

# [DRAFT]: Best Practices for Thermodynamic Property Prediction from Molecular Simulations [Article v0.1]

Owen C. Madin<sup>1</sup>, Bryce C. Manubay<sup>1</sup>, John D. Chodera<sup>2</sup>, Michael R. Shirts<sup>1\*</sup>

<sup>1</sup>University of Colorado Boulder; <sup>2</sup>Sloan Kettering Institute

*This LiveCoMS document is maintained online on GitHub at <https://github.com/ocmadin/best-practices-paper>; to provide feedback, suggestions, or help improve it, please visit the GitHub repository and participate via the issue tracker.*

*This version dated February 13, 2019*

**Abstract** This document describes a collected set of best practices for computing various physical properties from molecular simulations of liquid mixtures. When computing these properties, parameters throughout the simulation-calculation workflow affect the accuracy and precision of these simulated properties. We present rationale for allocation of computational resources and choice of calculation methods for commonly simulated thermophysical properties, including densities, heat capacities, and dielectric constants. For future updates, we plan to include a variety of change-of-phase and mixture properties. We hope that this document will be a useful guide for a variety of uses, including force field benchmarking.

## \*For correspondence:

[michael.shirts@colorado.edu](mailto:michael.shirts@colorado.edu) (MRS); [email2@example.com](mailto:email2@example.com) (FS)

<sup>†</sup>These authors contributed equally to this work

<sup>‡</sup>These authors also contributed equally to this work

**Present address:** <sup>1</sup>Department of Chemical and Biological Engineering, University of Colorado, Boulder, CO; <sup>2</sup>Computational Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY

## 1 Introduction

Although simple thermophysical properties are not generally the end-use case for most molecular simulations, they are important for several force field uses. For example, during force field parameterization, comparison to experimental physical properties is a criteria often used to determine whether parameters are “good enough” to use for experimental purposes, commonly referred to as benchmarking. Calculation of thermophysical properties is also an important sanity check for simulation users; if the system being studied does not

reproduce physical properties, the chance of it being useful for other purposes is very low. Uncertainty quantification is also an important part of property calculation, and many choices users need to make affect uncertainty. From allocating computational resources to selecting equilibrium samples to choosing property calculation methods, there are “pitfalls” that can increase uncertainty if users are not aware of them.

## 2 Prerequisites

Add content about basic MD things, fluctuations, finite differences, and MBAR - might want to migrate some content from section on thermo preliminaries:

## 3 Checklist

Here we use a full-page checklist with multiple sections, so it will appear on a separate page of the sample PDF. Other checklist formats are possible, as shown in the sample `sample-document.tex` in [github.com/livecomsjournal/article\\_templates/templates](https://github.com/livecomsjournal/article_templates/templates).

Your checklist should include a succinct list of steps that people should follow when carrying out the task in question. This is provided to ensure certain basic standards are followed and common but critical major errors are avoided. Note that a checklist is not intended to cover *all* important steps, but rather focus on the most common reasons for failure or incorrect results, or issues which are particularly crucial.

## 4 Definitions and Thermodynamic Preliminaries

### Definitions

- $V$ : Volume
- $U$ : Total energy (including potential and kinetic, excluding external energy such as due to gravity, etc)
- $S$ : Entropy
- $N$ : Number of particles
- $T$ : Temperature
- $P$ : Pressure
- $k_B$ : Boltzmann constant
- $\beta$ :  $(k_B T)^{-1}$
- $M$ : Molar mass
- $\rho$ : Density ( $M/V$ )
- $H$ : Enthalpy
- $G$ : Gibbs Free Energy (free enthalpy)
- $A$ : Helmholtz Free Energy
- $\mu$ : Chemical potential
- $D$ : Total dipole moment
- $u$ : reduced energy
- $f$ : reduced free energy

Macroscopically, the quantities  $V$ ,  $U$ ,  $N$  are constants (assuming the system is not perturbed in any way), as we assume that the fluctuations are essentially zero, and any uncertainty comes from our inability to measure that constant value precisely. For a mole of compound (about 18 mL for water), the relative uncertainty in any of these quantities is about  $10^{-12}$ , far lower than any thermodynamics experiment can actually measure.

However, in a molecular simulation, these quantities are not necessarily constant. For example, in an  $NVT$  equilibrium simulation,  $U$  is allowed to vary. For a long enough simulation (assuming ergodicity, which can pretty much always be assumed with correctly implemented simulations and simple fluids), then the ensemble average value of  $U = \langle U \rangle$  will converge to a constant value, and in the limit of large simulations/long time will converge to the macroscopic value  $U$ ; at least, the macroscopic value of that given model, though perhaps not the  $U$  for the real system. In an  $NVT$  simulation, clearly  $V$  is constant. In an  $NPT$  simulation, however,  $V$  is a variable, and we must estimate what the macroscopic value would be using the ensemble average  $\langle V \rangle$ .

The quantities  $T$ ,  $P$ , and  $\mu$  are typically *set* as constant during the equilibrium simulations and experiments of interest here. More precisely, the system is in contact with a thermal bath with a fixed  $T$  (or in the case of  $NPT$  simulations, in contact with a thermal and mechanical bath), and we sample from the systems in equilibrium with this bath. There are a number of quantities that can be used to ESTIMATE constants such as  $T$  and  $P$ . For example,  $\langle \frac{1}{3Nk_B} \sum_i m_i |v_i|^2 \rangle$ , where  $m$  is the mass of each particle and  $|v_i|$  is the magnitude of the velocity of each particle, is an estimate of  $T$  (the temperature of the bath), and its average will be equal to the  $T$ . But it is not the temperature. This quantity fluctuates, but  $T$  remains constant; otherwise the simulation could not be at constant temperature.

Ensemble averages of some quantity  $X$  ( $\langle X \rangle$ ) are assumed to be averages over the appropriate Boltzmann weighting, i.e. in the  $NVT$  ensemble with classical statistical mechanics, they would be  $\int X(\vec{x}, \vec{p}) e^{-\beta U(\vec{x}, \vec{p})} d\vec{x} d\vec{p}$ . We note that in the limit of very large systems,  $\langle X \rangle_{NPT} = \langle X \rangle_{NVT} = \langle X \rangle_{\mu VT}$ .

Ensemble averages can be computed by one of two ways. First, they can be computed directly, by running a simulation that produces samples with the desired Boltzmann distribution. In that case ensemble averages can be computed as simple averages,  $\langle V \rangle = \frac{1}{N} \sum_i V_i$ , where the sum is over all observations. Uncertainties can be estimated in a number of different ways, but usually require estimating the number of uncorrelated samples. Secondly, they can be calculated as reweighted estimates from several different simulations, as  $\langle V \rangle = \frac{1}{\sum_i w_i} \sum_i V_i w_i$  where  $w_i$  is a reweighting factor that can be derived from importance sampling theory.

To simplify our discussion of reweighting, we use some additional notation. We define the reduced potential  $u = \beta U(\vec{x})$  in the canonical ( $NVT$ ) ensemble,  $u = \beta U + \beta PV$  in the isobaric-isothermal ( $NPT$ ) ensemble, and  $u = \beta U - \beta N\mu$  in the grand canonical ensemble (similar potentials can be defined in other ensembles). We then define  $f = \int e^{-u} dx$ , where the integral is over all of the DOF of the system ( $x$  for  $NVT$ ,  $x, V$  for  $NPT$ , and  $x, N$  for  $\mu VT$ ). For  $NPT$ , we then have  $f = \beta G$ , and for  $NVT$  we

have  $f = \beta A$ , while for  $\mu VT$  we have  $f = -\beta \langle P \rangle V$ .

