Approach to Absolute Binding Free Energy Estimation

Joe Z. Wu,<sup>†,‡,||</sup> Solmaz Azimi,<sup>†,¶,||</sup> Sheenam Khuttan,<sup>†,¶,||</sup> Nanjie Deng,<sup>§</sup> and  
Emilio Gallicchio\*,<sup>†,‡,¶,||</sup>

<sup>†</sup>*Department of Chemistry, Brooklyn College of the City University of New York, New York, NY*

<sup>‡</sup>*Ph.D. Program in Chemistry, The Graduate Center of the City University of New York, New York, NY*

<sup>¶</sup>*Ph.D. Program in Biochemistry, The Graduate Center of the City University of New York, New York, NY*

<sup>§</sup>*Department of Chemistry and Physical Sciences, Pace University, New York, New York, 10038*

<sup>||</sup>*These authors contributed equally to this work*

E-mail: [egallicchio@brooklyn.cuny.edu](mailto:egallicchio@brooklyn.cuny.edu)

## Abstract

The Alchemical Transfer Method (ATM) for the calculation of standard binding free energies of non-covalent molecular complexes is presented. The method is based on a coordinate displacement perturbation of the ligand between the receptor binding site and the explicit solvent bulk, and a thermodynamic cycle connected by a symmetric intermediate in which the ligand interacts with the receptor and solvent environments with equal strength. While the approach is alchemical, the implementation

of ATM is as straightforward as for physical pathway methods of binding. The method is applicable in principle with any force field, as it does not require splitting the alchemical transformations into electrostatic and non-electrostatic steps, and it does not require soft-core pair potentials. We have implemented ATM as a freely available and open-source plugin of the OpenMM molecular dynamics library. The method and its implementation are validated on the SAMPL6 SAMPLing host-guest benchmark set. The work paves the way to streamlined alchemical relative and absolute binding free energy implementations on many molecular simulation packages and with arbitrary energy functions including polarizable, quantum-mechanical, and artificial neural network potentials.

## Introduction

The binding free energy of a molecular complex is a rigorous thermodynamic measure of the degree of affinity of two molecules for each other. Measurements of binding free energies are useful in a wide range of chemical and medicinal applications ranging from drug discovery to chemical detection and toxicology. The ability to estimate binding free energies by computational modeling adds an important dimension to this probe.<sup>1</sup> Relative binding free energy models are widely used, for example, in drug lead optimization.<sup>2</sup>

This work is concerned in particular with atomistic models of the absolute binding free energy. These methods can be divided into two classes, each of which is based on the nature of the thermodynamic path that is used to connect the bound and unbound states of the molecular complex.<sup>3</sup> Physical pathway methods define a spatial coordinate along which the reversible work for bringing the two molecules together is calculated. Conversely, alchemical methods connect the bound and unbound states by a series of unphysical intermediate states. The Single-Decoupling Method (SDM),<sup>4,5</sup> for example, is based on progressively turning on the effective interaction between the ligand and the receptor with an implicit representation of the solvent. The related Double-Decoupling Method (DDM),<sup>6</sup> which is widely used to

estimate the absolute binding free energy with an explicit representation of the solvent, estimates the binding free energy as the difference between the free energies of coupling the ligand to a pure solvent system and to the solvated receptor.

Physical and alchemical binding free energy methods are characterized by distinct challenges and limitations.<sup>7</sup> While it is broadly employed, the alchemical DDM approach is known to suffer from poor convergence and strong bias, especially for large and charged ligands, whose large and compensating decoupling free energies result in the amplification of statistical and systematic errors in the binding free energy estimate.<sup>8</sup> Another significant shortcoming of DDM is the large perturbation of the environment of the ligand in going from the solvated state to the vacuum intermediate state. Unless the ligand is conformationally restrained, the transition from a solvated state to vacuum can trigger substantial intramolecular conformational changes that relax slowly to the bound or unbound configurations. Other technical limitations of DDM stem from differences in the composition and size of the molecular systems used for the two decoupling legs,<sup>9</sup> and from the inconsistent treatment of long-range interactions.<sup>10,11</sup> Even for relatively simple systems, these and other alchemical transformations must be conducted with care to avoid singularities and slow convergence. It is recommended, for example, to couple electrostatic interactions separately from other interactions,<sup>12</sup> and to employ customized soft-core pair potentials<sup>13</sup> to avoid end-point singularities.

In addition to these challenges, software implementations of alchemical binding free energy methods also tend to be significantly complex. These methods are often based on modifying the parameters and the form of the energy function, and as a result, require in-depth knowledge of the data structures of the target molecular simulation package. For example, core energy routines are usually customized to implement the specific modified pair potentials that represent the interaction of the ligand with the rest of the system. These modified pair potentials depend on the alchemical progress parameter (generally denoted by  $\lambda$ ) which, together with other alchemical variables, become additional system param-

ters. Methods such as Thermodynamic Integration (TI)<sup>14,15</sup> require additional routines to implement the calculations of the gradients of the energy function with respect to the parameters that are alchemically transformed. A particular challenge is the implementation of alchemical transformations that involve many-body potentials, where the  $\lambda$  dependence affects more than individual pair interaction energies. These include alchemical applications with polarizable potentials,<sup>16</sup> Ewald long-range electrostatic treatments,<sup>17</sup> implicit solvent models,<sup>18</sup> as well as some conventional intramolecular potential terms.<sup>19</sup>

Physical pathway methods<sup>8,20,21</sup> address some of the limitations of DDM by physically moving the ligand from the solvent bulk to the binding site. Physical pathway binding free energy calculations are seen as preferable over DDM for large and charged ligands because they are typically performed as one continuous path in a single solvent box without transferring the ligand to a different phase. In addition, software implementations of physical pathway methods do not require as much customization of the underlying molecular simulation package as alchemical methods do. The primary limitation of physical pathway methods is the high computational cost due to the requirement of equilibrating the complex at many intermediate receptor-ligand separations that might not be of interest. Because these methods require a physical exit and entry channel, physical pathway methods are also not generally applicable to occluded binding sites.<sup>22</sup>

Building upon our Single Decoupling Method (SDM) for absolute binding free energy estimation<sup>23,24</sup> implemented in OpenMM,<sup>25</sup> we have been investigating techniques to streamline alchemical calculations with explicit solvation. SDM, which has been designed for binding free energy calculations with implicit solvation, computes the alchemical perturbation energy by translating the ligand from the solvent medium to the receptor binding site, rather than attempting to selectively turn-on and turn-off individual ligand-receptor interactions. The approach treats all interactions in one concerted step and employs the standard molecular mechanics force field without soft-core pair potentials. End-point singularities are addressed by a suitable non-linear alchemical energy function.<sup>24</sup> We have recently

shown that the same approach is applicable to the estimation of the concerted hydration free energies of drug-sized solutes in water droplets with explicit solvation.<sup>26</sup>

In this work, we extend this concerted alchemical scheme to the calculations of absolute binding free energies with explicit solvation. In the resulting alchemical scheme, called the Alchemical Transfer Method (ATM) (Fig. 1B), the unbound and bound states of the molecular complex are related by a translation vector that brings the ligand from the solvent bulk to the receptor binding site in a single solvent box. We show that the proposed method addresses some of the aforementioned challenges of binding free energy calculations by exploiting the best characteristics of the alchemical and physical methods. Like DDM, the method is based on alchemical transformations that accomplish efficient conformational sampling only in the solvent bulk and the receptor binding site, and like physical methods, ATM is based on moving the ligand in physical space, in a single simulation box, and without transferring the ligand to vacuum. The ATM method, implemented as a freely available plugin of the OpenMM molecular simulation package, does not require any modifications of the OpenMM core energy routines. Here, we validate the ATM approach on a rigorous benchmarking dataset developed by Rizzi et al.<sup>27</sup>

## Theory and Methods

### Alchemical Transformations

The Alchemical Transfer Method introduced in the next section is based on the thermodynamic cycle illustrated in Fig. 1B. The free energy changes of each leg of the cycle are estimated using alchemical transformations based on a  $\lambda$ -dependent potential energy functions  $U_\lambda(x)$  that brings the system from an initial state at  $\lambda = 0$ , described by the potential function  $U_0(x)$ , to a final state at  $\lambda = 1$ , corresponding to the potential function  $U_1(x)$ .

