Implicit ligand theory for relative binding free energies: II. An estimator based on control variates

Trung Hai Nguyen\textsuperscript{1,2} and David D L Minh\textsuperscript{3,4}*

1 Laboratory of Theoretical and Computational Biophysics, Ton Duc Thang University, Ho Chi Minh City, Vietnam
2 Faculty of Applied Sciences, Ton Duc Thang University, Ho Chi Minh City, Vietnam
3 Department of Chemistry, Illinois Institute of Technology, Chicago, IL 60616, United States of America
4 Author to whom any correspondence should be addressed.

E-mail: nguyentrunghai@tdtu.edu.vn and dminh@iit.edu

Keywords: binding free energy, protein-ligand, implicit ligand theory, statistical mechanics, control variates

Abstract

Implicit ligand theory describes the relationship between the noncovalent binding free energy and the binding free energy between a ligand and multiple rigid receptor conformations. We have previously shown that if the receptor conformations are sampled from or reweighed to a holo ensemble, the binding free energy relative to the ligand that defines the ensemble can be calculated. Here, we apply a variance reduction technique known as control variates to derive a new statistical estimator for the relative binding free energy. In applications to a data set of 6 reference ligands and 18 test ligands, statistically significant differences between the estimators are not observed for most systems. However, in cases where such differences are observed, the new estimator is more accurate, precise, and converges more quickly. Performance improvements are most consistent where there is a clear correlation, with a correlation coefficient greater than 0.3, between the control variate and the statistic being averaged.

1. Introduction

Noncovalent binding between small organic molecules and biological macromolecules is a ubiquitous process in biology and is critical to the mechanism of most drugs. Thus, there has been significant interest in developing computational methods to predict binding free energies, which quantify the strength of these interactions \cite{1–7}. A key application of these methods is rational drug design \cite{5, 7–9}.

Alchemical pathway methods \cite{10} are a class of theoretically rigorous but computationally expensive binding free energy methods. These methods involve sampling conformations of receptor-ligand complexes from a series of thermodynamic states along a possibly nonphysical pathway. They can be used to obtain either absolute binding free energies between a receptor and ligand ($\Delta G_{RL}^{\circ}$, as defined in equation (1)) or relative binding free energies between a receptor and two different ligands ($\Delta \Delta G_{RL}^{\circ}$, as defined in equation (2)). In $\Delta G_{RL}^{\circ}$ calculations, the pathway may involve decoupling or physically separating the ligand from the receptor \cite{11–14}. In $\Delta \Delta G_{RL}^{\circ}$ calculations, the ligand is transformed from one molecule to another \cite{15, 16}. Alchemical pathway methods have accurately predicted binding free energies for many systems including protein-ligand \cite{9, 17–22} and protein-protein complexes \cite{23, 24}. However, computational expense continues to restrict more widespread application of the methods.

Since 2012, we have been developing binding free energy methods based on implicit ligand theory (ILT) \cite{25}—a statistical mechanics framework for calculating absolute or relative binding free energies using a presampled set of rigid receptor conformations—that can quickly calculate binding free energies for a large set of ligands binding to the same receptor \cite{26–28}. Computing $\Delta G_{RL}^{\circ}$ involves drawing from or reweighing receptor conformations to the apo ensemble—the receptor alone, in the absence of ligand \cite{25}. In a $\Delta \Delta G_{RL}^{\circ}$ calculation, receptor configurations are drawn from a holo ensemble—where the receptor is bound to a ligand \cite{29}. After receptor configurations are selected, the next step of an ILT-based free energy calculation is to compute the
binding potential of mean force (BPMF)—the $\Delta G_{RL}^\circ$ between a flexible ligand and a rigid receptor—for the receptor configurations. The standard absolute $\Delta G_{RL}^\circ$, computed using the ‘apo estimator’, is an exponential average of the BPMF over the apo ensemble. $\Delta \Delta G_{RL}^\circ$ of a ligand relative to the reference ligand defining the holo ensemble, computed with the ‘holo estimator’ (equation (6)), is an exponential average of the BPMF over the reference holo ensemble.

ILT has several advantages over alternative binding free energy methods based on a flexible receptor. The main advantage of ILT is that once conformations of the receptor are thoroughly sampled, they can be used to calculate the binding free energies of many ligands. In contrast, in alchemical pathway calculations with a flexible receptor, receptor conformational sampling needs to be performed for every receptor-ligand complex. Another advantage of ILT is that BPMF calculations are much faster and more scalable than binding free energies to flexible receptors [28]. Protein-ligand interaction energies may be pre-computed on a grid and interpolated, reducing the number of nonbonded force calculations and essentially eliminating the relationship between receptor size and computational speed. Finally, it is worth noting that unlike most methods for $\Delta \Delta G_{RL}^\circ$ estimation, the holo estimator does not require one ligand to be transformed into another. Therefore it can be employed even for cases where the target ligand significantly differs from the reference one.