To calculate expectations at one set of parameters generated with parameters that give rise to a different set of probability distributions, we start with the definition of an ensemble average given a probability distribution  $p_i(x)$ .

$$\langle X \rangle_i = \int X(x) p_i(x) dx \quad (1)$$

We then multiply and divide by  $p_j(x)$ , to get

$$\langle X \rangle_i = \int X(x) p_i(x) \frac{p_j(x)}{p_j(x)} dx = \int X(x) p_j(x) \frac{p_i(x)}{p_j(x)} dx \quad (2)$$

We then note that this last integral can be estimated by the Monte Carlo estimate

$$\langle X \rangle_i = \int X(x) p_j(x) \frac{p_i(x)}{p_j(x)} dx = \frac{1}{N} \sum_{n=1}^N X(x_n) \frac{p_i(x_n)}{p_j(x_n)} \quad (3)$$

Where the  $x_k$  are sampled from probability distribution  $p_j(x)$

We now define the mixture distribution of  $K$  other distributions as  $p_m(x) = \frac{1}{N} \sum_{i=1}^N N_k p_k(x)$ , where  $N = \sum_k N_k$ . We can construct a sample from the mixture distribution by simply pooling all the samples from  $k$  individual simulations. The formula for calculating ensemble averages in a distribution  $p_i(x)$  from samples from the mixture distribution is:

$$\langle X \rangle_i = \sum_{n=1}^N X(x_n) \frac{p_i(x_n)}{\sum_{k=1}^{N_k} p_k(x_n)} \quad (4)$$

In the case of Boltzmann averages, then  $p_i(x) = e^{f_i - u_i(x)}$ , where the reduced free energy  $f$  is unknown. Reweighting from the mixture distribution becomes.

$$\langle X \rangle_i = \sum_{n=1}^N X(x_n) \frac{e^{f_i - u_i(x_n)}}{\sum_{k=1}^{N_k} e^{f_k - u_k(x_n)}} \quad (5)$$

which can be seen to be the same formula as the MBAR formula for expectations. The free energies can be obtained by setting  $X = 1$ , and looking at the  $K$  equations obtained by reweighting to the  $K$  different distributions.

Finite differences at different temperatures and pressures can be calculated by including states with different reduced potentials. For example,  $u_j(x) = \beta_j U(x) + \beta_j (P_j + \Delta P)V$ , or  $u_j = \frac{1}{k_B(T_j + \Delta T)} U(x) + \frac{1}{k_B(T_j + \Delta T)} P_j V$ . However, the relationship between  $f$  and  $G$  can be problematic when looking at differences in free energy with respect to temperature, because  $G_2 - G_1 = \beta_2 f_2 - \beta_1 f_1$ . We can in general write

$$\Delta G_{ij}(T) = k_B T \left( \Delta f_{ij}(T) - \Delta f_{ij}(T_{ref}) \right) + \frac{T}{T_{ref}} \Delta G_{ij}(T_{ref})$$

, where  $\Delta G_{ij}(T_{ref})$  is known at some temperature.

Since with MBAR, one can make the differences as small as one would like (you don't have to actually carry out a simulation at those points), we can use the simplest formulas: central difference for first derivatives:

$$\frac{dA}{dx} \approx \frac{1}{2\Delta x} (A(x + \Delta x) - A(x - \Delta x))$$

And for 2nd derivatives:

$$\frac{d^2 A}{dx^2} \approx \frac{1}{\Delta x^2} (A(x + \Delta x) - 2A(x) + A(x - \Delta x))$$

Thus, only properties at two additional points need to be evaluated to calculate both first and 2nd derivatives.

It may first appear that these finite difference calculations will propagate significant error as they subtract similar numbers. However, MBAR calculates the covariance matrix between  $\langle A \rangle$ ,  $A(x + \Delta x)$ , and  $A(x - \Delta x)$ , meaning in practice the uncertainty is far lower than would be expected by standard error propagation of uncorrelated observables.

Note that if the finite differences are re-evaluated using reweighting approaches, it is important that the simulation used generates the correct Boltzmann distribution. If not, reweighted observables will be incorrect, and the results of the finite difference approach will have significant error.

The following document details calculation of various mechanical observables by both direct methods pulled from literature sources and the use of reweighting techniques. Corrections in certain observables are also summarized where suggested by previous authors.

## 5 Single Phase Properties

### 5.1 Pure Solvent Properties

#### 5.1.1 Density

Direct calculation

Starting with the equation used to calculate the density experimentally,

$$\rho = \frac{M}{V} \quad (6)$$

We replace the average with the ensemble estimate (calculated either directly, or with reweighting) to obtain:

$$\rho = \frac{M}{\langle V \rangle} \quad (7)$$

Derivative Estimate

From the differential definition of the Gibbs free energy  $dG = VdP - SdT + \sum_i \mu_i dN_i$  that  $V$  can be calculated from the Gibbs free energy as:

$$V = \left( \frac{\partial G}{\partial P} \right)_{T,N} \quad (8)$$

The density can therefore be estimated from the Gibbs free energy.

$$\rho = \frac{M}{\left( \frac{\partial G}{\partial P} \right)_{T,N}} \quad (9)$$

The derivative can be estimated using a central difference numerical method utilizing Gibbs free energies reweighted to different pressures.

$$\left( \frac{\partial G}{\partial P} \right)_{T,N} \approx \frac{G_{P+\Delta P} - G_{P-\Delta P}}{2\Delta P} \quad (10)$$

The density can then finally be estimated.

$$\rho \approx \frac{M}{\frac{G_{P+\Delta P} - G_{P-\Delta P}}{2\Delta P}} \quad (11)$$

This can be calculated from the reduced free energy  $f$  if desired by simply substituting:

$$\rho \approx \frac{\beta M}{\frac{f_{P+\Delta P} - f_{P-\Delta P}}{2\Delta P}} \quad (12)$$

Intuitively, one would imagine that equation 12 would be a worse estimate of density given that the calculations involved have more room for error than direct simulations. That being said, this method should prove invaluable when estimating densities of unsampled states using MBAR.

### 5.1.2 Molar Enthalpy

This section is on the relation of enthalpy to Gibbs free energy (should we need it). This is not an experimental quantity, but will be helpful in calculating related properties of interest. The enthalpy,  $H$ , can be found from the Gibbs free energy,  $G$ , by the Gibbs-Helmholtz relation:

$$H = -T^2 \left( \frac{\partial(\frac{G}{T})}{\partial T} \right)_{P,N} \quad (13)$$

Transforming the derivative in the Gibbs-Helmholtz relation to be in terms of  $\beta$  instead of  $T$  yields:

$$H = -T^2 \frac{\beta^2}{\beta^2} \left( \frac{\partial(\frac{G}{T})}{\partial T} \frac{\partial T}{\partial \beta} \frac{\partial \beta}{\partial T} \right)_{P,N} \quad (14)$$

Recall that  $\beta = \frac{1}{k_B T}$ , therefore  $\frac{\partial \beta}{\partial T} = -\frac{1}{k_B T^2}$ . Substituting these values into the enthalpy equation gives:

$$\begin{aligned} H &= \frac{1}{k_B^3 T^2 \beta^2} \left( \frac{\partial(\frac{G}{T})}{\partial \beta} \right)_{P,N} \\ &= \frac{1}{k_B} \left( \frac{\partial(\frac{G}{T})}{\partial \beta} \right)_{P,N} = \frac{\partial f}{\partial \beta}_{P,N} \end{aligned} \quad (15)$$

### 5.1.3 Heat Capacity

The definition of the isobaric heat capacity is:

$$C_P = \left( \frac{\partial H}{\partial T} \right)_{P,N} \quad (16)$$

$$C_P = \frac{\partial \left( \frac{\partial f}{\partial \beta} \right)}{\partial T}_{P,N} \quad (17)$$

$$C_P = -k_B \beta^2 \frac{\partial^2 f}{\partial \beta^2} \quad (18)$$

This could be computed by finite differences approach or analytical derivation using MBAR

The enthalpy fluctuation formula can also be used to calculate  $C_P$ [1].

$$C_P = \frac{\langle H^2 \rangle - \langle H \rangle^2}{N k_B \langle T \rangle^2} \quad (19)$$

The form is equivalent for isochoric heat capacity, but with derivatives at constant volume rather than pressure.

