Keywords: force field; molecular dynamics simulation; parameterization; inference; metamodels

[OUTLINE]: Parameterization of Non-Bonded Classical Mechanics Potentials for Neat Organic Liquids using a Multi-fidelity Bayesian Inference Approach

Bryce C. Manubay[1, *] and Michael R. Shirts[1, †]

[1]*University of Colorado*

(Dated: September 14, 2017)

* bryce.manubay@colorado.edu

† Corresponding author; michael.shirts@colorado.edu

I. 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
- β: $(k_B T)^{-1}$
- M: Molar mass
- ρ: Density (M/V)
- H: Enthalpy
- G: Gibbs Free Energy (free enthalpy)
- A: Helmholtz Free Energy
- u: reduced energy
- f: reduced free energy

II. Introduction

- MD as a critical research tool
  - The development of force fields, which are readily transferable between dissimilar physical systems and are quantitatively accurate, are imperative for the use of molecular simulation driven studies to continue to proliferate.[1–3]
- Transferability issues
  - Transferability of MD force fields, and particularly sets of force field parameters, is an extremely popular topic, and current limitation, in the molecular simulation field.[4–7]
  - Inaccurate and poorly parameterized force fields have been shown to grossly misrepresent molecular systems. [8–10]
  - It has been shown that depending on the choice of force field, the same experiments can produce quantitatively different results, making the choice of force field far more important than it should be. [8, 9, 11–13]
- Parameterization efforts
  - Early
    * Until very recently, force fields have been parameterized manually, guided by the intuition of expert computational chemists.[14–25]
    * Despite attempts at improvement, many of the functional forms and parameters of popular force fields remain mostly unchanged due to the lack of clear, systematic methods for updating them.[26, 27]
    * Many early force fields were parameterized manually for narrow classes of molecules with large redundant parameter spaces.[23]
  - Second Gen
    * The parameterization of GAFF used a semi-automated genetic algorithm approach to select parameters.[28]
    * Force fields like AMBER *parm94* showed intuitive departure by shrinking parameter space with clever atom typing defined by expert computational chemists.[14]
  - Current efforts
    * A few notable attempts, such as GAAMP and ForceBalance, have been made in recent years towards the development of more automated and systematic force field parameterization methods.[29–33]
    * Each made important contributions to automated force field parameterization through clever use of objective function optimization, exploiting a variety of fitting data and allowing exploration of functional forms.
- Bayesian parameterization
  - Previous uses
    * Bayesian inference provides a robust statistical framework for force field parameterization. It has been shown that Bayesian approaches can be applied to a wide variety of data driven sciences. [34–42]
    * Would also include citations for UQ literature and the limited applications to MD parameterization thus far
  - Surrogate models/metamodels
    * Metamodeling has been critical in accelerating sampling driven processes which involve expensive calculations [**?**]
    * Need to include citations for the Bayes inference MD parameterization by the stats guys at ETH
    * Should also find a few more examples
  - What our ideas for parameterization are/paper overall thesis
    * Through systematically testing different multi-fidelity likelihood estimation workflows, we have found an optimal process which maximizes computational efficiency while yielding a force field nearly identical to that which would be produced by a parameterization utilizing solely MD simulation for optimization.

III. Methods

- Simulation protocol
- What parameters?
  - Non-bonded for cyclohexane and ethanol
  - Since we're switching to neat organic parameterization, I'm going to add a few more molecules, but try to keep the number of parameters capped at 10 (chain alkanes, cyclic alcohols, etc.)
  - 10
  - Specific SMIRKS:
    * $[\#8X2H1+0:1],[\#6X4:1],[\#1:1]-[\#6X4],[\#1:1]-[\#8],[\#1:1]-[\#6X4]-[\#7,\#8,\#9,\#16,\#17,\#35]$
- Property calculation
  - Densities Starting with the equation used to calculate the density experimentally,

$$\rho = \frac{M}{V} \tag{1}$$

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

$$\rho = \frac{M}{\langle V \rangle} \tag{2}$$

a. 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:

$$V = \left(\frac{\partial G}{\partial P}\right)_{T,N} \tag{3}$$

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

$$\rho = \frac{M}{\left(\frac{\partial G}{\partial P}\right)_{T,N}} \tag{4}$$

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

$$\left(\frac{\partial G}{\partial P}\right)_{T,N} \approx \frac{G_{P+\Delta P} - G_{P-\Delta P}}{2\Delta p} \tag{5}$$