Regarding receptor conformational sampling, the optimal choice of apo or holo ensemble depends on the extent to which ligand binding induces a consistent conformational change in the receptor. Both the apo and holo estimators are exponential averages that are dominated by the lowest BPMFs. The receptor configurations with the lowest BPMFs are those that are prevalent in the holo ensemble of interest. In contrast, BPMFs for receptor configurations that poorly accommodate a ligand may be approximated as infinite without a significant effect on the binding free energy estimate. Thus, converged $\Delta G_{RL}^\circ$ estimates with ILT require sampling receptor configurations that are prevalent in the holo ensemble for the ligand of interest. If ligand binding induces a significant conformational change, then simulating the apo ensemble is unlikely to yield such conformations. In this case, relevant configurations may be obtained by simulating one or more holo ensembles with different ligands [26]. In principle, enhanced and biased sampling techniques may also be used for receptor conformational sampling.

Here, we describe a variation of the holo estimator that is based on exploiting the BPMFs between a ligand and its own holo ensemble. These BPMFs may be used to estimate the relative binding free energy between a ligand and itself, the ‘self’ relative binding free energy, $\Delta \Delta G_{RL}^\circ$. By definition, $\Delta \Delta G_{RL}^\circ$ is zero. However, for a finite set of receptor configurations, $\Delta G_{RL}^\circ$ may be estimated as nonzero. In previous work [29], we corrected $\Delta G_{RL}^\circ$ estimates for all ligands by subtracting the estimate of $\Delta G_{RL}^\circ$. This correction ensures that the estimated $\Delta G_{RL}^\circ$ is always zero, regardless of the sample size. Although subtracting this term helped reduce the error of other $\Delta G_{RL}^\circ$ estimates, this strategy may not be the most statistically optimal approach to use the information contained in the BPMFs.

To improve the holo estimator, we apply a statistical variance reduction technique known as control variates. In the control variates technique, the bias in an estimate of a known expectation value is used to correct the bias in the estimate of an unknown expectation value. The expectation values are of quantities that can be calculated from the same samples. The method works best when the quantities are highly correlated and therefore bias in one estimate is related to bias in the other.

The structure of this paper is as follows: first, we introduce notation and review the holo estimator. We then develop a variation of the estimator based on control variates. We then apply the newly optimized estimator to $\Delta G_{RL}^\circ$ calculations for T4 lysozyme and 24 ligands. Finally, we compare the accuracy of the unoptimized and optimized estimators with the results from alchemical pathway calculations.

2. Theory

The absolute standard binding free energy of a ligand $L$ noncovalently bound to receptor $R$ to form a complex $RL$ is given by,

$$\Delta G_{RL}^\circ = -k_B T \ln \left( \frac{C_{RL} C^0}{C_R C_L} \right)$$  \hspace{1cm} (1)

where $k_B$ is Boltzmann’s constant and $T$ is the temperature in Kelvin. $C_R$, $C_L$ and $C_{RL}$ are equilibrium concentrations of the receptor, ligand and complex, respectively. $C^0$ is the standard concentration, 1 M.

In this work, we focus on $\Delta \Delta G_{RL}^\circ$, which is defined as the difference in binding free energies between a target ligand $L$ and a reference ligand $L_o$,

$$\Delta \Delta G_{RL}^\circ = \Delta G_{RL}^\circ - \Delta G_{RL_o}^\circ.$$  \hspace{1cm} (2)

The definition of $\Delta G_{RL}^\circ$ is the same as $\Delta G_{RL}^\circ$ except that $L_o$ replaces $L$. 
2.1. ILT estimator for relative binding free energies

Previously, we reported that \( \Delta \Delta G_{RL}^o \) can be written as an expectation value \([29]\),

\[
\exp \left[ - \frac{\Delta \Delta G_{RL}^o}{k_B T} \right] = \frac{\int \rho_{RL}(r_{RL}) e^{-\beta [\hat{B}(r_{RL}) - \Psi(r_{RL})]} dr_{RL}}{\int \rho_{RL}(r_{RL}) dr_{RL}} \\
\equiv \langle \exp \{ -\beta [B(r) - \Psi(r_{RL})] \} \rangle_{RL},
\]

(3)

where \( k_B \) is Boltzmann’s constant and \( T \) is the temperature. \( \rho_{RL}(r_{RL}) = I(\xi_o) J(\xi_o) e^{-\beta H(r_{RL})} \) is the unnormalized probability density of the receptor-ligand complex in the bound state. It includes: \( I(\xi_o) \), a function based on the ligand external coordinates \( \xi \) that specifies whether the ligand is in the binding site; \( J(\xi_o) \), the Jacobian for transforming the ligand external coordinates from Cartesian into the coordinate system used for \( \xi \); and \( U(r_{RL}) \), the potential energy of species \( X \in \{ R, L_o, RL_o \} \) in implicit solvent. The observable of the expectation \( \langle ... \rangle_{RL} \) includes \( B(r) \), the BPMF—the binding free energy of the flexible target ligand \( L \) to a rigid receptor conformation. It also includes \( \Psi(r_{RL}) = U(r_{RL}) - U(r_R) - U(r_L) \), the interaction energy between the reference ligand \( L_o \) and the receptor \( R \) when they form a bound complex \( r_{RL} \).