### 5.1.4 Isothermal Compressibility

The definition of isothermal compressibility is:

$$\kappa_T = -\frac{1}{V} \left( \frac{\partial V}{\partial P} \right)_T \quad (20)$$

#### First Derivative

Thus, it can be estimated by the finite difference of  $\langle V \rangle$

$$\kappa_T = -\frac{1}{2V(T, P)^2} (\langle V(P + \Delta P, T) \rangle - \langle V(P - \Delta P, T) \rangle) \quad (21)$$

Or by the finite differences evaluation of:

$$\kappa_T = -\frac{\left( \frac{\partial^2 G}{\partial P^2} \right)_{T,N}}{\left( \frac{\partial G}{\partial P} \right)_{T,N}} = -\frac{\left( \frac{\partial^2 f}{\partial P^2} \right)_{T,N}}{\left( \frac{\partial f}{\partial P} \right)_{T,N}} \quad (22)$$

$\kappa_T$  can also be estimated from the ensemble average and fluctuation of volume (in the NPT ensemble) or particle number (in the  $\mu VT$  ensemble)[2]:

$$\kappa_T = \beta \frac{\langle \Delta V^2 \rangle_{NTP}}{\langle V \rangle_{NTP}} = V \beta \frac{\langle \Delta N^2 \rangle_{VT}}{\langle N \rangle_{VT}} \quad (23)$$

### 5.1.5 Speed of Sound

The definition of the speed of sound is[3]:

$$c^2 = \left( \frac{\partial P}{\partial \rho} \right)_S = -\frac{V^2}{M} \left( \frac{\partial P}{\partial V} \right)_S \quad (24)$$

$$c^2 = \frac{V^2}{\beta M} \left[ \left( \frac{\gamma_V}{k_B} \right)^2 + \frac{\beta}{V \kappa_T} \right] \quad (25)$$

Where:

$$\gamma_V = \left( \frac{\partial P}{\partial T} \right)_V \quad (26)$$

$\gamma_V$  is known as the isochoric pressure coefficient.  $\kappa_T$  is the same isothermal compressibility from equation 20

An alternate derivation, applying the triple product rule to  $\left(\frac{\partial P}{\partial V}\right)_S$  yields the following.

$$\left(\frac{\partial P}{\partial V}\right)_S = \frac{\left(\frac{\partial S}{\partial V}\right)_P}{\left(\frac{\partial S}{\partial P}\right)_V} \quad (27)$$

$$\left(\frac{\partial S}{\partial V}\right)_P = \left(\frac{\partial S}{\partial T}\right)_P \left(\frac{\partial T}{\partial V}\right)_P = \frac{C_P}{T} \left(\frac{\partial T}{\partial V}\right)_P = \frac{C_P}{TV\alpha} \quad (28)$$

Where  $\alpha = \frac{1}{V} \left(\frac{\partial V}{\partial T}\right)_P = \left(\frac{\partial \ln V}{\partial T}\right)_P$  is the coefficient of thermal expansion. The second term in our triple product rule expansion,  $\left(\frac{\partial S}{\partial P}\right)_V$ , can be expressed as follows:

$$\left(\frac{\partial S}{\partial P}\right)_V = \left(\frac{\partial S}{\partial T}\right)_V \left(\frac{\partial T}{\partial P}\right)_V = \frac{C_V}{T} \left(\frac{\partial T}{\partial P}\right)_V = \frac{C_V}{T\gamma_V} \quad (29)$$

Thus our derivation yields:

$$\left(\frac{\partial P}{\partial V}\right)_S = \frac{C_P\gamma_V}{C_VV\alpha} \quad (30)$$

To avoid running an NVT simulation to determine  $C_V$ , the following relationship between  $C_P$  and  $C_V$  can be used:

$$C_P - C_V = TV \left(\frac{\alpha^2}{\kappa_T}\right) \quad (31)$$

This yields the following expression:

$$\left(\frac{\partial P}{\partial V}\right)_S = \left(\frac{C_P}{C_P - TV\frac{\alpha^2}{\kappa_T}}\right) \left(\frac{\gamma_V}{V\alpha}\right) \quad (32)$$

Horn et al set out several ways for calculating  $\alpha$ [1]:

Analytical derivative of density with respect to temperature

$$\alpha = -\frac{d \ln \langle \rho \rangle}{dT} \quad (33)$$

Numerical derivative of density over range of T of interest

The same finite differences approach as shown for isothermal compressibility can be applied here, thus:

$$\alpha = -\frac{d \ln \langle \rho \rangle}{dT} = -\frac{1}{2\rho(T, P)} (\ln \langle \rho(P, T + \Delta T) \rangle - \ln \langle \rho(P, T - \Delta T) \rangle) \quad (34)$$

Using the enthalpy-volume fluctuation formula

$$\alpha = \frac{\langle VH \rangle - \langle V \rangle \langle H \rangle}{k_B \langle T \rangle^2 \langle V \rangle} \quad (35)$$

Finite differences approximations and/or analytical derivation can also be used to calculate  $\gamma_V$  or by note of the relation:

$$\gamma_V = -\frac{\alpha}{\kappa_T} \quad (36)$$

Substituting Equations (24),(32), and (36), the following easily calculable expression for speed of sound is obtained:

$$c^2 = \frac{V}{M\kappa_T} \left( \frac{C_P}{C_P - TV\frac{\alpha^2}{\kappa_T}} \right) \quad (37)$$

Which can also be written in the simplified form:

$$c^2 = \frac{\gamma}{\rho\kappa_T} \quad (38)$$

Where  $\gamma = \frac{C_P}{C_V}$ .

### 5.1.6 Dielectric Constant

The dielectric constant is a simple and important quantity of interest in parameterization, as it can capture some of the accuracy of the electrostatics of the system. Fundamentally, the dielectric constant is related to the polarization of a material when an external field is applied: [? ]

$$\vec{P} = \frac{1}{4\pi} (\epsilon(0) - 1) \vec{E}^{ext} \quad (39)$$

Fundamentally, there are two ways that this dielectric constant can be calculated.