In this work we adopt the family of alchemical potential energy functions introduced in

references 24 and 26. Briefly, the alchemical potential energy function is expressed as

$$U_\lambda(x) = U_0(x) + W_\lambda(u) \quad (1)$$

where  $x$  represents the set of atomic coordinates of the system,

$$u(x) = U_1(x) - U_0(x) \quad (2)$$

is the perturbation energy, and  $W_\lambda(u)$  is the generalized softplus alchemical perturbation function

$$W_\lambda(u) = \frac{\lambda_2 - \lambda_1}{\alpha} \ln [1 + e^{-\alpha(u_{\text{sc}}(u) - u_0)}] + \lambda_2 u_{\text{sc}}(u) + w_0. \quad (3)$$

The parameters  $\lambda_2$ ,  $\lambda_1$ ,  $\alpha$ ,  $u_0$ , and  $w_0$  are functions of  $\lambda$  (see Computational Details), and

$$u_{\text{sc}}(u) = \begin{cases} u & u \leq u_0 \\ (u_{\text{max}} - u_0) f_{\text{sc}} \left[ \frac{u - u_0}{u_{\text{max}} - u_0} \right] + u_0 & u > u_0 \end{cases} \quad (4)$$

with

$$f_{\text{sc}}(y) = \frac{z(y)^a - 1}{z(y)^a + 1}, \quad (5)$$

and

$$z(y) = 1 + 2y/a + 2(y/a)^2 \quad (6)$$

is the soft-core perturbation energy. The soft-core function is a monotonic map that avoids singularities near the initial state of the alchemical transformation at  $\lambda = 0$  without affecting the distribution of perturbation energies at the final state at  $\lambda = 1$ . As expressed in Eq. (4), the soft-core perturbation energy is designed to cap the perturbation energy  $u(x)$  to a maximum positive value  $u_{\text{max}}$  and to be equal to the perturbation energy when this is below a cutoff value  $u_0$ . The  $u_0$  cutoff is selected to be sufficiently large so that  $u_{\text{sc}}(u) = u$  for all observed samples collected at the end state. The specific values of  $u_0$ ,  $u_{\text{max}}$ , and of the

scaling parameter  $a$  used in this work are listed in the Computational Details.

In order to reproduce the desired end points, it is necessary that the alchemical perturbation function is defined such that  $W_0(u) = 0$  and  $W_1(u) = u$  at  $\lambda = 0$  and  $\lambda = 1$ , respectively. This requirement is satisfied by the linear function,  $W_\lambda(u) = \lambda u_{\text{sc}}(u)$ , which is special case of the softplus function in Eq. (3) for which  $\lambda_1 = \lambda_2 = \lambda$ . The linear function is the standard choice for the alchemical perturbation function. As it can be verified from Eq. (3), in general the end-point requirement is satisfied whenever  $\lambda_1 = \lambda_2 = 0$  at  $\lambda = 0$  and  $\lambda_1 = \lambda_2 = 1$  at  $\lambda = 1$ .

## The Alchemical Transfer Method for Binding Free Energy Estimation

Consider the non-covalent binding process between a receptor R and a ligand L. The standard free energy of binding,  $\Delta G_b^\circ$ , is defined as the difference in free energy between the bound complex and the unbound components at the standard concentration of  $C^\circ = 1$  M,

$$\Delta G_b^\circ = \Delta G_{\text{site}}^\circ + \Delta G_b^*. \quad (7)$$

where  $\Delta G_b^*$  is the excess component, defined as the reversible work for transferring the ligand to the binding site region of the receptor of volume  $V_{\text{site}}$  from a region of the same volume in the solvent bulk (Fig. 1) plus a concentration-dependent term

$$\Delta G_{\text{site}}^\circ = -k_B T \ln C^\circ V_{\text{site}} \quad (8)$$

that corresponds to the free energy of transfer of a ligand molecule from an ideal solution at concentration  $C^\circ$  to a region of volume  $V_{\text{site}}$  in the solvent.<sup>3</sup> In the remainder, we will focus on the calculation of the excess free energy component  $\Delta G_b^*$  by alchemical molecular simulations.

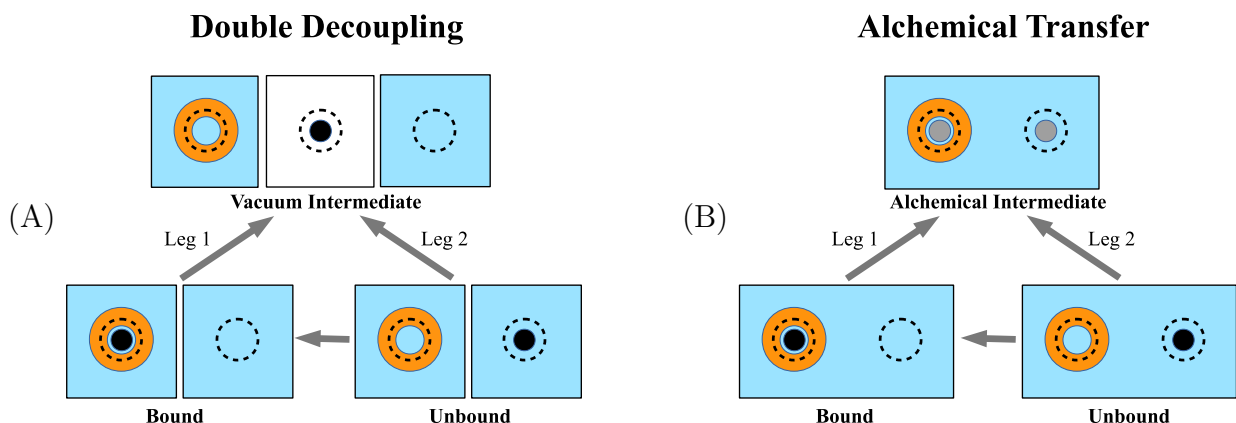


Figure 1: Schematic illustrations of (A) the double decoupling (DDM) and (B) the Alchemical Transfer (ATM) methods for the calculation of the binding free energy between a molecular receptor (orange doughnut) and a ligand (black circle). The dashed circle within the receptor represents the binding site region. The blue boxes represent the solvent. The unbound and bound end states for the two methods are considered thermodynamically equivalent. In both methods, the end states are transformed to a common intermediate state and the excess binding free energy is the difference of the free energy changes of the two legs,  $\Delta G_b^* = \Delta G_2 - \Delta G_1$ . Double decoupling defines an intermediate state in which the ligand is in vacuum (white). The alchemical transformations in the Alchemical Transfer Method are instead performed in the same solvent box and the ligand does not leave the solvated system.

The Double Decoupling Method (DDM)<sup>6</sup> has emerged as one of the gold standards for the alchemical calculation of absolute binding free energies in condensed phase. DDM is based on the thermodynamic cycle illustrated in Fig. 1A, whereby the bound and unbound states of the molecular complex are thermodynamically linked by an intermediate state in which the ligand is placed in vacuum. The excess binding free energy is expressed as the difference of the free energies of alchemically decoupling the ligand from the unbound (Leg 1) and bound (Leg 2) states into the intermediate vacuum state, as

$$\Delta G_b^* = \Delta G_2 - \Delta G_1. \quad (9)$$

The Alchemical Transfer Method (ATM) proposed here avoids the vacuum intermediate and requires only one molecular system. As illustrated in Fig. 1B, ATM is based on an alchemical intermediate in which the ligand interacts simultaneously with the solvent bulk and the receptor. We assume, without loss of generality, that there is a suitable coordinate frame attached to the receptor and that the binding site region (represented by the dashed circle in Fig. 1) is fixed relative to this coordinate frame. Under these assumptions, every point in the binding site region maps to a unique point in a region of the same shape in the solvent bulk by means of a constant translation vector  $h$  (Fig. 2). The bound state of the system is defined as any configuration in which the ligand is in the binding site region. Conversely, the unbound state of the system is defined as any configuration of the system in which the ligand is placed into the bulk solvent region (Fig. 1B).

Under these assumptions, any configuration of the bound state maps to a unique configuration of the unbound state by a rigid translation of the ligand atoms by the vector  $h$  (Fig. 2). The reverse is also true, such that any configuration of the unbound system maps uniquely to a configuration of the bound state by translation of the vector  $-h$ . Hence, the translation vector  $h$  can be used as a perturbation parameter to connect, in a statistical thermodynamic sense, the bound and unbound states. The direct transfer of the bound

ligand to a remote bulk solvent location (or vice versa) circumvents the problems associated with physical barriers for ligand extraction and entry that often hamper physical pathway methods.<sup>22</sup>

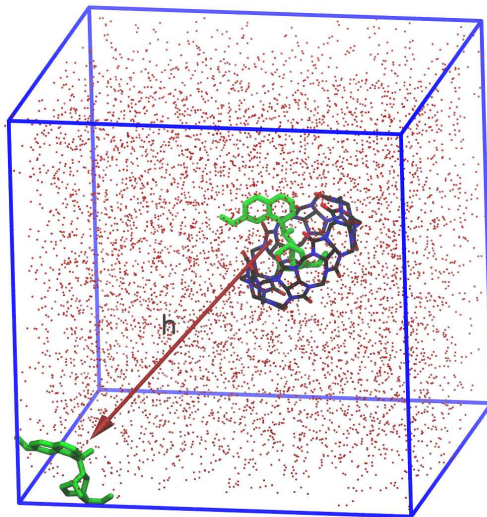


Figure 2: Illustration of the Alchemical Transfer step of leg 1 (Figure 1) to connect the bound and unbound states of the CB8-G3 complex. The CB8 host is located at the center. The bound and unbound G3 guests are in green. The unbound guest is obtained by translating the bound guest by the displacement vector  $h$  shown in dark red. The small red spheres surrounding the host and both guests in the box represent the oxygen atoms of the water molecules of the solvent. This particular configuration is extracted from the ensemble at the symmetric intermediate  $\lambda = 1/2$  state.