If configurations of the complex are drawn from a different statistical distribution than the bound state, then importance sampling may be used to estimate \( \Delta \Delta G_{RL}^o \),

\[
\exp \left[ - \frac{\Delta \Delta G_{RL}^o}{k_B T} \right] = \frac{\langle w(r_{RL}) \exp \{ -\beta [B(r) - \Psi(r_{RL})] \} \rangle_0}{\langle w(r_{RL}) \rangle_0},
\]

(5)

where \( w(r_{RL}) = \rho_{RL}(r_{RL})/\rho_0(r_{RL}) \) is the ratio of the statistical weight in the holo state, \( \rho_{RL}(r_{RL}) \), over the statistical weight in the original distribution that configurations are sampled from, \( \rho_0(r_{RL}) \).

Using a sample mean to estimate the expectation value in equation (5) leads to,

\[
\Delta \Delta \hat{G}_{RL}^o = -k_B T \ln \left\{ \frac{\sum r_{RL} w(r_{RL}) e^{-\beta [\hat{B}(r) - \Psi(r_{RL})]}}{\sum r_{RL} w(r_{RL})} \right\},
\]

(6)

where the summation is over conformations of the receptor. \( \hat{B}(r) \) is an estimate of the BPMF.

When \( \hat{B}(r) \) is the BPMF of the reference ligand itself, we obtain an estimate of the ‘self’ relative binding free energy,

\[
\Delta \Delta \hat{G}_{RL}^o = -k_B T \ln \left\{ \frac{\sum r_{RL} w(r_{RL}) e^{-\beta [\hat{B}(r) - \Psi(r_{RL})]}}{\sum r_{RL} w(r_{RL})} \right\},
\]

(7)

In the limit of infinite sample size, \( \Delta \Delta \hat{G}_{RL}^o \) should converge to zero. However, it can be nonzero for a finite number of samples. In our previous study \([29]\), we subtracted \( \Delta \Delta \hat{G}_{RL}^o \) from \( \Delta \Delta G_{RL}^o \) to ensure that the corrected \( \Delta \Delta G_{RL}^o \) estimate is always zero, regardless of the sample size:

\[
\Delta \Delta \hat{G}_{RL}^o = \Delta \Delta G_{RL}^o - \Delta \Delta \hat{G}_{RL}^o = \Delta \Delta G_{RL}^o - \Delta \Delta \hat{G}_{RL}^o = -k_B T \ln \left\{ \frac{\sum r_{RL} w(r_{RL}) e^{-\beta [\hat{B}(r) - \Psi(r_{RL})]}}{\sum r_{RL} w(r_{RL})} \right\},
\]

(8)

We will refer to this expression as estimator A. Although estimator A ensures that the ‘self’ relative binding free energy estimate is zero, it may not be statistically optimal.

2.2. ILT estimator based on control variates

To apply the control variates technique, we need a property with an exactly known expectation value. In the special case that the ligand is the reference ligand, then the left hand side of equation (5) is unity. Rearrangement and combination of the expectation values yields \( \langle g(r_{RL}) \rangle_0 = 0 \), where,

\[
g(r_{RL}) = w(r_{RL}) \left( e^{-\beta [\hat{B}(r) - \Psi(r_{RL})]} - 1 \right).
\]

(10)

For notational convenience, we also define a function based on the key observable in equation (5),

\[
h(r_{RL}) = w(r_{RL}) e^{-\beta [\hat{B}(r) - \Psi(r_{RL})]}. \]

(11)

In the control variates technique, \( \langle h(r_{RL}) \rangle \) is estimated by the sample mean of a new variable, \( m(r_{RL}) \), which is given by,
between T4 lysozyme and 24 small organic molecules was previously calculated [26] using YANK, a software package developed by the Chodera group [12]. In these calculations, the AMBER ff14 force field [30] was used for T4 lysozyme and the generalized AMBER force field [31] with Bondi radii [32] and AM1 BCC partial charges [33, 34] for ligands. Solvent was treated with alchemical pathway method. The absolute binding free energy \( \Delta G_{RL} \) between T4 lysozyme and the generalized AMBER ff14 force field is given by equation (13).

\[
m(t_{RL}) = h(t_{RL}) + C(g(t_{RL})) = h(t_{RL}) - Cg(t_{RL}),
\]

(12)

where \( C \) is an arbitrary constant. In the standard control variates method, \( C \) is optimized to minimized the variance of the sample mean of \( m(t_{RL}) \). In this paper, we select \( C \) to minimize the variance of a new estimator for relative binding free energy, which is given by,

\[
\Delta \Delta G_{RL}^{c,B} = -k_B T \ln \frac{\bar{m}}{\bar{w}}.
\]

(13)

The bar over the variable name denotes the sample mean, such that \( \bar{m} = \frac{1}{N} \sum m(t_{RL}) \) and \( \bar{w} = \frac{1}{N} \sum w(t_{RL}) \). \( N \) is the number of sampled receptor configurations.