The first, and most common, is to use the fluctuations of the system dipole moment  $\vec{M}$  in the limit of low field strength. This method allows for the calculation of dielectrics from equilibrium simulations without an applied electric field, and with long-range interactions calculated via PME. For this method, the equation was provided by a literature reference authored by CJ Fennell[4] and is the standard for calculating the dielectric constant. Below,  $\epsilon(0)$  is the zero frequency dielectric constant,  $V$  is the system volume and  $D$  is the total system dipole moment.

$$\epsilon(0) = 1 + \frac{4\pi}{3k_B T \langle V \rangle} (\langle D^2 \rangle - \langle D \rangle^2) \quad (40)$$

A second way of doing this is by applying an external electric field and measuring the polarization response in the direction of that field. The equation for this method, given by Riniker [? ] is as follows:

$$\epsilon(0) = 1 + 4\pi \frac{\langle M_z \rangle}{V E_z^{ext}} \quad (41)$$

This method takes a bit more care to implement than the fluctuation method, as it is important to ensure that the response of the Polarization is in the linear regime (low field strength), where this regime changes from material to material. This method is also generally used with a dielectric continuum rather than periodic boundary conditions, which may affect its utility. However, when implemented properly, this method can yield the same results as the fluctuation method with an order of magnitude less time. For fluctuations, simulations of 10 ns or longer are recommended for convergence, while Riniker[?] found that only 500 ps per field strength (calculated with multiple field strengths) was sufficient. Kolafa [?] suggests the metric of saturation of polarization  $S$  as a metric to assess whether the response is in the linear regime. The saturation of polarization is defined as the polarization of the system divided by the total possible polarization of the system, i.e. all molecules are aligned.

The saturation is given by the following two equations, for fluctuations and external field respectively:

$$S = \frac{\langle D^2 \rangle^{1/2}}{N\mu} \quad (42)$$

$$S = \frac{\langle D_z \rangle}{N\mu} \quad (43)$$

Where  $\mu$  is the dipole moment of the individual molecule, and  $N$  is the number of molecules in the system. Kolafa suggests a target value of  $S \leq 0.1$  to keep systematic error from the saturation to a minimum.

## 5.2 Binary Mixture Properties

### 5.2.1 Mass Density, Speed of Sound and Dielectric Constant

The methods for these calculations are the same for a multi-component system.

### 5.2.2 Activity Coefficient

The definition of chemical potential in a pure substance is:

$$\mu(T, P) = \left( \frac{\partial G}{\partial N} \right)_{T, P} \quad (44)$$

which is a function of only temperature and pressure.

Then the definition of the chemical potential  $\mu_i$  of compound  $i$  in a mixture is:

$$\mu_i(T, P, \vec{N}) = \left( \frac{\partial G}{\partial N_i} \right)_{T, P, N_{j \neq i}} \quad (45)$$

$N_i$  refers to a molecule of component  $i$  and  $N_{j \neq i}$  refers to all molecules other than component  $i$ , with  $\vec{N}$  the vector of all component numbers. Since  $\mu_i$  is intensive, this is equivalently a function of the vector of mole fractions  $\vec{x}_i$  instead of simply

of  $N_i$ .

For an ideal solution, the chemical potential  $\mu_i$  can be related to the pure chemical potential by

$$\mu_i(T, P, \vec{x}_i) = \mu(T, P) + k_B T \ln(\gamma_i) \quad (46)$$

By analogy to this form, we can say

$$\mu_i(T, P, \vec{x}_i) = \mu(T, P) + k_B T \ln(x_i \gamma_i) \quad (47)$$

Where  $\gamma_i$  is the activity coefficient of component  $i$ , and is a function of  $T, P$ , and  $\vec{x}_i$ . Rearrangement of the previous equation yields:

$$\gamma_i = \frac{e^{\left( \frac{\mu_i(T, P, \vec{x}_i) - \mu(T, P)}{k_B T} \right)}}{x_i} \quad (48)$$

Although chemical potentials cannot be directly calculated from simulation, chemical potential residuals can. We can calculate the difference  $\mu_i(T, P, \vec{x}_i) - \mu(T, P)$  by calculating  $\Delta\mu(T, P)_{liquid} - \Delta\mu(T, P)_{gas}$  using a standard alchemical simulation of the pure substance, followed by the calculation of  $\mu_i(T, P, \vec{x}_i)_{liquid} - \Delta\mu(T, P, \vec{x}_i)_{gas}$ , and assuming that  $\Delta\mu(T, P, \vec{x}_i)_{gas} = \Delta\mu(T, P)_{gas}$ . Note: there are a few subtleties here relating to the  $\ln x_i$  factor, but it appears that with alchemical simulations with only one particle that is allowed to change, this will cancel out (need to follow up).

Several of these alchemical simulation methods for calculating activity coefficients have been pioneered by Andrew Paluch [5]. A method detailing the calculation of infinite dilution activity coefficients  $\gamma_i^{inf}$  for binary a mixture follows directly:

$$\ln \gamma_2^\infty(T, P, x_2 = 0) = \beta \mu_2^{res, \infty}(T, P, N_1, N_2 = 1) + \ln \left[ \frac{RT}{V_1(T, P)} \right] - \ln f_2^0(T, P) \quad (49)$$

Where  $\beta \mu_2^{res, \infty}$  is the dimensionless residual chemical potential of component 2 at infinite dilution. The residual is defined here as the difference between the liquid and ideal gas state.  $V_1(T, P)$  is the molar volume of component 1 at  $T$  and  $P$ .  $\ln f_2^0(T, P)$  is the natural logarithm of the pure liquid fugacity of component 2 and is defined as:

$$\ln f_2^0(T, P) = \beta \mu_2^{res}(T, P) + \ln \left[ \frac{RT}{V_2(T, P)} \right] \quad (50)$$



Paluch et al. use a multistage free energy perturbation approach utilizing MBAR in order to calculate the residual chemical potentials (recall that the chemical potential is the partial molar Gibbs free energy and dimensionless Gibbs free energy differences between multiple states are readily computed with MBAR). The idea is to connect two states of interest. In the case of a pure liquid, connecting a system of pure liquid molecules with  $N-1$  interacting molecules and one fully decoupled molecule to a system of  $N$  fully interacting molecules. The coupling/decoupling process is detailed by Paluch et al [6], but involves a linear alchemical switching function where LJ and electronic interactions are slowly turned on for the decoupled molecule until they are fully on. The free energy of this coupling is calculated by simply summing the free energy changes along this path.

### 5.2.3 Excess Molar Properties

The general definition of an excess molar property can be stated as follows:

$$y^E = y^M - \sum_i x_i y_i \quad (51)$$

Where  $y^E$  is the excess molar quantity,  $y^M$  is the mixture quantity,  $x_i$  is the mole fraction of component  $i$  in the mixture and  $y_i$  is the pure solvent quantity. In general, the simplest methods for calculating excess molar properties for binary mixtures will require three simulations. One simulation is run for each pure component and a third will be run for the specific mixture of interest. We note that only one set of pure simulations are needed to calculate excess properties at all compositions.

### 5.2.4 Excess Molar Heat Capacity and Volume

Excess molar heat capacities and volume will be calculated using the methods for the pure quantities in section 1 in combination with the general method for excess property calculation above.

### 5.2.5 Excess Molar Enthalpy

Excess molar enthalpy can be calculated using the general relation of molar enthalpy as it relates to Gibbs Free Energy from section 1 and the general method of excess molar property calculation above or by the following[7]:

$$H^E = \langle E^M \rangle + PV^E - \sum_i x_i \langle E_i \rangle \quad (52)$$

Where  $\langle E \rangle$  denotes an ensemble average of total energy and  $V^E$  is calculated using the general method of excess molar properties.