For example, consider Leg 1 of the ATM cycle in Fig. 1B. Denoting the potential energy function of the system as  $U(x)$ , with  $x = (x_S, x_L)$  being the coordinates of the bound system, where  $x_L$  are the coordinates of the ligand in the receptor binding site and  $x_S$  are

the coordinates of the receptor and the solvent (the surroundings), the potential energy functions  $U_0$  and  $U_1$  of the initial and final states of the leg (the bound and unbound states in this case) are respectively

$$U_0(x_S, x_L) = U(x_S, x_L) \quad (10)$$

and

$$U_1(x_S, x_L) = U(x_S, x_L + h). \quad (11)$$

To connect the bound and unbound states, we consider the hybrid alchemical potential of Eq. (1) with the perturbation energy

$$u_1(x_S, x_L) = U(x_S, x_L + h) - U(x_S, x_L), \quad (12)$$

which is defined as the change in the potential energy of the system for rigidly translating the ligand atoms from the binding site region to the bulk solvent region while all other degrees of freedom of the system remain unchanged.

With these definitions, the alchemical hybrid potential used for the first leg of the ATM cycle is:

$$\text{Leg 1 : } U_\lambda(x_S, x_L) = U(x_S, x_L) + W_\lambda(u_1), \quad 0 \leq \lambda \leq 1/2 \quad (13)$$

where the alchemical perturbation function  $W_\lambda(u)$  is defined in Eq. (3) and the perturbation energy  $u_1$  is defined in Eq. (12). As indicated in Eq. (13), the alchemical pathway for the first leg is terminated at  $\lambda = 1/2$ , where the ligand interacts with half strength with both the receptor environment and the solvent bulk. This ensures that severe steric clashes are not likely to occur at  $\lambda = 1/2$ . The  $u_0$  parameter of the soft-core perturbation potential is set to a large enough value so that the perturbation potential  $u$  does not exceed  $u_0$  at  $\lambda = 1/2$ . Under these conditions, the original and the soft-core perturbation potentials are

the same and it follows from Eq. (13) that the alchemical potential energy at  $\lambda = 1/2$  is

$$U_{1/2}(x_S, x_L) = \frac{1}{2}U(x_S, x_L) + \frac{1}{2}U(x_S, x_L + h). \quad (14)$$

Eq. (14) defines the potential energy function of the alchemical intermediate of the ATM cycle in Fig. 1B. Accordingly, the ligand which resides in the binding site interacts symmetrically with the receptor (via regular nonbonded interactions) and with the bulk solvent environments (because of the translation mapping vector  $h$ ) in the alchemical intermediate ensemble. Conversely in this state, the receptor atoms and the bulk solvent molecules interact with the ligand at half strength. The total momentum and angular momentum of the system are therefore always conserved. The alchemical calculation that corresponds to the alchemical potential Eq. (13) yields the free energy change  $\Delta G_1$  in going from the bound state to the ATM alchemical intermediate.

To formulate Leg 2 of the ATM cycle connecting the unbound state to the intermediate state (Fig. 1), the role of the end states is reversed relative to the first leg. As before,  $x_L$  describes the coordinates of the ligand in the receptor binding site and  $x_L + h$  describes the coordinates of the ligand in the solvent bulk and the initial and final states are now defined by the potential energy functions of, respectively, the unbound and bound states:

$$U_0(x_S, x_L) = U(x_S, x_L + h) \quad (15)$$

and

$$U_1(x_S, x_L) = U(x_S, x_L). \quad (16)$$

The alchemical potential energy function for the second leg of the ATM cycle is

$$\text{Leg 2 : } U_\lambda(x_S, x_L) = U(x_S, x_L + h) + W_\lambda(u_2), \quad 0 \leq \lambda \leq 1/2 \quad (17)$$

with the perturbation energy

$$u_2(x_S, x_L) = U(x_S, x_L) - U(x_S, x_L + h) \quad (18)$$

that corresponds to the potential energy change of transferring the ligand from the bulk to the receptor binding site. Note that under the same assumptions that led to Eq. (14), Eq. (17) reaches  $\lambda = 1/2$  at the same alchemical intermediate as Eq. (13). Thus, Eq. (9) holds for the ATM thermodynamic cycle in Fig. 1B. Note also that the ligand intramolecular interactions are present along the thermodynamic pathway in ATM. This in contrast to the situations in DDM, where for large ligands the intramolecular ligand interactions often need to be turned off, since such interactions can lead to kinetically trapped vacuum conformations (Mark Abraham, Berk Hess, David van der Spoel, and Erik Lindahl, GROMACS Reference Manual, 2016). However, the need to turn off the ligand intramolecular interactions introduces additional perturbation to DDM calculations and hampers the convergence of the calculation.

We end the presentation of the Alchemical Transfer Method by discussing the requirement of splitting the alchemical path into two legs. Although the symmetric hybrid potential Eq. (14) can formally cover the direct path from the unbound to the bound states by extending the  $\lambda$  range from 0 to 1, in practice, it suffers severe end-point singularities at both of the end states. The potential energy when the ligand is placed in the solvent bulk region,  $U(x_S, x_L + h)$ , is ill-defined near  $\lambda = 0$  due to the atomic clashes that occur when ligand-solvent interactions are turned off. Conversely, near  $\lambda = 1$ , the potential energy when the ligand is placed in the receptor site,  $U(x_S, x_L)$ , is ill-defined due to clashes between ligand atoms and receptor atoms. The soft-core protocol we employ is based on an asymmetric definition of the perturbation potential [see Eqs. (12) and (18)] that can address the singularity at one end-point or the other but not both simultaneously with only one continuous alchemical perturbation potential energy function.

## Software Implementation

The method is implemented as an integrator plugin (`github.com/rajatkrpal/openmm_sdm_plugin`)<sup>24</sup> of the OpenMM library.<sup>25</sup> The integrator is based on the Langevin thermostat and high-level routines that displace the ligand, issue calls to energy and forces calculation routines, and compute the alchemical potential energy [Eq. (3)] and its gradients by combining the returned system energies and forces. For Leg 1 for example, the plugin first computes the potential energy and the forces at each MD step when the ligand is in the binding site and saves them in temporary buffers. Then it displaces the ligand in the bulk by the displacement vector  $h$  and recalculates the energy and forces (except for the binding site restraining potential, see below). The perturbation energy,  $u_1$  [Eq. (12)], and its gradients are obtained by taking the corresponding differences before and after the displacement. The gradients of the alchemical perturbation energy [Eq. (13)] are derived from those of  $u_1$  and of the undisplaced potential  $U(x_S, x_L)$  by application of the chain rule. The resulting forces are then used to propagate atomic coordinates by one MD step. The same process is used for Leg 2, except that the ligand is initially placed in the solvent bulk and it is translated into the binding site by reversing the displacement. In each case, the binding restraint potential (see below) is applied when the ligand is in the binding site. No modifications of the core OpenMM energy routines are applied.

## Benchmark Systems

We validated ATM on the stringent SAMPL6 SAMPLing host-guest benchmark set.<sup>27</sup> The set includes pre-configured systems of the well-studied octa-acid (OA) and cucurbit[8]uril (CB8) hosts that have been featured in previous host-guest binding SAMPL challenges. The set includes three small-molecule guests resembling conventional drug-like fragments. The SAMPLing benchmark set includes the 5-hexenoic acid (OA-G3) and 4-methylpentanoic acid (OA-G6) for the OA host, and quinine (CB3-G3) for the CB3 host, which are indicated in Fig. 3. Despite the name, the G3 guest in the CB8-G3 complex is distinct from the G3

guest in the OA-G3 complex. The parametrized systems, including their solvent descriptions, partial charges, and initial geometries, are provided at the github SAMPL6 SAMPLing site: ([https://github.com/samplchallenges/SAMPL6/tree/master/host\\_guest/SAMPLing](https://github.com/samplchallenges/SAMPL6/tree/master/host_guest/SAMPLing)).