To optimize \( C \), we use error propagation formulae to obtain the variance of \( \Delta \Delta G_{RL}^{c,B} \) with respect to \( C \), set its derivative to zero, and solve for \( C \). The variance of the estimator is given by,

\[
\mathrm{Var}[\Delta \Delta G_{RL}^{c,B}] = (k_B T)^2 \left( \frac{\bar{m}^2}{N} \right)^2 \mathrm{Var} \left[ \frac{\bar{m}}{\bar{w}} \right]
\]

\[
= (k_B T)^2 \left[ \frac{\mathrm{Var}[\bar{m}]}{\bar{m}^2} + \frac{\mathrm{Var}[\bar{w}]}{\bar{w}^2} - \frac{2}{(\bar{w})(\bar{m})} \mathrm{Cov}[\bar{m}, \bar{w}] \right]
\]

\[
= \frac{(k_B T)^2}{N} \left[ \frac{\mathrm{Var}[m]}{m^2} + \frac{\mathrm{Var}[w]}{w^2} - \frac{2}{(w)(m)} \mathrm{Cov}[m, w] \right].
\]

(14)

Note that only \( m \) depends on \( C \). Taking derivative of \( \mathrm{Var}[\Delta \Delta G_{RL}^{c,B}] \) with respect to \( C \) yields,

\[
\frac{d}{dC} \mathrm{Var}[\Delta \Delta G_{RL}^{c,B}] = \frac{(k_B T)^2}{N} \left[ \frac{d}{dC} \left[ \frac{\mathrm{Var}[m]}{m^2} \right] - \frac{2}{m} \frac{d}{dC} \left[ \frac{\mathrm{Cov}[m, w]}{w} \right] \right] = 0
\]

(15)

The derivative can be set to zero, resulting in,

\[
\frac{d}{dC} \left[ \frac{\mathrm{Var}[m]}{m^2} \right] - \frac{2}{m} \frac{d}{dC} \left[ \frac{\mathrm{Cov}[m, w]}{w} \right] = 0
\]

(16)

Substituting derivatives of the ratios into equation (16) and rearranging gives,

\[
(\bar{w})(\bar{m}) \frac{d}{dC} \frac{\mathrm{Var}[m]}{m^2} - 2\bar{m}^2 \frac{d}{dC} \frac{\mathrm{Cov}[m, w]}{w} + 2(\bar{m}\bar{w}) \frac{d}{dC} \frac{\mathrm{Var}[m]}{m} = 0,
\]

(17)

where

\[
m = h - Cg, \quad \bar{m} = \bar{h} - C\bar{g}, \quad \frac{d}{dC} \bar{m} = -\bar{g},
\]

\[
\frac{d}{dC} \mathrm{Var}[m] = \frac{d}{dC} [\mathrm{Var}[h] + C^2 \mathrm{Var}[g] - 2C \mathrm{Cov}[h, g]],
\]

\[
\frac{d}{dC} \mathrm{Cov}[m, w] = \frac{d}{dC} [\mathrm{Cov}[h, w] - C \mathrm{Cov}[g, w]],
\]

\[
\frac{d}{dC} \mathrm{Cov}[m, w] = -\mathrm{Cov}[g, w].
\]

Substituting these identities into equation (17), rearranging, and solving for \( C \) gives,

\[
C = \frac{(\bar{w})(\bar{h}) \mathrm{Cov}[h, g] - (\bar{w})(\bar{g}) \mathrm{Var}[h] - \bar{h} \bar{g} \mathrm{Cov}[g, w] + (\bar{h})(\bar{g}) \mathrm{Cov}[h, w]}{(\bar{w})(\bar{h}) \mathrm{Var}[g] - (\bar{w})(\bar{g}) \mathrm{Var}[h, g] - (\bar{h})(\bar{g}) \mathrm{Cov}[h, w] + \bar{h} \bar{g} \mathrm{Cov}[h, w]}
\]

(18)

We will refer to equation (13) with \( C \) given by equation (18) as estimator B. Estimator B is the key theoretical result of the paper. In the remainder of the paper, we will compare the accuracy and precision of estimator A (equation (9)) and estimator B (equation (13)).

3. Methods

All of the quantities needed to calculate binding free energies using estimators A (equation (9)) and B (equation (13)) and to assess their quality were obtained in previous studies. For the reader’s convenience, we describe some key details in this section.