## 5.3 Suggested Corrections

### 5.3.1 Heat Capacity

Horn et al suggest a number of vibrational corrections be applied to the calculation of  $C_p$  due to a number of approximations made during the simulation of the liquid [1]. The following terms were added as a correction.

$$\left( \frac{\partial E_{vib,l}}{\partial T} \right)_p = \left( \frac{\partial E_{vib,l,intra}^{QM}}{\partial T} \right)_p + \left( \frac{\partial E_{vib,l,inter}^{QM}}{\partial T} \right)_p - \left( \frac{\partial E_{vib,l,inter}^{CM}}{\partial T} \right)_p \quad (53)$$

Where:

$$\left( \frac{\partial E_{vib}^{CM}}{\partial T} \right)_p = k_B n_{vib} \quad (54)$$

$$\left( \frac{\partial E_{vib}^{QM}}{\partial T} \right)_p = \sum_{i=1}^{n_{vib}} \left( \frac{h^2 v_i^2 e^{\frac{h v_i}{k_B T}}}{k_B T^2 \left( e^{\frac{h v_i}{k_B T}} - 1 \right)^2} \right) \quad (55)$$

Above,  $n_{vib}$  is the number of vibrational modes,  $h$  is Planck's constant and  $v_i$  is the vibrational frequency of mode  $i$ .

## 6 Properties Involving Change of Phase

### 6.1 Pure Solvent Properties

#### 6.1.1 Enthalpy of Vaporization

The definition of the enthalpy of vaporization is[8]:

$$\Delta H_{vap} = H_{gas} - H_{liq} = E_{gas} - E_{liq} + P(V_{gas} - V_{liq}) \quad (56)$$

If we assume that  $V_{gas} \gg V_{liq}$  and that the gas is ideal (and therefore kinetic energy terms cancel):

$$\Delta H_{vap} = E_{gas,potential} - E_{liq,potential} + RT \quad (57)$$

## 6.2 Suggested Corrections

### 6.2.1 Enthalpy of Vaporization

An alternate, but similar, method for calculating the enthalpy of vaporization is recommended by Horn et al [1].

$$\Delta H_{vap} = -\frac{E_{liq,potential}}{N} + RT - PV_{liq} + C \quad (58)$$

In the above equation  $C$  is a correction factor for vibrational energies, polarizability, non-ideality of the gas and pressure. It can be calculated as follows.

$$\begin{aligned} C_{vib} &= C_{vib,intra} + C_{vib,inter} \\ &= \left( E_{vib,QM,gas,intra} - E_{vib,QM,liq,intra} \right) \\ &\quad + \left( E_{vib,QM,liq,inter} - E_{vib,CM,liq,inter} \right) \end{aligned} \quad (59)$$

The  $QM$  and  $CM$  subscripts stand for quantum and classical mechanics, respectively.

$$C_{pol} = \frac{N}{2} \frac{(d_{gas} - d_{liq})^2}{\alpha_{p,gas}} \quad (60)$$

Where  $d_i$  is the dipole moment of a molecule in phase  $i$  and  $\alpha_{p,gas}$  is the mean polarizability of a molecule in the gas phase.

$$C_{ni} = P_{vap} \left( B - T \frac{dB}{dT} \right) \quad (61)$$

Where  $B$  is the second virial coefficient.

$$C_x = \int_{P_{ext}}^{P_{vap}} [V(P_{ext}) [1 - (P - P_{ext})^{\kappa_T}] - TV\alpha] dP \quad (62)$$

Where  $P_{ext}$  is the external pressure and  $V(P_{ext})$  is the volume at  $P_{ext}$ .

This is frequently done as a single simulation calculation by assuming the average intramolecular energy remains constant during the phase change, which is rigorously correct for something like a rigid water molecule (intramolecular energies are zero), but less true for something with structural rearrangement between gas and liquid phases.

As discussed by myself and MRS, we have decided to not initially begin the parametrization process using enthalpy of vaporization data. While force field parametrization is commonly done using said property we have ample reason to

not follow classical practice. First of all, the enthalpy data is usually not collected at standard temperature and pressure, but at the saturation conditions of the liquid being vaporized [9]. This would require corrections to be made to get the property at STP (the process will be explained below) using fitted equations for heat capacity. Not only is this inconvenient, but it adds an unknown complexity when adjusting experimental uncertainties due to the added correction. Often times the uncertainties of these "experimental" enthalpies are unrecorded because they are estimated from fitted Antoine equation coefficients [9].

An additional issue is the necessity of having to use gas phase simulation data in order to validate a parametrization process meant for small organic liquids and their mixtures. Following an example of Wang et al. [10] we plan to instead use enthalpy of vaporization calculations as an unbiased means of testing the success of the parametrization. If the parametrization procedure is expanded to use enthalpy of vaporization, corrections can be made to the experimental data in order to get a value at STP using the following equation.

$$\Delta H_{vap}(T) = \Delta H_{vap}^{ref} + \int_{T_{ref}}^T (C_{p,gas} - C_{p,liq}) dT \quad (63)$$

## 7 Numerical Considerations

### 7.1 Quality of Timeseries Data

With the finite computing resources that are available today, it is often necessary to make trade-offs between the uncertainty of the result and the amount of time/resources spent. As a major assumption of Molecular Dynamics simulations is the Ergodic Hypothesis, sampling over a timeseries will only be equivalent in the limit of infinite time. In all other cases, the timeseries generated from MD will function as an *approximation* to phase space sampling. As this is the case, there is an inherent decision to be made in the collection of all timeseries data: how much effort does one need to spend to get a result with the appropriate confidence for the specific problem being addressed?

For bulk thermophysical properties, there are two major factors that consume computational resources: the length of the simulation, often measured in time units such as nanoseconds or microseconds, and the size of the system that is simulated, measured in number of molecules. Because timescales vary based on the nature of the system, a more general metric that we will use in this discussion is  $N$ , the number of uncorrelated equilibrium samples. We will denote the number of molecules as  $M$  in this discussion to distinguish the two.



### 7.1.1 Equilibrium Sampling

The relationship between  $N$  and uncertainty is straightforward; as one collects more samples of the same system, the standard deviation decreases as  $\mathcal{O}(N^{-\frac{1}{2}})$ , as would be expected for the standard formula for sample standard deviation. It is worth noting that while  $N$  follows this rule, the length of the simulation will not follow it exactly. This is because the detection of equilibration and the choice of uncorrelated equilibrium samples will vary between users and systems.

For a timeseries to satisfy the Ergodic Hypothesis and allow for calculation of equilibrium samples, the system needs to have settled into an equilibrium region, following a "burn-in" period in which the system starts at some non-equilibrium configuration and relaxes into an equilibrium configuration. This portion of a timeseries needs to be removed; the question as to how to do so is slightly trickier. For users running many simulations of similar systems, often a burn-in time is determined by inspection and then set for all subsequent simulations. However, this can backfire based on the simulation parameters being varied. For example, systems with small numbers of molecules can equilibrate much slower than larger systems, making the choice of equilibrium for one invalid for another, or discarding valuable equilibrium data. Another option is an automatic equilibration detection algorithm, like the one proposed by Chodera. In this work, this automatic equilibration algorithm is used in all analysis. It is important to note, however, that algorithms like this should not be used in a vacuum; the user should periodically check to make sure that the algorithm is functioning properly, and have a rudimentary understanding of how it works.

Another important factor in the calculation of thermophysical properties is the collection of *uncorrelated* equilibrium samples from an equilibrated timeseries. Although the system may be at equilibrium constantly after the burn-in period, samples close to each other in time are likely correlated. Two common methods for obtaining uncorrelated samples are through subsampling and block averaging. In block averaging, the timeseries is separated into "blocks" of data and observations are taken from the mean of each block. This requires for a size of block to be chosen, and as the block size increases, the variance of the observable will eventually reach a plateau where samples are uncorrelated. Another method is subsampling, where the statistical inefficiency  $g$  is estimated and then snapshots are chosen based on this parameter, without averaging. In this work, subsampling is used for all calculations.