Host-guest systems are practical alternatives to protein-ligand systems for benchmarking novel computational methods because of their minimal atom count and the improbability of the molecules undergoing major conformational reorganization. These systems thus enable the investigation of novel simulation techniques, such as longer timescales and appropriate sampling of multiple equivalent binding modes. Each host-guest complex is provided with five initial conformations that differ in the position of the guest within the binding site, as well as in the conformation of the guest. The procedure of obtaining five replicate free energy calculations allows for assessing the statistical uncertainty and reproducibility of the methodology of ATM. The benchmark systems employed here create an accessible platform to improve both the predictive accuracy and computational efficiency of free energy calculations.<sup>27</sup>

## Computational Details

The ATM calculations employed the host-guest molecular complexes parametrized using the GAFF1.8/AM1-BCC force-field and the TIP3P water model as provided by Rizzi et al.<sup>27</sup> The Cartesian components of the displacement vector  $h$  were set to approximately half the dimensions of the simulation box in order to ensure that the ligand is placed in the corner of the solvent box (Fig. 2). This position is the farthest from the host, which the solvent box is centered around, and its periodic images. The complexes were energy minimized and thermalized at 300 K. Then, using the ATM alchemical potential energy function for Leg 1 [Eq. (13)] and starting at the bound state at  $\lambda = 0$ , the systems were annealed to the symmetric intermediate  $\lambda = 1/2$  for 250 ps. The purpose of this step is to obtain a suitable initial configuration without severe unfavorable repulsion interactions at either end of the alchemical paths in order to start the molecular dynamics replica exchange alchemical

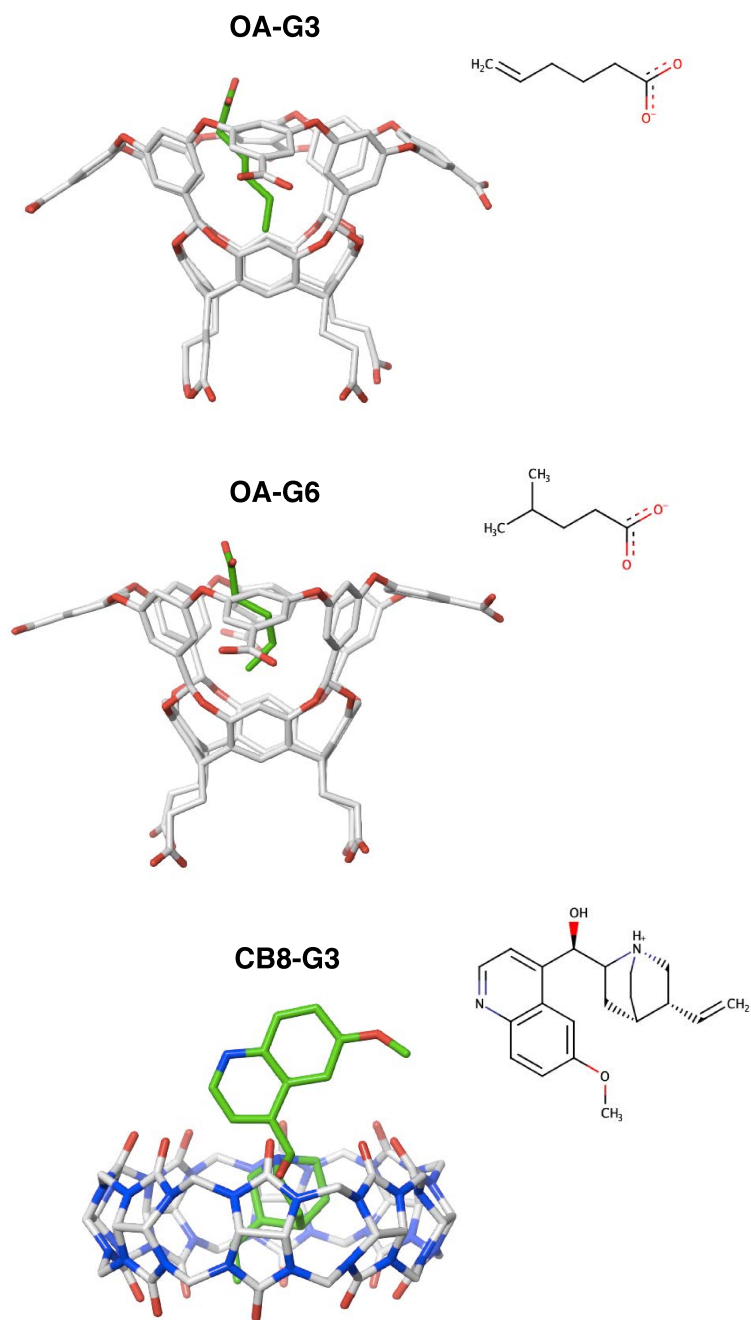


Figure 3: Benchmark SAMPL6 host-guest systems applied in this study. Guests for the OA hosts include 5-hexenoic acid and 4-methylpentanoic acid, respectively, and that of CB8 includes quinine. Red atoms represent oxygen, blue nitrogen, and white carbon. The carbon atoms of the guests are represented in green.

calculations for each leg (see below).

To limit the fluctuations of the position of the bulk solvent region, which would impact convergence, the position and orientation of the receptor was loosely restrained with a flat-bottom harmonic potential of force constant 25.0 kcal/(mol Å<sup>2</sup>) and a tolerance of 1.5 Å on all of the heavy atoms of the lower portion of the receptor (the first 40 atoms of the host as listed in the provided files).

Polar hydrogen atoms with zero Lennard-Jones parameters were modified to  $\sigma_{\text{LJ}} = 0.1$  Å and  $\epsilon_{\text{LJ}} = 10^{-4}$  kcal/mol to avoid large attractive interactions between opposite charges at small distances in nearly uncoupled states.<sup>26</sup> We established that the change in potential energy of the system in the unbound, bound, and symmetric intermediate states due to this modification of the Lennard-Jones parameters is below single floating point precision. Alchemical MD calculations employed the OpenMM<sup>25</sup> MD engine and the SDM integrator plugin ([github.com/rajatkrpal/openmm\\_sdm\\_plugin.git](https://github.com/rajatkrpal/openmm_sdm_plugin.git)) using the OpenCL platform. The ASyncRE software,<sup>28</sup> customized for OpenMM and SDM ([github.com/egallicc/async\\_re-openmm.git](https://github.com/egallicc/async_re-openmm.git)), was used for the Hamiltonian Replica Exchange in  $\lambda$  space for each ATM leg.

The linear alchemical perturbation potential,  $W_\lambda(u) = \lambda u_{\text{sc}}(u)$ , corresponding to Eq. (3) with  $\lambda_1 = \lambda_2 = \lambda$ , was used for the octa-acid (OA) systems with 11  $\lambda$ -states uniformly distributed between  $\lambda = 0$  and  $1/2$  for each of the two ATM legs. The ATM calculations for the more challenging CB8-G3 complex employed the softplus perturbation potential Eq. (3) with 24  $\lambda$ -states and the parameters listed in Tables 1 and 2 for each leg. The parameters of the softplus perturbation potential were optimized using the scheme described previously,<sup>24,26</sup> which involved running trial calculations with the linear potential, fitting the analytical model of binding<sup>29</sup> to each transformation, and adjusting the parameters of the softplus potential to the resulting  $\lambda$ -function<sup>24</sup> to obtain a smooth alchemical transition.

The soft-core perturbation energy Eq. (4) was used for all calculations with  $u_{\text{max}} = 300$  kcal/mol and  $u_0 = 100$  kcal/mol. The ligand was sequestered within the binding site by

Table 1: Alchemical schedule of the softplus perturbation function for leg 1 for the CB8/G3 complex.

| $\lambda$ | $\lambda_1$ | $\lambda_2$ | $\alpha^a$ | $u_0^b$ | $w_0^b$ |
|-----------|-------------|-------------|------------|---------|---------|
| 0.000     | 0.000       | 0.000       | 0.250      | 230.0   | 0.000   |
| 0.022     | 0.000       | 0.050       | 0.250      | 230.0   | 0.000   |
| 0.043     | 0.000       | 0.100       | 0.250      | 220.0   | 0.000   |
| 0.065     | 0.000       | 0.150       | 0.250      | 215.0   | 0.000   |
| 0.087     | 0.000       | 0.200       | 0.250      | 210.0   | 0.000   |
| 0.109     | 0.000       | 0.250       | 0.250      | 205.0   | 0.000   |
| 0.130     | 0.000       | 0.300       | 0.250      | 200.0   | 0.000   |
| 0.152     | 0.000       | 0.350       | 0.250      | 195.0   | 0.000   |
| 0.174     | 0.000       | 0.400       | 0.250      | 188.0   | 0.000   |
| 0.196     | 0.000       | 0.400       | 0.250      | 178.0   | 0.000   |
| 0.217     | 0.000       | 0.400       | 0.200      | 170.0   | 0.000   |
| 0.239     | 0.000       | 0.400       | 0.150      | 160.0   | 0.000   |
| 0.261     | 0.000       | 0.400       | 0.150      | 152.0   | 0.000   |
| 0.283     | 0.000       | 0.425       | 0.140      | 145.0   | 0.000   |
| 0.304     | 0.000       | 0.450       | 0.130      | 140.0   | 0.000   |
| 0.326     | 0.000       | 0.475       | 0.140      | 135.0   | 0.000   |
| 0.348     | 0.000       | 0.500       | 0.150      | 128.0   | 0.000   |
| 0.369     | 0.000       | 0.500       | 0.200      | 120.0   | 0.000   |
| 0.391     | 0.000       | 0.500       | 0.200      | 112.0   | 0.000   |
| 0.413     | 0.100       | 0.500       | 0.200      | 110.0   | 0.000   |
| 0.435     | 0.200       | 0.500       | 0.150      | 110.0   | 0.000   |
| 0.457     | 0.300       | 0.500       | 0.150      | 110.0   | 0.000   |
| 0.478     | 0.400       | 0.500       | 0.100      | 110.0   | 0.000   |
| 0.500     | 0.500       | 0.500       | 0.100      | 110.0   | 0.000   |