To benchmark the accuracy of the two estimators, we used absolute binding free energies obtained via an alchemical pathway method. The absolute binding free energy \( \Delta G_{RL} \) between T4 lysozyme and 24 small organic molecules was previously calculated [26] using YANK, a software package developed by the Chodera group [12]. In these calculations, the AMBER ff14 force field [30] was used for T4 lysozyme and the generalized AMBER force field [31] with Bondi radii [32] and AM1 BCC partial charges [33, 34] for ligands. Solvent was treated with
the OBC2 generalized Born/surface area implicit solvent model [35]. Calculations were repeated 3 times for each system. Standard deviations were less than 0.5 kcal/mol for 22 systems and less than 1 kcal/mol for all 24 systems [26].

We considered one reference set and two test sets: test set I to benchmark consistency with YANK and test set II to benchmark against experimental results. The reference set consisted of methylpyrrole, benzene, p-xylene, phenol, n-hexylbenzene and DL-camphor. Test set I consisted of 18 remaining ligands from YANK calculations. Test set II included 18 ligands having experimental binding free energies reported in references [36–38]. As seven ligands were part of both test sets, we considered a total of 29 ligands.

ILT-based binding free energies were also computed using quantities obtained in previous studies. In previous work, 96 receptor snapshots were selected from every alchemical state of YANK calculations with the reference ligands. The BPMF between each ligand and the selected receptor configurations was calculated [26] using our software package AlGDock [28, 39]. The same force field and implicit solvent model were used in the YANK and AlGDock calculations. Interaction energies \( Y_rRLo \) were calculated using the OpenMM 6.3.1 library [40, 41]. Since receptor snapshots selected for AlGDock calculations were drawn from all alchemical states, not just the holo state, we needed to reweigh them to the holo state. To do this, we used the multistate Bennett Acceptance Ratio (MBAR) [42] to calculate the weights \( w(r_{RL}) \) in the holo ensemble.

Estimator A (equation (9)) and B (equation (13)) were used to calculate relative binding free energies. Relative binding free energies were converted to absolute binding free energies by adding them to the YANK binding free energy of the appropriate reference ligand.

The accuracy of the two estimators was compared via several metrics. The root mean square error (RMSE) and Pearson’s R were computed with respect to YANK results. We also compared the estimators using the difference in absolute deviation from YANK,

\[
d = |\Delta \tilde{G}^{B}_{RL} - \Delta \Delta \tilde{G}^{A}_{RL} - \Delta \Delta \tilde{G}^{A,Y}_{RL}| - |\Delta \Delta \tilde{G}^{B}_{RL} - \Delta \Delta \tilde{G}^{A,Y}_{RL}|,
\]  

Figure 1. Binding free energies for ligands in test set I relative to all reference ligands estimated by YANK (x-axis) and AlGDock (y-axis) using estimator A (left column) or estimator B (right column). Error bars denote the standard deviation from three independent YANK calculations (x-axis) or from bootstrapping BPMFs (y-axis), with the range of error bars representing a single standard deviation. A least-squares linear regression is shown as a dashed line.

Table 1. Comparing the accuracy of estimators A and B with respect to YANK free energies. The RMSE is in kcal/mol.

| Reference ligand | Estimator A | Estimator B |
|------------------|-------------|-------------|
|                  | Pearson’s R | RMSE        | Pearson’s R | RMSE        |
| methylpyrrole    | 0.94(0.03)  | 1.20(0.31)  | 0.94(0.03)  | 1.59(0.32)  |
| benzene          | 0.89(0.06)  | 1.48(0.46)  | 0.83(0.08)  | 1.68(0.45)  |
| p-xylene         | 0.83(0.07)  | 1.72(0.36)  | 0.83(0.05)  | 1.77(0.30)  |
| phenol           | 0.82(0.07)  | 1.94(0.33)  | 0.84(0.07)  | 1.73(0.37)  |
| DL-camphor       | 0.91(0.04)  | 3.98(0.26)  | 0.92(0.04)  | 3.53(0.26)  |
| n-hexylbenzene   | 0.89(0.05)  | 7.65(0.33)  | 0.86(0.07)  | 5.81(0.33)  |
| overall          | 0.65(0.05)  | 3.76(0.29)  | 0.72(0.05)  | 3.11(0.21)  |

The accuracy of the two estimators was compared via several metrics. The root mean square error (RMSE) and Pearson’s R were computed with respect to YANK results. We also compared the estimators using the difference in absolute deviation from YANK.

The accuracy of the two estimators was compared via several metrics. The root mean square error (RMSE) and Pearson’s R were computed with respect to YANK results. We also compared the estimators using the difference in absolute deviation from YANK.

The accuracy of the two estimators was compared via several metrics. The root mean square error (RMSE) and Pearson’s R were computed with respect to YANK results. We also compared the estimators using the difference in absolute deviation from YANK.