## 7.2 System Size

For the number of molecules  $M$ , relationships with uncertainty are more complicated and vary from property to property. This can lead to unintuitive results in which a larger  $M$  may

reduce uncertainty in one physical quantity, but for a *derived* property of said quantity, the same increase in  $M$  may not improve estimates.

### 7.2.1 Enthalpy and Heat Capacity

This is best illustrated with the example of energy and heat capacity. From Eq. 19, the *extensive* heat capacity is directly proportional to the variance in the enthalpy, which is defined as  $\langle H^2 \rangle - \langle H \rangle^2$ . For any extensive property, its value has a linear relationship with  $M$ ; as you go from 100 molecules to 200 molecules, the total heat capacity (or enthalpy, or volume, etc.) will double.

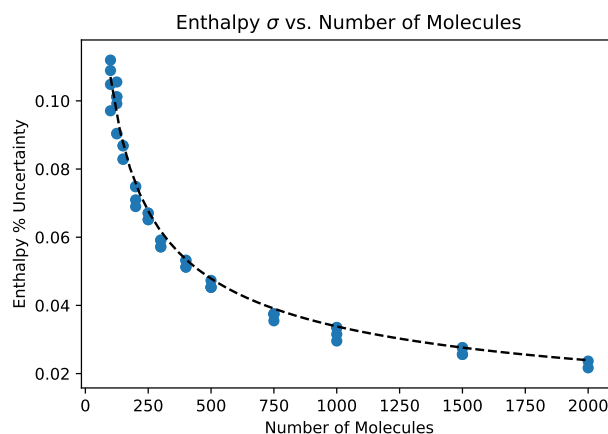
Therefore, the heat capacity goes as  $C_p \sim \mathcal{O}(M)$ , and subsequently, so does the variance of the enthalpy:  $\text{Var}(H) \sim C_p \sim \mathcal{O}(M)$ . If we wish to examine the variance of the *molar* enthalpy, an *intensive* quantity, we can make use of this standard formula for variances:

$$\text{Var}(aX) = a^2 \text{Var}(X) \quad (64)$$

Since molar enthalpy is defined as  $h = \left(\frac{H}{M}\right)$ , the variance of the molar enthalpy is given by:

$$\text{Var}(h) = \text{Var}\left(\frac{H}{M}\right) = \frac{1}{M^2} \text{Var}(H) \sim \frac{\mathcal{O}(M)}{M^2} \sim \mathcal{O}(1/M) \quad (65)$$

So, the standard deviation of the molar enthalpy goes as  $\mathcal{O}(M^{-\frac{1}{2}})$ , meaning that quadrupling the size of the system will cut the uncertainty in half, just as quadrupling  $N$  would. This relationship is shown in figure:



**Figure 1.** Figure shows enthalpy % uncertainty (calculated as  $(\sigma_h/h) \times 100\%$ ) vs.  $M$  for simulations of liquid cyclohexane at 293.15 K, 1.01 bar, calculated via fluctuations.  $\mathcal{O}(M^{-\frac{1}{2}})$  on  $M$ .

Now, we examine the uncertainty behavior of  $C_p$ , the extensive heat capacity. Since we know  $C_p \sim \text{Var}(H)$ , it follows that  $\text{Var}(C_p) \sim \text{Var}(\text{Var}(H))$ . For normally distributed samples, the variance of the variance is given by:

$$\text{Var}(\text{Var}(X)) \approx \frac{2(N-1)}{N^2} (\text{Var}(X))^2 \quad (66)$$

Holding  $N$  constant and applying to the enthalpy, we obtain the following:

$$\text{Var}(C_p) \sim \text{Var}(\text{Var}(H)) \sim (\text{Var}(H))^2 \sim \mathcal{O}(M^2) \quad (67)$$

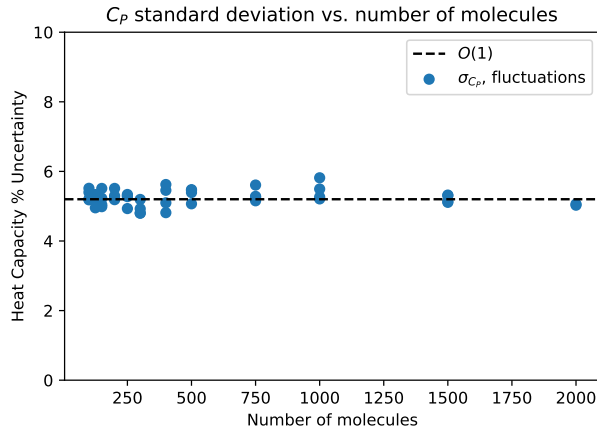
As we examine the *molar* heat capacity  $c_p = C_p/M$  and apply the variance formula from above, we obtain:

$$\text{Var}(c_p) = \text{Var}\left(\frac{C_p}{M}\right) = \frac{\text{Var}(C_p)}{M^2} \sim \frac{\mathcal{O}(M^2)}{M^2} \sim \mathcal{O}(1) \quad (68)$$

So, even though estimates of the molar enthalpy  $h$  improve as the size of the system increases, estimates of the molar heat capacity  $c_p$ , which is alternatively defined as the derivative of  $h$  with respect to  $T$ , do not. Note that while this derivation is for NPT systems, a similar derivation follows for NVT systems, replacing  $H$  with  $E$  and  $C_p$  with  $C_v$ .

### 7.2.2 Volume and compressibility

For the volume  $V$ , the standard deviation also goes as  $\mathcal{O}(M^{-1/2})$ . This follows from the information we already know about energy and enthalpy and the relationship between the two. Since they are related by  $H = E + PV$ , and the variance of  $H$  and  $E$  both go as  $\mathcal{O}(M)$  due to their relationship with the heat capacities  $C_p$  and  $C_v$ , the variance of  $V$  must also go as  $\mathcal{O}(M)$ . By similar analysis to that for the molar enthalpy  $h$ , it follows that the standard deviation of the molar volume  $v$  goes as  $\mathcal{O}(M^{-1/2})$ , and this claim is supported by experimental data in figure.



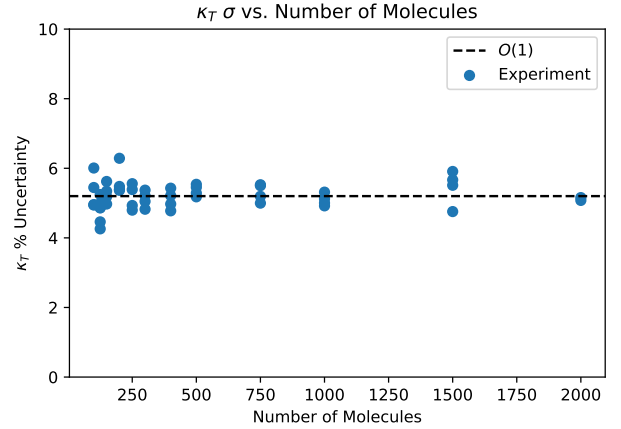
**Figure 2.** Figure shows heat capacity % uncertainty (calculated as  $(\sigma_{C_p}/C_p) \times 100\%$ ) vs.  $M$  for simulations of liquid cyclohexane at 293.15 K, 1.01 bar, calculated via fluctuations. No dependence on  $M$ .