<sup>a</sup>In (kcal/mol)<sup>-1</sup>. <sup>b</sup>In kcal/mol.

means of a flat-bottom harmonic potential between the centers of mass of the host and the ligand with a force constant of 25 kcal/mol Å<sup>2</sup> applied for separation greater than 4.5 Å. Perturbation energy samples and trajectory frames were saved every 5 ps. Hamiltonian replica exchanges in  $\lambda$ -space were performed every 5 ps. The Langevin thermostat with a time constant of 2 ps was used to maintain the temperature at 300 K. Each replica was simulated for a minimum of 10 ns. Binding free energies and the corresponding uncertainties were computed from the perturbation energy samples using UWHAM,<sup>30</sup> discarding the first 5 ns of trajectory, followed by the addition of the concentration-dependent term  $\Delta G_{\text{site}}^{\circ} =$

Table 2: Alchemical schedule of the softplus perturbation function for leg 2 for the CB8/G3 complex.

| $\lambda$ | $\lambda_1$ | $\lambda_2$ | $\alpha^a$ | $u_0^b$ | $w_0^b$ |
|-----------|-------------|-------------|------------|---------|---------|
| 0.000     | 0.000       | 0.000       | 0.250      | 175     | 0.000   |
| 0.022     | 0.000       | 0.084       | 0.250      | 172.5   | 0.000   |
| 0.043     | 0.000       | 0.167       | 0.250      | 170     | 0.000   |
| 0.065     | 0.000       | 0.233       | 0.225      | 165     | 0.000   |
| 0.087     | 0.000       | 0.333       | 0.200      | 160     | 0.000   |
| 0.109     | 0.000       | 0.417       | 0.175      | 155     | 0.000   |
| 0.130     | 0.000       | 0.500       | 0.150      | 150     | 0.000   |
| 0.152     | 0.000       | 0.550       | 0.125      | 145     | 0.000   |
| 0.174     | 0.000       | 0.600       | 0.100      | 140     | 0.000   |
| 0.196     | 0.000       | 0.650       | 0.095      | 135     | 0.000   |
| 0.217     | 0.000       | 0.700       | 0.090      | 130     | 0.000   |
| 0.239     | 0.000       | 0.700       | 0.085      | 125     | 0.000   |
| 0.261     | 0.000       | 0.700       | 0.080      | 120     | 0.000   |
| 0.283     | 0.000       | 0.700       | 0.075      | 112     | 0.000   |
| 0.304     | 0.050       | 0.700       | 0.070      | 110     | 0.000   |
| 0.326     | 0.100       | 0.700       | 0.0675     | 102.5   | 0.000   |
| 0.348     | 0.150       | 0.700       | 0.065      | 105     | 0.000   |
| 0.369     | 0.200       | 0.700       | 0.060      | 100     | 0.000   |
| 0.391     | 0.250       | 0.650       | 0.055      | 100     | 0.000   |
| 0.413     | 0.300       | 0.625       | 0.050      | 100     | 0.000   |
| 0.435     | 0.350       | 0.600       | 0.045      | 100     | 0.000   |
| 0.457     | 0.400       | 0.575       | 0.040      | 100     | 0.000   |
| 0.478     | 0.450       | 0.550       | 0.035      | 100     | 0.000   |
| 0.500     | 0.500       | 0.500       | 0.030      | 100     | 0.000   |

<sup>a</sup>In (kcal/mol)<sup>-1</sup>. <sup>b</sup>In kcal/mol.

$-k_B T \ln C^\circ V_{\text{site}} = 0.87$  kcal/mol that corresponds to 300 K temperature and the volume  $V_{\text{site}}$  of a sphere of radius 4.5 Å. The replica exchange simulations were run on the XSEDE Comet GPU HPC cluster at the San Diego Supercomputing Center, each using four GPUs per node.

Equilibration analysis (Fig. 4) was based on monitoring the excess binding free energy estimates [Eq. 9] averaged over the five starting conditions as a function of the time  $t_{\text{eq}}$  discarded from the beginning of the trajectories of each replica-exchange thread. The uncertainty of each estimate was taken as the 95% confidence interval of the mean based on

Table 3: Standard binding free energy estimates and corresponding computational effort for the three host-guest complexes with the Alchemical Transfer Method compared to experimental and reference computed values.

| Complex | $\Delta G_b^\circ(\text{exp})^{a,b}$ | $\Delta G_b^\circ(\text{ref})^{a,c,d}$ | $n_{\text{eval}}(\text{ref})^{c,e}$ | $\Delta G_b^{\circ a,e}$ | $n_{\text{eval}}^f$ |
|---------|--------------------------------------|----------------------------------------|-------------------------------------|--------------------------|---------------------|
| OA-G3   | $-5.18 \pm 0.02$                     | $-6.3 \pm 0.1$                         | $458 \times 10^6$                   | $-5.88 \pm 0.41$         | $220 \times 10^6$   |
| OA-G6   | $-4.97 \pm 0.02$                     | $-6.8 \pm 0.1$                         | $305 \times 10^6$                   | $-6.43 \pm 0.17$         | $220 \times 10^6$   |
| CB8-G3  | $-6.45 \pm 0.06$                     | $-10.5 \pm 0.6$                        | $2135 \times 10^6$                  | $-8.47 \pm 0.61$         | $480 \times 10^6$   |

<sup>a</sup>In kcal/mol. <sup>b</sup>From references 33 and 34. <sup>c</sup>From reference 27. <sup>d</sup>APR method. <sup>e</sup>This work, from Tables 4, 5, and 6. <sup>f</sup>This work.

the t-test distribution with 4 degrees of freedom. A valid equilibration time is such that the estimate at that time is statistically consistent with the estimates at all later times.<sup>23,31,32</sup> In this work, we established that an equilibration time of 5 ns, corresponding to half of the trajectory of each replica, is suitable for all three systems.

## Results

The standard binding free energy estimates,  $\Delta G_b^\circ$ , for the host-guest systems obtained using the ATM method are listed in Table 3. These results are compared with the corresponding experimental measurements,  $\Delta G_b^\circ(\text{exp})$ , as well as the reference computational estimates obtained through the attach-pull-release (APR) methodology.<sup>27</sup> Table 3 also lists the number of energy and forces evaluations per replicate (a proxy for the computational cost) for the reference APR calculations,  $n_{\text{eval}}(\text{ref})$ , and for the ATM calculations reported here ( $n_{\text{eval}}$ ). The APR method was selected as representative for this comparison because APR, similar to ATM and unlike DDM, displaces the ligand in a position in the solvent bulk at a finite distance from the host.

The ATM binding free energy estimates reported in Table 3 are obtained as the average of five replicates starting from different initial conformations as reported in Tables 4, 5, and 6. These tables also report the calculated free energies,  $\Delta G_1$  and  $\Delta G_2$ , of the two alchemical legs for each replicate. The binding free energy of each replicate is the difference between

Table 4: Free energy estimates for the two legs of the Alchemical Transfer Method and corresponding standard binding free energy estimates for the OA-G3 complex starting with each of the the five initial SAMPL6 SAMPLing conformations.

| Conformation | $\Delta G_1^a$   | $\Delta G_2^a$   | $\Delta G_{\text{site}}^{\circ a}$ | $\Delta G_b^{\circ a}$ |
|--------------|------------------|------------------|------------------------------------|------------------------|
| OA-G3-0      | 57.00            | 50.62            | 0.87                               | -5.51                  |
| OA-G3-1      | 57.94            | 51.03            | 0.87                               | -6.04                  |
| OA-G3-2      | 57.80            | 50.91            | 0.87                               | -6.02                  |
| OA-G3-3      | 57.59            | 51.17            | 0.87                               | -5.55                  |
| OA-G3-4      | 57.26            | 50.11            | 0.87                               | -6.27                  |
| Average      | $57.52 \pm 0.48$ | $50.77 \pm 0.52$ | 0.87                               | $-5.88 \pm 0.41$       |

<sup>a</sup>In kcal/mol. <sup>b</sup>With t-test 95% confidence intervals with 4 degrees of freedom based on the standard deviation of the mean.