The accuracy of the two estimators was compared via several metrics. The root mean square error (RMSE) and Pearson’s R were computed with respect to YANK results. We also compared the estimators using the difference in absolute deviation from YANK.
where $\Delta \Delta G_{RL}^{A}$ and $\Delta \Delta G_{RL}^{B}$ are estimator A (equation (9)) and B (equation (13)), respectively. $\Delta \Delta G_{RL}^{x,Y}$ is the $\Delta \Delta G_{RL}$ obtained from YANK. A negative value of $d$ indicates that estimator B reduces the absolute deviation from YANK compared to estimator A and a positive value means the opposite.

Bootstrapping was used to estimate statistical errors. For relative binding free energies, random sets of samples were drawn with replacement from the set of 96 receptor snapshots. The standard deviation of free energy estimates from 100 such sets was reported as the uncertainty. To estimate error bars for RMSE and Pearson’s R values, the set of 18 test ligands was resampled with replacement 100 times.

### 4. Results

#### 4.1. Estimator B improves consistency with YANK

Based on binding free energies of all test ligands in test set I relative to all reference ligands, estimator B is more consistent with YANK than estimator A (figure 1 and table 1). The RMSE of free energies with respect to YANK
is lower with estimator B (3.11 ± 0.21 kcal/mol) than with estimator A (3.76 ± 0.29 kcal/mol). The difference in RMSE is larger than the estimated error, suggesting that the improvement is statistically significant. While the Pearson’s R of estimator B (0.72 ± 0.05) is also slightly higher than that of estimator A (0.65 ± 0.05), the improvement is not large compared to the estimated error.

Although estimator B has a better performance overall, differences in quality metrics are only statistically significant for binding free energies relative to two reference ligands (figures 2 and 3 and table 1). For three out of six reference ligands (methylpyrrole, benzene and p-xylene) estimator B yields slightly higher RMSEs compared to estimator A, but the differences are smaller than the estimated error. For another three systems (phenol, DL-camphor and n-hexylbenzene) estimator B is more consistent with YANK than A. For phenol, the improvement is slight and is unlikely to be statistically significant. However, for DL-camphor and n-hexylbenzene, which are the two most difficult cases yielding the largest RMSEs, estimator B shows a significant reduction in RMSEs. For n-hexylbenzene, the RMSE is dramatically reduced from 7.65 ± 0.33 kcal/mol to 5.82 ± 0.35 kcal/mol. On the other hand, differences between the RMSE of the two estimators are smaller than estimated error (table 1).

4.2. Estimator B reduces errors with respect to experiment

Overall, estimator B reproduces experimental results more accurately than estimator A (figure 4 and table 2). The RMSE with respect to experiment of estimator B (2.77 ± 0.21 kcal/mol) is lower than the RMSE of
estimator A (3.53 ± 0.28 kcal/mol). The difference in RMSE is statistically significant because it is much larger than the estimated errors. The Pearson’s R of estimator B (0.32 ± 0.07) is slightly higher than that of estimator A (0.25 ± 0.08), but the difference is small compared to the estimated errors. When considering each reference ligand separately (figures 5, 6 and table 2), the RMSE of estimator B with respect to experiment is significantly lower than the RMSE of estimator A in three of six cases. In particular, it significantly improves the most difficult case, based on the reference ligand n-hexylbenzene. For two reference ligands, methylpyrrole and p-xylene, estimator A has a somewhat lower RMSE than estimator B. The two estimators perform equally well in the case of the reference ligand benzene. Due to large estimated errors, differences in the Pearson’s R between the two estimators are not statistically significant.

4.3. Estimator B converges faster for the most difficult case
For all reference ligands except for n-hexylbenzene, the rate at which the binding free energy estimators converge is comparable (figure 7). Convergence was assessed by the rate at which the RMSE with respect to YANK is reduced as the number of receptor snapshots is increased. Convergence behavior is very similar for five out six reference ligands: methylpyrrole, benzene, p-xylene, phenol and DL-camphor; the curves look rather similar and level off starting at almost the same number of receptor snapshots. For p-xylene and DL-camphor, estimator B appears to converge more closely than estimator A, but differences are well within error bars. On the other hand, estimator B shows significantly faster convergence in the RMSE for the hardest system, n-hexylbenzene. With estimator A, increasing the number of snapshots actually increases the RMSE with respect to YANK.

4.4. Overall, estimator B is more precise than A
In addition to being more accurate and converging more quickly than estimator A, estimator B is also more precise. In the majority of systems, 63%, the error estimated for estimator B is smaller than for estimator A (figure 8). The mean and median values of error estimates for estimator A are 0.89 kcal/mol and 0.81 kcal/mol respectively; for estimator B, they are 0.74 kcal/mol and 0.66 kcal/mol, respectively.