For the isothermal compressibility, the variance is easy to calculate once the variance of the volume is known. From Eq. 23, we can see that with constant temperature, the variance of  $\kappa_T$  is as such:

$$\text{Var}(\kappa_T) \sim \text{Var}\left(\frac{\text{Var}(V)}{V}\right) = \frac{1}{V^2} \text{Var}(\text{Var}(V)) \quad (69)$$

$$\text{Var}(\kappa_T) \sim \frac{1}{V^2} (\text{Var}(V))^2 \sim \frac{\mathcal{O}(M^2)}{\mathcal{O}(M^2)} \sim \mathcal{O}(1) \quad (70)$$

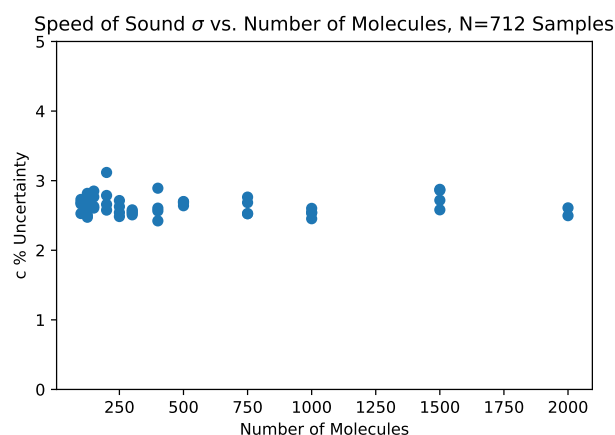
So, the uncertainty of the compressibility is independent of the number of molecules used to simulate the system. This claim is supported by experimental data in figure 3:



**Figure 3.** Figure shows speed of sound % uncertainty (calculated as  $(\sigma_{\kappa_T}/\kappa_T) \times 100\%$ ) vs.  $M$  for simulations of liquid cyclohexane at 293.15 K, 1.01 bar, calculated via fluctuations. No dependence on  $M$ .

### 7.2.3 Speed of Sound

For thermophysical properties that are calculated from several other basic properties, such as speed of sound  $c = \frac{\gamma}{\rho \kappa_T}$ , it is important to examine the relative contributions of the error to determine the overall uncertainty behavior of the property. For example,  $\gamma$  and  $\kappa_T$  are both properties with uncertainties that do not depend on  $M$ , but  $\rho$  should have an  $\mathcal{O}(N^{-1/2})$  uncertainty behavior, similar to molar volume. Therefore, one would expect the overall uncertainty behavior to be dependent on  $M$ , but data from experiment in Figure 4 shows no correlation.



**Figure 4.** Figure shows speed of sound % uncertainty (calculated as  $(\sigma_c/c) \times 100\%$ ) vs.  $M$  for simulations of liquid cyclohexane at 293.15 K, 1.01 bar, calculated via fluctuations. No apparent dependence on  $M$ .

#### 7.2.4 Dielectric Constant

For calculations of the relative permittivity from the fluctuation method, the main relationship of interest is the dependence of the saturation  $S$  on the number of molecules  $N$ . With appropriately chosen values of  $S$ , the statistical error is minimized; it is therefore of interest to understand this relationship. It is important to note that the electric field of the continuum of permittivity that surrounds the Ewald boundary can also affect  $S$ , but because this is not typically changed in equilibrium PME simulations, it is not considered here.

#### 7.2.5 Discussion

From these analyses, we can see that properties taken from direct averages of timeseries data,  $(V, E, H)$  have standard deviation that goes as  $\mathcal{O}(M^{-1/2})$ . The molar versions of these properties subsequently have standard deviation that goes as  $\mathcal{O}(M^{-1/2})$ . However, this does not imply the same for properties calculated via fluctuations. Fluctuation properties can have standard deviation as  $\mathcal{O}(M)$ , like  $C_p$ , or with no dependence on  $M$ , like  $c_p$  or  $\kappa_T$ . Additionally, they can have a very weak dependence on  $M$ , such as speed of sound  $c$ . Because of these varying dependencies between quantities, and the extra computational cost of a larger system size, we recommend that additional computational resources be used to collect more equilibrium samples, which has a consistent effect on uncertainty for all properties. This recommendation comes with the caveat that systems need to be large enough to avoid finite size effects. When systems are too small, systems may not equilibrate very quickly, amongst other problems. For a more in-depth discussion of finite size effects, see. It is also important to remember when modifying the number of molecules in a system that Molecular Dynamics programs report energies in units of energy/mole of box, rather than en-

ergy per mole of compound. For pure compounds, one must divide the "moles of box" value by the number of molecules in order to get a true molar energy.

### 7.3 Comparison of Calculation Methods

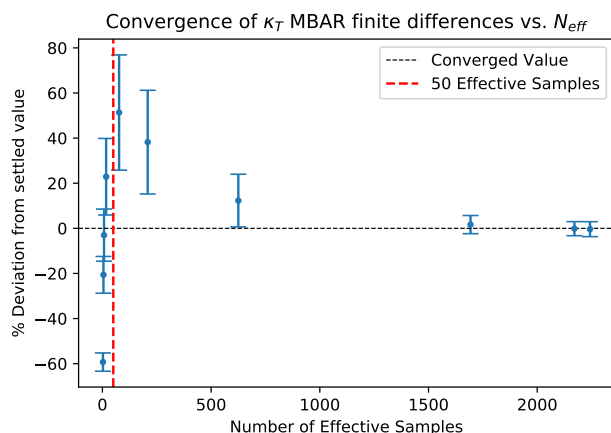
For the properties discussed in this paper, the methods of calculation generally fall into one of two categories: derivatives via finite differences or calculations from fluctuation properties. Therefore, it is worth investigating whether these all methods are equivalent, or if there are differences in methods which may make them more or less attractive, depending on the situation.

#### 7.3.1 Comparison of MBAR finite differences and direct finite differences

Suppose we are interested in calculating a thermodynamic property  $B$  at state  $X$ , defined as  $B = dA/dX$ . When calculating this property from timeseries data via centered finite difference (such as in Eq. 10,21), the simplest and most nature choice is to choose a finite difference  $\Delta X$ , perform simulations at  $X - \Delta X$  and  $X + \Delta X$ , take the expectations  $\langle A(X - \Delta X) \rangle$  and  $\langle A(X + \Delta X) \rangle$ , and then calculate the finite difference from these two values ("Direct Finite Differences"). Another possibility is take a simulation at  $X$ , and use reweighting via the MBAR method [?] to estimate  $A$  at  $X - \Delta X$  and  $X + \Delta X$ , then calculate the finite difference from these estimates ("MBAR Finite Differences"). The choice of finite difference is important here, as  $\Delta X$  will determine the accuracy of the calculation. Theoretically, we would like the smallest finite difference possible to minimize potential bias in the calculation. For direct finite differences, however, a small  $\Delta X$  will provide unreliable estimates, as timeseries data at very similar state points are highly correlated, leading to high uncertainty. As  $\Delta X$  increases, the timeseries data becomes less correlated and uncertainty goes down, but there is potential for bias, which depends on the behavior of the function.