Table 5: Free energy estimates for the two legs of the Alchemical Transfer Method and corresponding standard binding free energy estimates for the OA-G6 complex starting with each of the the five initial SAMPL6 SAMPLing conformations.

| Conformation | $\Delta G_1^a$   | $\Delta G_2^a$   | $\Delta G_{\text{site}}^{\circ a}$ | $\Delta G_b^{\circ a}$ |
|--------------|------------------|------------------|------------------------------------|------------------------|
| OA-G6-0      | 57.88            | 50.66            | 0.87                               | -6.34                  |
| OA-G6-1      | 58.26            | 50.73            | 0.87                               | -6.66                  |
| OA-G6-2      | 58.33            | 51.03            | 0.87                               | -6.43                  |
| OA-G6-3      | 58.70            | 51.50            | 0.87                               | -6.33                  |
| OA-G6-4      | 58.31            | 51.04            | 0.87                               | -6.39                  |
| Average      | $58.29 \pm 0.36$ | $50.99 \pm 0.41$ | 0.87                               | $-6.43 \pm 0.17$       |

<sup>a</sup>In kcal/mol. <sup>b</sup>With t-test 95% confidence intervals with 4 degrees of freedom based on the standard deviation of the mean.

those of the two legs (the excess component) plus the standard state term  $\Delta G_{\text{site}}^{\circ}$ , which in this case measures out to be approximately 0.87 (see Computational Details). The statistical fluctuations of the binding free energies among the five conformations for each system were consistently smaller than those of each of the legs, suggesting some level of systematic error cancellation.

The ATM results obtained for the OA-G3 and OA-G6 complexes are in good agreement (within 0.5 kcal/mol) with the reference values and well within the range of estimates obtained with other methods.<sup>27</sup> The ATM binding free energy estimate for the more challenging CB-G3 complex deviates more substantially from the APR reference (2 kcal/mol less

Table 6: Free energy estimates for the two legs of the Alchemical Transfer Method and corresponding standard binding free energy estimates for the CB8-G3 complex starting with each of the the five initial SAMPL6 SAMPLing conformations.

| Conformation | $\Delta G_1^a$   | $\Delta G_2^a$   | $\Delta G_{\text{site}}^{\circ a}$ | $\Delta G_b^{\circ a}$ |
|--------------|------------------|------------------|------------------------------------|------------------------|
| CB8-G3-0     | 70.50            | 61.06            | 0.87                               | -8.58                  |
| CB8-G3-1     | 70.65            | 60.95            | 0.87                               | -8.83                  |
| CB8-G3-2     | 71.41            | 61.56            | 0.87                               | -8.98                  |
| CB8-G3-3     | 71.82            | 62.81            | 0.87                               | -8.14                  |
| CB8-G3-4     | 69.09            | 60.42            | 0.87                               | -7.80                  |
| Average      | $70.70 \pm 1.30$ | $61.36 \pm 1.12$ | 0.87                               | $-8.47 \pm 0.61$       |

<sup>a</sup>In kcal/mol. <sup>b</sup>With t-test 95% confidence intervals with 4 degrees of freedom based on the standard deviation of the mean.

favorable) and from those of the other methods applied to this benchmark system.<sup>27</sup> The origin of this discrepancy is not obvious. However all ATM estimates appear to generally underestimate binding affinities relative to the other methods and thus bringing them, perhaps coincidentally, closer to the experimental measurements. The deviations between the experimental standard binding free energies and the ATM estimates are 0.70 kcal/mol, 1.46 kcal/mol, and 2.02 kcal/mol for the OA-G3, OA-G6, and CB8-G3 complexes, respectively, compared to 1.12 kcal/mol, 1.83 kcal/mol, and 4.05 kcal/mol with APR.

The range of the spread between ATM replicates obtained here for the octa-acid systems is generally larger than with APR and other methods,<sup>27</sup> albeit at a generally higher computational cost. For CB8-G3, ATM yields a spread similar to APR and the other methods with significantly less computational cost (480 vs. 2,135 million energy evaluations as compared to APR, Table 3). We also note that the free energy differences for both legs shown in Table 4 and Table 5 are about 50 kcal/mol smaller than the values of decoupling free energies for similarly charged carboxylic molecules computed using DDM.<sup>8</sup>

Reverse cumulative equilibration profiles<sup>23,31,32</sup> of the ATM excess binding free estimates of the host-guest systems show a relatively weak dependence on the simulation production length (Fig. 4). For the octa-acid systems OA-G3 and OA-G6, the estimates vary within a small fraction of a kcal/mol when discarding from 0% to 75% of the initial data of 10 ns

trajectories of each replica-exchange replica. The estimate for the more challenging CB8-G3 system displays greater fluctuations, in addition to a likely systematic drift towards more negative values at equilibration times longer than 6 ns per replica (Fig. 4). Preliminary conformational analysis (not shown) suggests that the downward drift could be due to the late rebinding of the ligand in the alternative bound state in which G3 is flipped vertically with respect to the symmetry plane of the CB8 host relative to the starting conformation of the complex (Fig. 3).

The equilibration profiles are used to determine the time after which the simulation overcomes the bias associated to the starting conformation of each system. The binding free energy estimates reported here correspond to an equilibration time that yields a free energy value that is statistically consistent with the subsequent equilibration times. As shown in Fig. 4 (red horizontal line), according to this analysis, all three systems achieve equilibration after 5 ns or less of simulation time per replica, corresponding to the values presented in Tables 3–6.

In the equilibration analysis, the uncertainty of each estimate is obtained from the spread of the binding free energy from each of the five replicates as described in Computational Details. Generally, uncertainties increase with increasing  $t_{\text{eq}}$  due to the progressively fewer samples included in the estimate. However, the spread among the replicates of the OA-G6 system display an unusual non-monotonic behavior (Fig. 4) reaching nearly zero at 4.3 ns before increasing at later equilibration times. It appears that this a random occurrence as a result of the small number of runs in which the five replicates happen, by chance, to independently converge near a single value of the binding free energy.

Overall, ATM yields standard binding free energy estimates within the general range displayed by the established methods tested on the SAMPL6 SAMPLing benchmark set at a similar computational expense.<sup>27</sup> These initial results confirm the validity of the ATM approach.

