Best Practices for Property Prediction from Molecular Simulations

1 Preliminaries

Definitions

- V: Volume

- U: Total energy (including potential and kinetic, excluding external energy such as due to gravity, etc)

- S: Entopy

- N: Number of particles

- T: Temperature

- P: Pressure

- k_B: Boltzmann constant

- β: $(k_B T)^{-1}$

- M: Molar mass

- ρ: Density (M/V)

- H: Enthalpy

- G: Gibbs Free Energy (free enthalpy)

- A: Helmholtz Free Energy

- μ: Chemical potential

- 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 precisly. For a mole of compound (about 18 mL for water), the relative uncertainty in any of these quantites is about 10^{-12}, far lower than any thermodynamics experiment.

However, in a molecular simulation, these quantites are not necessarily constant. For example, in a NVT simulation, U is allowed to vary. For a long enough simulation (assuming ergodicity, which can pretty much always be assumed with correct simulations and simple fluids), then the ensemble average value of $U = \langle U \rangle$ will be constant, 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 a NPT simulation, however, V is a variable, and we must estimate what the macroscopic value would be with the ensemble estimate $\langle V \rangle$.

The quantities T, P, and μ are *always* constants in both simulation in experiment. There are a number of quantities that can be used to ESTIMATE these constants. 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 estimator of T, and it's average will be equal to the temperature. But it is not the temperature. This quantity fluctuates, but the tempature 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 $\langle V \rangle = \frac{1}{\sum_i w_i} V_i w_i$ 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. For NPT, we then have $f = \beta G$, and for NVT we have $f = \beta A$, while for μVT we have $f = -\beta \langle P \rangle V$.

To calculate expectations at one set of parameters generated with paramters that give rise to a different set of probability distributions, we start with the

definition of an ensemble average given a probability distribution $p_i(x)$.

$$\langle X \rangle_i = \int X(x) p_i(x) dx \tag{1}$$

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

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

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

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

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:

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

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 distibution becomes.

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

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, $u_j(x) = \beta_i U(x) + \beta_i (P_i + \Delta P) V$, or $u_j = \frac{1}{k_B(T_i + \Delta T)} U(x) + \frac{1}{k_B(T_i + \Delta T)} P_i V$. However, the relationship between f and G can be problematic when looking at differences in free energy with respect to temperature, because $G_2 - G_1 = \beta_2 f_2 - \beta_1 f_1$. [MRS: needs to find notes on how this was dealt with last time]

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:

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

And for 2nd deriatives:

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

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

Note that if the finite differences are reevaluated 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.

2 Pure Solvent Properties

2.1 Density

2.1.1 Direct calculation

Starting with the equation used to calculate the density experimentally,

$$\rho = \frac{M}{V} \tag{6}$$

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

$$\rho = \frac{M}{\langle V \rangle} \tag{7}$$

2.1.2 Derivative Estimate

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

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

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

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

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

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

The density can then finally be estimated.

$$\rho \approx \frac{M}{\frac{G_{P+\Delta P} - G_{P-\Delta P}}{2\Delta P}} \tag{11}$$

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

$$\rho \approx \frac{\beta M}{\frac{f_{p+\Delta p} - f_{p-\Delta p}}{2\Delta p}} \tag{12}$$

2.2 Isothermal Compressibility

The definition of isothermal compressibility is:

$$\kappa_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \tag{13}$$

2.2.1 First Derivative

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

$$\kappa_T = -\frac{1}{2V(T,P)^2}\left(\langle V(P+\Delta P, T)\rangle - \langle V(P-\Delta P)\rangle\right) \tag{14}$$

Or by the finite differences evaluation of:

$$\kappa_T = -\frac{\left(\frac{\partial^2 G}{\partial P^2}\right)_{T,N}}{\left(\frac{\partial G}{\partial P}\right)_{T,N}} \tag{15}$$

κ_T can also be estimated from the ensemble average and fluctuation of volume (in the NPT ensemble) or particle number (in the μVT ensemble):[1]

$$\kappa_T = \beta\frac{\langle \Delta V^2\rangle_{NTP}}{\langle V\rangle_{NTP}} = V\beta\frac{\langle \Delta N^2\rangle_{VT}}{\langle N\rangle_{VT}} \tag{16}$$

2.3 Molar Enthalpy

Section on relation of enthalpy to Gibbs free energy (should we need it). The enthalpy, H, can be found from the Gibbs free energy, G, by the Gibbs-Helmholtz relation:

$$H = -T^2 \left(\frac{\partial(\frac{G}{T})}{\partial T} \right)_{P,N} \tag{17}$$

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

$$H = -T^2 \frac{\beta^2}{\beta^2} \left(\frac{\partial(\frac{G}{T})}{\partial T} \frac{\partial T}{\partial \beta} \frac{\partial \beta}{\partial T} \right)_{P,N} \tag{18}$$

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:

$$H = \frac{1}{k_B^3 T^2 \beta^2} \left(\frac{\partial(\frac{G}{T})}{\partial \beta} \right)_{P,N} = \frac{1}{k_B} \left(\frac{\partial(\frac{G}{T})}{\beta} \right)_{P,N} = \frac{\partial f}{\partial \beta}_{T,P,N} \tag{19}$$

Applying the quotient rule to the partial derivative yields

$$H = \frac{T}{k_B} \left(\frac{\partial G}{\partial \beta} \right)_{P,N} - \frac{G}{k_B} \left(\frac{\partial T}{\partial \beta} \right)_{P,N} = \frac{1}{T k_B} \left(\frac{\partial G}{\partial \beta} \right) - \frac{G}{T^2 k_B} \left(\frac{\partial T}{\partial \beta} \right) \tag{20}$$