When using MBAR finite differences, the problem is opposite: MBAR estimates are inherently decorrelated, leading to decent uncertainty at small  $\Delta X$ , but MBAR requires significant phase overlap between the simulated state and the desired state, so estimates become more unreliable as  $\Delta X$  increases. However, since MBAR finite differences are stable over a large range of  $\Delta X/X$  ( $10^{-6}$  –  $10^{-3}$ ) for heat capacities, this is less of an issue. However, it does make comparisons of direct finite differences and MBAR finite differences difficult, as the minimum  $\Delta X$  in direct finite differences that produces uncertainties equivalent to MBAR finite differences is roughly  $10^{-2}$  in heat capacity; MBAR estimates are already poor at this point. In order to quantify the reliability of the MBAR finite difference, the number of effective samples ( $N_{eff}$ ) metric can be used. This metric quantifies how many of the sam-

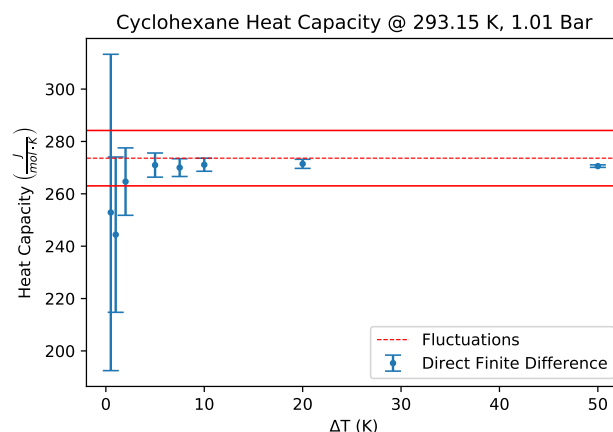
ples from simulated state A could be drawn from reweighted state B. Previous use of this metric has shown that a rough minimum of  $N_{eff} = 50$  is required for statistics to be stable, but for *accurate* finite differences, the percentage of effective samples ( $(N_{eff}/N) \times 100\%$ ) is a better metric, but may depend on the property being calculated. For example, MBAR finite differences for  $C_p$  agree with fluctuations with roughly 10 % of effective samples, but is more like 75% for  $\kappa_T$ .



**Figure 5.** Percent deviation from fluctuation value as a function of  $N_{eff}$ . Dashed red line indicates minimum requirement of 50 samples. Error bars indicate bootstrapped 95% CI

Since the number of effective samples is easily adjusted, as a best practice, we recommend a percentage of effective samples of 90 % or higher. It is also important to recognize that for properties that require reweighting in multiple variables, such as speed of sound, the number of effective samples for all variables needs to be taken into account (in the example of speed of sound,  $T$  and  $P$ ). In general, comparison with fluctuation calculations is the ultimate metric for accuracy of MBAR finite differences.

To test this, we performed a 10 ns simulation of cyclohexane at 293.15 K and 1.01 bar, as well as several pairs of 5 ns cyclohexane simulations, also at 1.01 bar, but spaced at intervals of 0.5, 1, 2, 5, 7.5, and 10 K apart from 293.15 K. Heat capacity calculations from finite differences were performed via MBAR on the 10 ns simulation, and directly on the pairs of 5 ns simulations.



**Figure 6.** Comparison of direct finite difference methods and MBAR reweighted finite differences for liquid cyclohexane  $C_p$  at 293.15 K, 1.01 bar. Red lines show MBAR estimate and uncertainty, error bars indicate bootstrapped 95 % CI.

### 7.3.2 Finite differences and fluctuations

For properties that can be calculated via finite differences or fluctuations, such as  $\kappa_T$  (Eq. 21, 23),  $C_p$  (Eq. 16, 19), or  $\alpha$  (Eq. 33,35), the two methods should be theoretically equivalent, as fluctuation formulas are derived by substituting statistical mechanical expressions for energy, enthalpy, or volume into derivative formulas and evaluating the derivative. In the limit of infinitely small finite differences and infinitely long timeseries, this will be exactly true; however, since neither condition can be met, it is important to understand if these methods are approximately equivalent under normal conditions.

Because of the large number of particles and the chaotic nature of the interactions between them, replicate timeseries are often slightly different, and will provide substantially different sets of uncorrelated equilibrium samples. Therefore, agreement between finite differences and fluctuation methods are not based on agreement to some "grand truth" but agreement between the data used to compute the derivative and the fluctuations. This point is illustrated with the following relation:

$$\frac{H(T_2) - H(T_1)}{T_2 - T_1} \approx \left( \frac{\partial H}{\partial T} \right)_P = \left( \frac{\text{Var}(H)}{RT^2} \right)_\infty \approx \left( \frac{\text{Var}(H)}{RT^2} \right) \quad (71)$$

As fluctuation calculations and finite differences are both approximations to the "grand truth" of true derivatives and true fluctuations, they will not agree unless they come from the same data source. For direct finite differences, since simulations are taken at different conditions than the simulation used for the fluctuation calculation ( $X + \Delta X$  and  $X - \Delta X$  instead of  $X$ ), the agreement between fluctuations and finite differences is significant. In the case of MBAR finite differences, the

same data can be used to calculate the finite differences and the fluctuations, so these two methods should agree, up to numerical error caused by the finite difference size chosen for MBAR. In this way, MBAR finite differences and fluctuations are a good sanity check, as they should be almost identical if implemented correctly.

## 8 Author Contributions

(Explain the contributions of the different authors here)

For a more detailed description of author contributions, see the GitHub issue tracking and changelog at <https://github.com/ocmadin/best-practices-paper>.

## 9 Other Contributions

(Explain the contributions of any non-author contributors here) For a more detailed description of contributions from the community and others, see the GitHub issue tracking and changelog at <https://github.com/ocmadin/best-practices-paper>.

## 10 Potentially Conflicting Interests

Declare any potentially conflicting interests here, whether or not they pose an actual conflict in your view.

## 11 Funding Information

FMS acknowledges the support of NSF grant CHE-1111111.

## References

- [1] **et al H.** Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *Journal of Chemical Physics*. 2004; 120(20):9665–9678.
- [2] **et al D.** Insights into Protein Compressibility from Molecular Dynamics Simulations. *The Journal of Physical Chemistry B*. 2000; 105(3):715–724.
- [3] **Lustig R.** Direct molecular NVT simulation of the isobaric heat capacity, speed of sound and Joule-Thomson coefficient. *Molecular Simulation*. 2011; 37(6):457–465.
- [4] **Fennell CJ.** Simple liquid models with corrected dielectric constants. *The Journal of Physical Chemistry B*. 2012; 116(23):6936–6944.
- [5] **et al P.** Developing a Predictive Form of MOSCED for Non-electrolyte Solids Using Molecular Simulation: Application to Acetanilide, Acetaminophen, and Phenacetin. *Industrial and Engineering Chemistry Research*. 2016; 55(18):5415–5430.
- [6] **et al P.** Calculating the Fugacity of Pure, Low Volatile Liquids via Molecular Simulation with Application to Acetanilide, Acetaminophen, and Phenacetin. *Industrial and Engineering Chemistry Research*. 2015; 54(36):9027–9037.
- [7] **et al D.** Enthalpies of mixing predicted using molecular dynamics simulations and OPLS force field. *Fluid Phase Equilibria*. 2010; 289(2):156–165.
- [8] **et al W.** Application of Molecular Dynamics Simulations in Molecular Property Prediction. 1. Density and Heat of Vaporization. *Journal of Chemical Theory and Computation*. 2011; 7(7):2151–2165.
- [9] **et al C.** Enthalpies of Vaporization of Organic and Organometallic Compounds, 1880–2002. *Journal of Physical and Chemical Reference Data*. 2003; 32(2):519–878.
- [10] **et al W.** Development of Polarizable Models for Molecular Mechanical Calculations IV: van der Waals parameterization. *Journal of Physical Chemistry B*. 2012; 116(24):7088–7101.