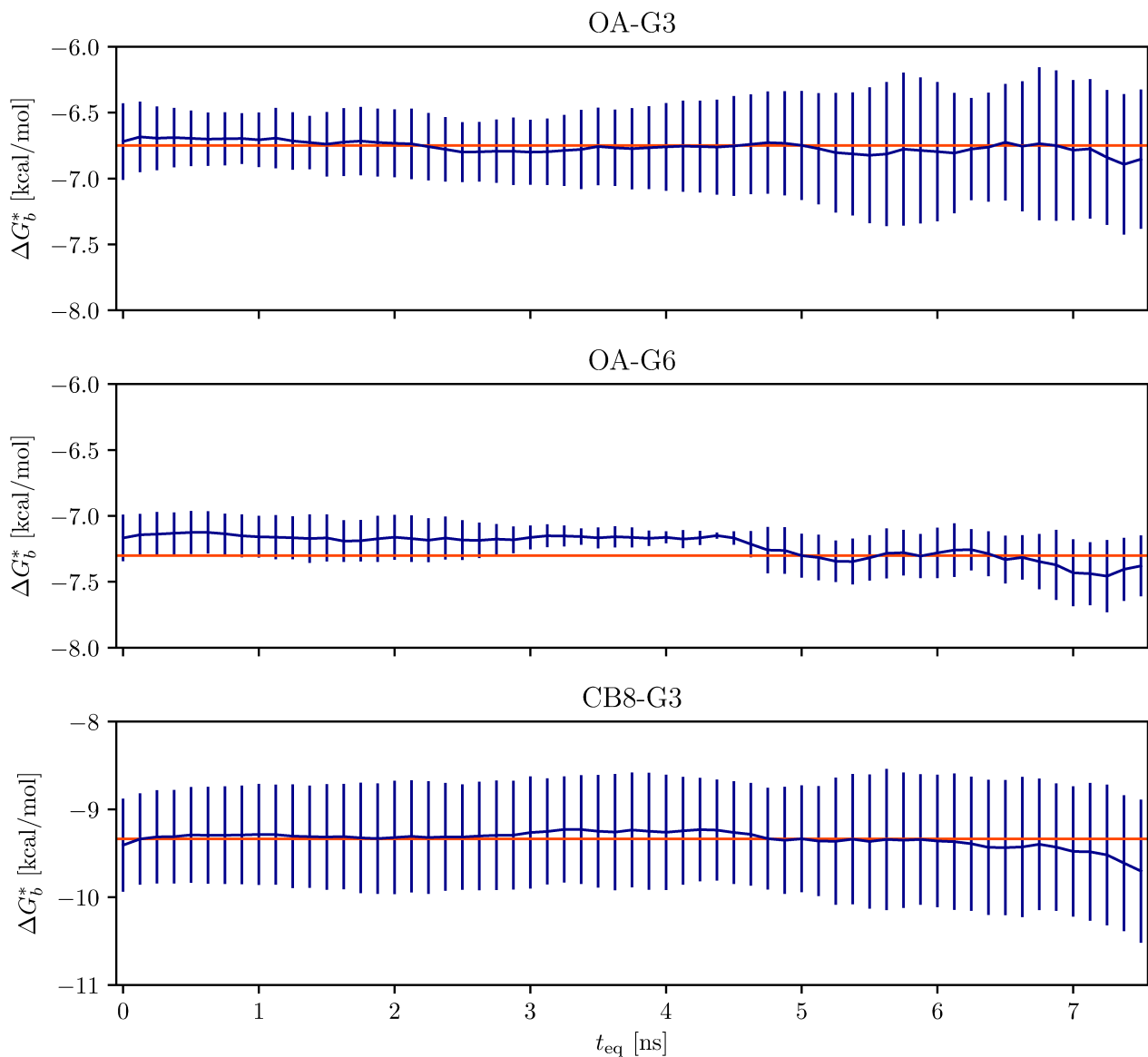


Figure 4: Reverse cumulative equilibration profiles<sup>23</sup> of the standard binding free energy estimates of the SAMPL6 SAMPLing host-guest complexes calculated as described in Computational Details. The horizontal line corresponds to the binding free energy estimate at  $t_{\text{eq}} = 5$  ns reported in Table 3. These estimates are within the statistical uncertainties of the estimates at all later equilibration times.

# Discussion

The Alchemical Transfer Method (ATM) presented here implements a perturbation potential based on rigidly displacing the coordinates of the ligand atoms from a region in the solvent bulk to the receptor binding site and vice versa. Like in smart-darting Monte Carlo<sup>35</sup> and the generalized alteration of structure and parameters (GASP) method,<sup>36</sup> this is accomplished using a one-to-one map relating the coordinates of the two conformational macrostates of interest. In ATM, a simple coordinate displacement is used to design a thermodynamic pathway in which the ligand disappears from one place and appears in another in a way that is physically not achievable or, in other words, "alchemical". The change in potential energy of the system due to the displacement of the ligand is the perturbation energy of the  $\lambda$ -dependent alchemical potential energy function that is used for conformational sampling and free energy estimation. While not presented here, the method is applicable to the calculation of the relative free energy of binding between two ligands<sup>14</sup> by swapping their positions in the bulk and in the receptor site. This work is ongoing and will be reported in a forthcoming publication.

Similarly to physical pathway methods,<sup>20,22,37</sup> ATM is relatively easy to implement in molecular simulation packages because it does not require modifications of system parameters nor customized single- and dual-topologies setups that characterize conventional alchemical binding free energy methods. The method is illustrated here using a plugin implementation on top of the core OpenMM library.<sup>25</sup> The method is agnostic of the underlying energy function. It has been validated here with explicit solvation and Particle Mesh Ewald (PME) long range electrostatics. It is conceivably applicable without approximations to any kind of many-body potential function, including polarizable,<sup>16,38</sup> quantum-mechanical,<sup>39-41</sup> and artificial neural network<sup>42,43</sup> potentials.

Unlike alchemical approaches such as double-decoupling,<sup>6</sup> which require two systems, and dual-system single box alchemical methods,<sup>44-47</sup> which require dual topologies, ATM works with a single standard model of the receptor-ligand complex solvated in a solvent box as in

conventional molecular dynamics applications. In addition, ATM does not require soft-core pair potentials nor the splitting of the alchemical transformations into separate electrostatic and steric/dispersion steps.<sup>24,26</sup>

Similar to single-box alchemical approaches<sup>44-47</sup> ATM avoids alchemical transformations that place the ligand in vacuum,<sup>6,48</sup> which are particularly problematic for large and charged ligands.<sup>22</sup> The desolvation step of double-decoupling, for example, removes all of the ligand-water interactions, even though those of the solvent-exposed region of the ligand are likely to form again in the subsequent coupling step with the receptor. With ATM, instead, existing hydration interactions in the bulk are more likely to be replaced by similar interactions with the ligand displaced into the binding site. Moreover, unless the ligand is properly restrained, the vacuum intermediate is likely to introduce hard to converge free energy terms related to the reorganization of the ligand conformational ensemble from vacuum to the solvated environment.

ATM has some drawbacks, some of which are technical in nature and are likely to be addressed in the future. Because it calculates the system energy and forces twice for each MD time-step,<sup>4</sup> once with the ligand in the bulk and again with the ligand in the receptor pocket, the method is a factor of two slower per step than standard molecular dynamics. The two energy evaluations are however independent and can be conceivably run in parallel on two attached computational devices for added performance. Currently the method also requires the recalculation of the non-bonded neighbor list after each ligand displacement, resulting in an additional 10 to 15% slow-down per step with OpenMM for these systems. As we observed here for the CB8-G3 system, the binding of bulky ligands requires optimized softplus alchemical perturbation functions trained on trial calculations with the linear alchemical potential.<sup>24,26,29</sup> In future work, we plan to implement an adaptive algorithm to refine the parameters of the alchemical potential function on the fly.

Here we have validated ATM on the rigorous SAMPL6 SAMPLing benchmark.<sup>27</sup> The dataset includes well-studied systems prepared with a single set of force field parameters

and in different initial conformations to probe both systematic and statistical errors. The binding free energies of the systems have been computed and validated with a diverse collection of approaches, including alchemical and physical pathway methods.<sup>27</sup> The ATM results for the octa-acid systems obtained here are well within the range of estimates reported in reference 27, thereby adding confidence that the ATM approach is sound and that it has been implemented correctly. In particular, we observed good agreement with the Attach Pull Release (APR) method,<sup>49</sup> a physical pathway approach in which the guest is progressively displaced into the solvent bulk to a comparable distance from the host as in this work. ATM yields a statistically significant, less favorable binding free energy estimate than the other methods for the more challenging CB8-G3 system. The source of the deviation is unclear, however, ATM appears to generally yield binding free energies of smaller magnitude and closer to the experimental measurements than the other methods. Taking into account the relative computational expense, the statistical uncertainties obtained here indicate that ATM estimates have a comparable level of reproducibility and computational efficiency as the methods tested in reference 27.

## Conclusions

We have presented the Alchemical Transfer Method (ATM) for the calculation of standard binding free energies of non-covalent molecular complexes. The method is based on a coordinate displacement perturbation of the ligand between the receptor binding site and the bulk solvent and a thermodynamic cycle connected by a symmetric intermediate in which the ligand interacts with the receptor and solvent environments equally. While the approach is alchemical, ATM’s implementation is as straightforward as physical pathway methods of binding. ATM does not require splitting the alchemical transformations into electrostatic and non-electrostatic steps and it does not employ soft-core pair potentials. We have implemented ATM as a freely available and open-source plugin of the OpenMM molecular

dynamics library. The method and its implementation have been validated on the SAMPL6 SAMPLing host-guest benchmark set.

## Acknowledgments

We acknowledge support from the National Science Foundation (NSF CAREER 1750511 to E.G.). Molecular simulations were conducted on the Comet GPU supercomputer cluster at the San Diego Supercomputing Center supported by NSF XSEDE award TG-MCB150001.

## References

- (1) Simonson, T. The physical basis of ligand binding. *In Silico Drug Discovery and Design* **2016**, 3–43.
- (2) Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc Chem Res* **2009**, 42, 724–733.
- (3) Gallicchio, E.; Levy, R. M. Recent Theoretical and Computational Advances for Modeling Protein-Ligand Binding Affinities. *Adv. Prot. Chem. Struct. Biol.* **2011**, 85, 27–80.
- (4) Gallicchio, E.; Lapelosa, M.; Levy, R. M. Binding Energy Distribution Analysis Method (BEDAM) for Estimation of Protein-Ligand Binding Affinities. *J. Chem. Theory Comput.* **2010**, 6, 2961–2977.
- (5) Pal, R. K.; Ramsey, S.; Gadhiya, S.; Cordone, P.; Wickstrom, L.; Harding, W. W.; Kurtzman, T.; Gallicchio, E. Inclusion of Enclosed Hydration Effects in the Binding Free Energy Estimation of Dopamine D3 Receptor Complexes. *PLoS One* **2019**, 14, e0222092.
- (6) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-

- Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72*, 1047–1069.
- (7) Heinzelmann, G.; Gilson, M. K. Automation of absolute protein-ligand binding free energy calculations for docking refinement and compound evaluation. *Scientific Reports* **2021**, *11*, 1–18.
  - (8) Deng, N.; Cui, D.; Zhang, B. W.; Xia, J.; Cruz, J.; Levy, R. Comparing alchemical and physical pathway methods for computing the absolute binding free energy of charged ligands. *Phys. Chem. Chem. Phys.* **2018**, *20*, 17081–17092.
  - (9) Rocklin, G. J.; Mobley, D. L.; Dill, K. A.; Hünenberger, P. H. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *The Journal of chemical physics* **2013**, *139*, 11B606.1.
  - (10) Pan, A. C.; Xu, H.; Palpant, T.; Shaw, D. E. Quantitative characterization of the binding and unbinding of millimolar drug fragments with molecular dynamics simulations. *J. Chem. Theory Comput.* **2017**, *13*, 3372–3377.
  - (11) Ohlknecht, C.; Perthold, J. W.; Lier, B.; Oostenbrink, C. Charge-Changing Perturbations and Path Sampling via Classical Molecular Dynamic Simulations of Simple GuestHost Systems. *J. Chem. Theory Comput.* **2020**, Article ASAP.
  - (12) Lee, T.-S.; Lin, Z.; Allen, B. K.; Lin, C.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Sherman, W.; York, D. M. Improved Alchemical Free Energy Calculations with Optimized Smoothstep Softcore Potentials. *J. Chem. Theory Comput.* **2020**,
  - (13) Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J Chem Phys* **2007**, *127*, 214108.