4.5. Performance improvement of estimator B over A is related to correlation between $h(r_{RL})$ and $g(r_{RL})$
Following expectations for the control variates technique, estimator B exhibits the most consistent improvement over estimator A when the statistic $h(r_{RL})$ is most correlated with the control variate $g(r_{RL})$ (figure 9 and table 3). Here, the metric that we consider is $d$ (equation (19)), the difference in the absolute deviation from YANK. Note that negative $d$ indicates that estimator B has a smaller absolute deviation from YANK than estimator A. When the absolute correlation between $h(r_{RL})$ and $g(r_{RL})$ is weak, such that $|\text{Corr}(h, g)| < 0.3$, $d$ are almost equally spread on both positive and negative sides; estimator B improves over A in 58% of the binding free energy calculations. On other other hand, when the correlation is stronger, $|\text{Corr}(h, g)| > 0.3$, $d$ are negative in 79% of calculations. Averaging over all systems, the mean of $d$ is -0.3 kcal/mol, consistent with other metrics demonstrating that estimator B is an improvement over estimator A.
5. Conclusions

Using the control variates technique, we have developed a new statistical estimator for using ILT to calculate binding free energies relative to a reference ligand that defines the holo ensemble. We have applied the new estimator to previously-collected simulation data for the binding of 6 reference ligands and 18 test ligands to T4 lysozyme. In most cases, differences in the performance of the new and old estimator are not statistically significant. In the cases where statistically significant differences are observed, the new estimator outperforms the old one; it is more accurate, precise, and converges more quickly. As anticipated for the control variates technique, performance improvements are greatest in cases where there is a strong correlation between the statistic \( h(\text{FLL}) \) and control variate \( g(\text{FLL}) \).

We recommend using the estimator over the previous one. Python code for the new estimator is included in the SI and on github at https://github.com/nguyentrunghai/RelBfe_control_variates. Data analyzed in this study has been deposited in the github repository.

Figure 5. Binding free energies measured by experiment (x-axis) and estimated by AlGDock (y-axis) using estimator A (left column) or estimator B (right column) for ligands in test set II. Error bars denote the standard deviation from bootstrapping BPMFs (y-axis), with the range of error bars representing a single standard deviation. A least-squares linear regression is shown as a dashed line. Each row corresponds to a reference ligand. Results for other three reference ligands are shown in figure 6.
Figure 6. Binding free energies continued from figure 5.

Table 2. Comparing the accuracy of estimators A and B with respect to experimental free energies. The RMSE is in kcal/mol.

| Reference ligand   | Estimator A | Estimator B |
|-------------------|-------------|-------------|
|                   | Pearson’s R | RMSE        | Pearson’s R | RMSE        |
| methylpyrrole     | 0.50(0.18)  | 1.04(0.12)  | 0.52(0.18)  | 1.70(0.20)  |
| benzene           | 0.41(0.22)  | 1.20(0.19)  | 0.43(0.19)  | 1.03(0.19)  |
| p-xylene          | 0.75(0.09)  | 1.34(0.17)  | 0.72(0.10)  | 1.78(0.28)  |
| phenol            | 0.22(0.25)  | 2.01(0.22)  | 0.42(0.20)  | 1.48(0.17)  |
| DL-camphor        | 0.42(0.11)  | 3.78(0.33)  | 0.35(0.21)  | 2.94(0.28)  |
| n-hexylbenzene    | 0.63(0.18)  | 7.22(0.37)  | 0.57(0.17)  | 5.29(0.42)  |
| overall           | 0.25(0.08)  | 3.53(0.28)  | 0.32(0.07)  | 2.77(0.21)  |
Figure 7. Convergence of binding free energies. RMSE with respect to YANK binding free energies for estimator A (blue) and B (red) for ligands in test set I. Error bars denote the standard deviation from bootstrapping. Different panels correspond to different reference ligands.

Figure 8. Bootstrap errors of estimator B versus A.
Figure 9. Difference in absolute deviation from YANK, \( d \), versus correlation between \( g(r_{50}) \) and \( h(r_{50}) \).

Table 3. Statistics related to difference in absolute deviation from YANK, \( d \), and absolute correlation coefficient corr\((g, h)\). \( d \) is in kcal/mol.

| statistic                     | \( |\text{Corr}(g, h)| < 0.3 \) | \( |\text{Corr}(g, h)| \geq 0.3 \) |
|-------------------------------|---------------------------------|----------------------------------|
| percentage of calculations    | \( d < 0 \) 58\%                | \( d \geq 0 \) 42\%               |
|                              | \( d < 0 \) 79\%                | \( d \geq 0 \) 21\%               |
| mean of \( d \) within subset | \( d < 0 \) -0.84                | \( d \geq 0 \) 0.58               |
|                              | \( d < 0 \) -0.64                | \( d \geq 0 \) 0.38               |

Acknowledgments

This research was supported in part by the National Institutes of Health (R01GM127712).