Recall that $\left(\frac{\partial T}{\partial \beta} \right)_{P,N} = -k_B T^2$, which allows the enthalpy to be simplified to:

$$H = \beta \left(\frac{\partial G}{\partial \beta} \right)_{P,N} + G \tag{21}$$

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

$$\left(\frac{\partial G}{\partial \beta} \right)_{P,N} \approx \frac{G_{\beta + \Delta\beta} - G_{\beta - \Delta\beta}}{2\Delta\beta} \tag{22}$$

2.4 Heat Capacity

The definition of the isobaric heat capacity is:

$$C_P = \left(\frac{\partial H}{\partial T} \right)_{P,N} \tag{23}$$

6

$$C_P = -k_B \beta^3 \left(\frac{\partial^2 G}{\partial \beta^2} \right)_{P,N} \tag{24}$$

Finite difference approach? See section on Molar Enthalpy

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

2.5 Speed of Sound

The definition of the speed of sound is:[2]

$$c^2 = \sqrt{\left(\frac{\partial P}{\partial \rho} \right)_S} = -\frac{V^2}{M} \left(\frac{\partial P}{\partial V} \right)_S \tag{25}$$

$$c^2 = \frac{V^2}{\beta M} \left[\frac{\left(\frac{\gamma_V}{k_B} \right)^2}{\frac{C_V}{k_B}} + \frac{\beta}{V \kappa_T} \right] \tag{26}$$

Where:

$$\gamma_V = \left(\frac{\partial P}{\partial T} \right)_V = \frac{C_V}{T \left(\frac{\partial S}{\partial P} \right)_V} \tag{27}$$

γ_V is known as the isochoric pressure coefficient. κ_T is the same isothermal compressibility from section A.1.3

2.6 Enthalpy of Vaporization

The definition of the enthalpy of vaporization is:[3]

$$\Delta H_{vap} = H_{gas} - H_{liq} = E_{gas} - E_{liq} + P(V_{gas} - V_{liq}) \tag{28}$$

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

$$\Delta H_{vap} = E_{gas,potential} - E_{liq,potential} + RT \tag{29}$$

This is frequently done as a single simulation calculation by assuming the average intramolecular energies 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.

2.6.1 Dielectric Constant

This equation was provided by a literature reference authored by CJ Fennell.[4] Below, $\epsilon(0)$ is the zero frequency dielectric constant, V is the system volume and M is the total system dipole moment.

$$\epsilon(0) = 1 + \frac{4\pi}{3k_B T \langle V \rangle}(\langle M^2 \rangle - \langle M \rangle^2) \tag{30}$$

3 Binary Mixture Properties

3.1 Mass Density, Speed of Sound and Dielectric Constant

The methods for these calculations are the same for a multicomponent system.

3.1.1 Activity Coefficient

The definition of chemical potential in a pure substance is:

$$\mu(T, P) = \left(\frac{\partial G}{\partial N}\right)_{T,P} \tag{31}$$

which is a function of only temperaure and pressure.

Then the definition of the chemical potential μ_i of compound i in a mixture is:

$$\mu_i(T, P, \vec{N}) = \left(\frac{\partial G}{\partial N_i}\right)_{T,P,N_{j\neq i}} \tag{32}$$

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 μ_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 μ_i can be related to the pure chemical potential by

$$\mu_i(T, P, \vec{x}_i) = \mu(T, P) + k_B T \ln(\gamma_i) \tag{33}$$

By analogy to this form, we can

$$\mu_i(T, P, \vec{x}_i) = \mu(T, P) + k_B T \ln(x_i \gamma_i) \tag{34}$$

Where γ_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:

$$\gamma_i = \frac{e^{\left(\frac{\mu_i(T, P, \vec{x}_i) - \mu(T, P)}{k_B T}\right)}}{x_i} \tag{35}$$

Although chemical potentials cannot be directly calculated from simulation, chemical potential differences 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 subleties here relating to the $\ln x_i$ factor, but it appears that with alchemical simulations with a only one particle that is allowed to change, this will cancel out (need to follow up).

3.1.2 Excess Molar Properties

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

$$y^E = y^M - \sum_i x_i y_i \tag{36}$$

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.

3.1.3 Excess Molar Heat Capacity and Volume

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

3.1.4 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 $A.1$ and the generalized method of ecess molar property calculation above or by the following:[5]

$$H^E = \langle E^M \rangle + PV^E - \sum_i x_i \langle E_i \rangle \tag{37}$$

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

References

[1] Dadarlat et al. Insights into protein compressibility from molecular dynamics simulations. *The Journal of Physical Chemistry B*, 105(3):715–724, 12 2000.

[2] Rolf Lustig. Direct molecular nvt simulation of the isobaric heat capacity, speed of sound and joule-thomson coefficient. *Molecular Simulation*, 37(6):457–465, 5 2011.

[3] Wang et al. Application of molecular dynamincs simulations in molecular property prediction. 1. density and heat of vaporization. *Journal of Chemical Theory and Computation*, 7(7):2151–2165, 5 2011.

[4] C.J. Fennell. Simple liquid models with corrected dielectric constants. *The Journal of Physical Chemistry B*, 116(23):6936–6944, 3 2012.

[5] Dai et al. Enthalpies of mixing predicted using molecular dynamics simulations and opls force field. *Fluid Phase Equilibria*, 289(2):156–165, 3 2010.