- (14) Mey, A. S. J. S.; Allen, B. K.; Macdonald, H. E. B.; Chodera, J. D.; Hahn, D. F.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living Journal of Computational Molecular Science* **2020**, *2*, 18378.
- (15) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee Jr, T. D.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J. Chem. Inf. Model.* **2020**,
- (16) Harger, M.; Li, D.; Wang, Z.; Dalby, K.; Lagardère, L.; Piquemal, J.-P.; Ponder, J.; Ren, P. Tinker-OpenMM: Absolute and relative alchemical free energies using AMOEBA on GPUs. *J. Comp. Chem.* **2017**, *38*, 2047–2055.
- (17) Darden, T. A.; York, D. M.; Pedersen, L. G. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (18) Gallicchio, E.; Paris, K.; Levy, R. M. The AGBNP2 Implicit Solvation Model. *J. Chem. Theory Comput.* **2009**, *5*, 2544–2564.
- (19) Zou, J.; Tian, C.; Simmerling, C. Blinded prediction of protein–ligand binding affinity using Amber thermodynamic integration for the 2018 D3R grand challenge 4. *J. Comp. Aid. Mol. Des.* **2019**, *33*, 1021–1029.
- (20) Woo, H.-J.; Roux, B. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6825–6830.
- (21) Limongelli, V.; Bonomi, M.; Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci.* **2013**, *110*, 6358–6363.
- (22) Cruz, J.; Wickstrom, L.; Yang, D.; Gallicchio, E.; Deng, N. Combining Alchemical Transformation with a Physical Pathway to Accelerate Absolute Binding Free Energy

- Calculations of Charged Ligands to Enclosed Binding Sites. *J. Chem. Theory Comput.* **2020**, *16*, 2803–2813.
- (23) Kilburg, D.; Gallicchio, E. Assessment of a Single Decoupling Alchemical Approach for the Calculation of the Absolute Binding Free Energies of Protein-Peptide Complexes. *Frontiers in Molecular Biosciences* **2018**, *5*, 22.
- (24) Pal, R. K.; Gallicchio, E. Perturbation potentials to overcome order/disorder transitions in alchemical binding free energy calculations. *J. Chem. Phys.* **2019**, *151*, 124116.
- (25) others,, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comp. Bio.* **2017**, *13*, e1005659.
- (26) Khuttan, S.; Azimi, S.; Wu, J. Z.; Gallicchio, E. Alchemical Transformations for Concerted Hydration Free Energy Estimation with Explicit Solvation. *J. Chem. Phys* **2021**, In press.
- (27) Rizzi, A.; Jensen, T.; Slochower, D. R.; Aldeghi, M.; Gapsys, V.; Ntekoimes, D.; Bosio, S.; Papadourakis, M.; Henriksen, N. M.; De Groot, B. L.; Cournia, Z.; Dickson, A.; Michel, J.; Gilson, M. K.; Shirts, M. R.; Mobley, D. L.; Chodera, J. D. The SAMPL6 SAMPLing challenge: Assessing the reliability and efficiency of binding free energy calculations. *J. Comp. Aid. Mol. Des.* **2020**, 1–33.
- (28) Gallicchio, E.; Xia, J.; Flynn, W. F.; Zhang, B.; Samlalsingh, S.; Menten, A.; Levy, R. M. Asynchronous replica exchange software for grid and heterogeneous computing. *Computer Physics Communications* **2015**, *196*, 236–246.
- (29) Kilburg, D.; Gallicchio, E. Analytical Model of the Free Energy of Alchemical Molecular Binding. *J. Chem. Theory Comput.* **2018**, *14*, 6183–6196.
- (30) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of binless multi-state free

- energy estimation with applications to protein-ligand binding. *J. Chem. Phys.* **2012**, *136*, 144102.
- (31) Yang, W.; Bitetti-Putzer, R.; Karplus, M. Free energy simulations: use of the reverse cumulative averaging to determine the equilibrated region and the time required for convergence. *J. Chem. Phys.* **2003**, *120*, 2618–2628.
- (32) Chodera, J. D. A simple method for automated equilibration detection in molecular simulations. *J. Chem. Theory Comput.* **2016**, *12*, 1799–1805.
- (33) Murkli, S.; McNeill, J. N.; Isaacs, L. Cucurbit[8]uril guest complexes: blinded dataset for the SAMPL6 challenge. *Supramolecular Chemistry* **2019**, *31*, 150–158.
- (34) Sullivan, M. R.; Yao, W.; Gibb, B. C. The thermodynamics of guest complexation to octa-acid and tetra-endo-methyl octa-acid: reference data for the sixth statistical assessment of modeling of proteins and ligands (SAMPL6). *Supramolecular chemistry* **2019**, *31*, 184–189.
- (35) Andricioaei, I.; Straub, J. E.; Voter, A. F. Smart darting monte carlo. *J. Chem. Phys.* **2001**, *114*, 6994–7000.
- (36) Severance, D. L.; Essex, J. W.; Jorgensen, W. L. Generalized alteration of structure and parameters: A new method for free-energy perturbations in systems containing flexible degrees of freedom. *J. Comp. Chem.* **1995**, *16*, 311–327.
- (37) Henriksen, N. M.; Fenley, A. T.; Gilson, M. K. Computational calorimetry: high-precision calculation of host–guest binding thermodynamics. *J. Chem. Theory Comput.* **2015**, *11*, 4377–4394.
- (38) Panel, N.; Villa, F.; Fuentes, E. J.; Simonson, T. Accurate PDZ/peptide binding specificity with additive and polarizable free energy simulations. *Biophys. J.* **2018**, *114*, 1091–1102.

- (39) Beierlein, F. R.; Michel, J.; Essex, J. W. A simple QM/MM approach for capturing polarization effects in protein- ligand binding free energy calculations. *J. Phys. Chem. B* **2011**, *115*, 4911–4926.
- (40) Lodola, A.; De Vivo, M. *Adv. Prot. Chem. Struct. Biol.*; Elsevier, 2012; Vol. 87; pp 337–362.
- (41) Hudson, P. S.; Woodcock, H. L.; Boresch, S. Use of interaction energies in QM/MM free energy simulations. *J. Chem. Theory Comput.* **2019**, *15*, 4632–4645.
- (42) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Comm.* **2019**, *10*, 1–8.
- (43) Rufa, D. A.; Macdonald, H. E. B.; Fass, J.; Wieder, M.; Grinaway, P. B.; Roitberg, A. E.; Isayev, O.; Chodera, J. D. Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning/molecular mechanics potentials. *BioRxiv* **2020**,
- (44) Heinzelmann, G.; Chen, P.-C.; Kuyucak, S. Computation of standard binding free energies of polar and charged ligands to the glutamate receptor GluA2. *The Journal of Physical Chemistry B* **2014**, *118*, 1813–1824.
- (45) Gapsys, V.; Michielssens, S.; Peters, J. H.; de Groot, B. L.; Leonov, H. *Molecular Modeling of Proteins*; Springer, 2015; pp 173–209.
- (46) Ekimoto, T.; Yamane, T.; Ikeguchi, M. Elimination of finite-size effects on binding free energies via the warp-drive method. *Journal of chemical theory and computation* **2018**, *14*, 6544–6559.

- (47) Macchiagodena, M.; Pagliai, M.; Karrenbrock, M.; Guarnieri, G.; Iannone, F.; Procacci, P. Virtual Double-System Single-Box: A Nonequilibrium Alchemical Technique for Absolute Binding Free Energy Calculations: Application to Ligands of the SARS-CoV-2 Main Protease. *J. Chem. Theory Comput.* **2020**,
- (48) Mobley, D. L.; Gilson, M. K. Predicting binding free energies: frontiers and benchmarks. *Ann. Rev. Bioph.* **2017**, *46*, 531–558.
- (49) Velez-Vega, C.; Gilson, M. K. Overcoming dissipation in the calculation of standard binding free energies by ligand extraction. *J. Comp. Chem.* **2013**, *34*, 2360–2371.