ORCID iDs

Trung Hai Nguyen https://orcid.org/0000-0003-1848-3963
David D L Minh https://orcid.org/0000-0002-4802-2618

References

[1] Jorgensen W L 2004 Science 303 1313–8
[2] Gilson M K and Zhou H X 2007 Annu. Rev. Biophys. Biomol. Struct. 36 21–42
[3] Michel J and Essex J W 2010 J. Comput.-Aided Mol. Des. 24 639–58
[4] Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K and Pande VS 2011 Curr. Opin. Struct. Biol. 21 150–60
[5] Mobley DL and Klimov PV 2012 J. Chem. Phys. 137 230901
[6] Abel R, Wang L, Mobley DL and Friesner RA 2017 Curr. Top. Med. Chem. 17 2577–85
[7] Mobley DL and Gilson MK 2017 Annu. Rev. Biophys. 46 531–58 PMID: 28399632
[8] Christ CD and Fox T 2014 J. Chem. Inf. Model. 54 108–20 PMID: 24256082
[9] Wang L et al 2015 JACS 137 2695–703 PMID:25625324
[10] Gilson MK, Given JA, Bush BL and McCammon JA 1997 Biophys. J. 72 1047–69
[11] Deng Y and Roux B 2009 J. Phys. Chem. B 113 2234–46
[12] Wang K, Chodera JD, Yang Y and Shirts MR 2013 J. Comput.-Aided Mol. Des. 27 989–1007
[13] Heinzelmann G, Henriksen NM and Gilson MK 2017 J. Chem. Theory Comput. 13 3260–75
[14] Bekker GJ, Kamyi N, Araki M, Fukuida I, Okuno Y and Nakamura H 2017 J. Chem. Theory Comput. 13 2389–99
[15] Wong CF and McCammon JA 1986 JACS 108 3830–2
[16] Bash P, Singh U, Langridge R and Kollman P 1987 Science 235 574–6
[17] Michel J and Essex J W 2008 J. Med. Chem. 51 6654–64
[18] Boyce SE, Mobley DL, Rocklin GJ, Graves AP, Dill KA and Shoichet BK 2009 J. Mol. Biol. 394 747–63
[19] Ge X and Roux B 2010 J. Phys. Chem. B 114 9525–39
[20] Wang L, Berne BJ and Friesner RA 2012 Proc. Natl. Acad. Sci. USA 109 1937–42
[21] Zhu S, Travis SM and Elcock AH 2013 J. Chem. Theory Comput. 9 3151–64
[22] Aldeghi M, Heitz M, Bodkin MJ, Knapp S and Biggin PC 2016 Chem. Sci. 7 207–18
[23] Gumbart JC, Roux B and Chipot C 2013 J. Chem. Theory Comput. 9 3789–98
[24] Rodriguez RA, Yu L and Chen LY 2015 J. Chem. Theory Comput. 11 4427–38 PMID: 26366131
[25] Minh DD L 2012 J. Chem. Phys. 137 104106
[26] Xie B, Nguyen T H and Minh DD L 2017 J. Chem. Theory Comput. 13 2930–44
[27] Nguyen TH, Zhou HX and Minh DD L 2018 J. Comput. Chem. 39 621–36
[28] Minh DD L 2020 J. Comput. Chem. 41 715–30
[29] Nguyen T H and Minh D D L. 2018 J. Chem. Phys. 148 104114
[30] Ponder J W and Case D A. 2003 Adv. Protein Chem. 66 27–85
[31] Wang J, Wang W, Kollman P A and Case D A. 2006 J. Mol. Graph. Model. 25 247–60
[32] Bondi A. 1964 J. Phys. Chem. 68 441–51
[33] Jakalian A, Bush B L, Jack D B and Bayly C I. 2000 J. Comput. Chem. 21 132–46
[34] Jakalian A, Jack D B and Bayly C I. 2002 J. Comput. Chem. 23 1623–41
[35] Onufriev A, Bashford D and Case D A. 2004 Proteins Struct. Funct. Bioinf. 55 383–94
[36] Morton A, Baase W A and Matthews B W. 1995 Biochemistry 34 8564–75 PMID: 7 612 598
[37] Mobley D L, Graves A P, Chodera J D, McReynolds A C, Shoichet B K and Dill K A. 2007 J. Mol. Biol. 371 1118–34
[38] Liu L, Baase W A and Matthews B W. 2009 J. Mol. Biol. 385 595–605
[39] Minh D D L. 2020 Journal of Computational Chemistry 41 715–30
[40] Eastman P and Pande V S. 2010 Comput. Sci. Eng. 12 34–9
[41] Eastman P, Friedrichs M and Chodera J. 2013 J. Chem. Theory Comput. 9 461–9
[42] Shirts M R and Chodera J D. 2008 J. Chem. Phys. 129 124105