The density can then finally be estimated.

$$\rho \approx \frac{M}{\frac{G_{P+\Delta P} - G_{P-\Delta P}}{2\Delta P}} \tag{6}$$

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

$$\rho \approx \frac{\beta M}{\frac{f_{P+\Delta P} - f_{P-\Delta P}}{2\Delta P}} \tag{7}$$

1. 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:

$$H = -T^2 \left(\frac{\partial(\frac{G}{T})}{\partial T} \right)_{P,N} \tag{8}$$

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

$$H = -T^2 \frac{\beta^2}{\beta^2} \left(\frac{\partial(\frac{G}{T})}{\partial T} \frac{\partial T}{\partial \beta} \frac{\partial \beta}{\partial T} \right)_{P,N} \tag{9}$$

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:

$$H = \frac{1}{k_B^3 T^2 \beta^2} \left(\frac{\partial(\frac{G}{T})}{\partial \beta} \right)_{P,N} = \frac{1}{k_B} \left(\frac{\partial(\frac{G}{T})}{\beta} \right)_{P,N} = \frac{\partial f}{\partial \beta}_{P,N} \tag{10}$$

2. Heat Capacity

The definition of the isobaric heat capacity is:

$$C_P = \left(\frac{\partial H}{\partial T} \right)_{P,N} \tag{11}$$

$$C_P = \frac{\partial \left(\frac{\partial f}{\partial \beta} \right)}{\partial T}_{P,N} \tag{12}$$

$$C_P = k_B \beta^2 \frac{\partial^2 f}{\partial \beta^2} \tag{13}$$

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.

$$C_P = \frac{\langle H^2 \rangle - \langle H \rangle^2}{N k_B \langle T \rangle^2} \tag{14}$$

- **More properties for verification (not in training data)**
- Had discussed enthalpy of vaporization calculations being used in verification

3. Enthalpy of Vaporization

The definition of the enthalpy of vaporization is:

$$\Delta H_{vap} = H_{gas} - H_{liq} = E_{gas} - E_{liq} + P(V_{gas} - V_{liq}) \tag{15}$$

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

$$\Delta H_{vap} = E_{gas,potential} - E_{liq,potential} + RT \tag{16}$$

A. Suggested Corrections

1. Enthalpy of Vaporization

An alternate, but similar, method for calculating the enthalpy of vaporization is recommended by Horn et al [33].

$$\Delta H_{vap} = -\frac{E_{liq,potential}}{N} + RT - PV_{liq} + C \tag{17}$$

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.

$$\begin{aligned} C_{vib} = C_{vib,intra} + C_{vib,inter} \\ &= (E_{vib,QM,gas,intra} - E_{vib,QM,liq,intra}) \\ &\quad + (E_{vib,QM,liq,inter} - E_{vib,CM,liq,inter}) \end{aligned} \tag{18}$$

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

$$C_{pol} = \frac{N}{2}\frac{(d_{gas} - d_{liq})^2}{\alpha_{p,gas}} \tag{19}$$

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.

$$C_{ni} = P_{vap}\left(B - T\frac{dB}{dT}\right) \tag{20}$$

Where B is the second virial coefficient.

$$C_x = \int_{P_{ext}}^{P_{vap}} \left[V\left(P_{ext}\right)\left[1 - \left(P - P_{ext}\right)\kappa_T\right] - TV\alpha\right] dP \tag{21}$$

Where P_{ext} is the external pressure and $V\left(P_{ext}\right)$ 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.

- Methods for metamodeling
  - MBAR
  - Surrogate models
    * Physically motivated models
      · Density

$$\frac{D_{new}}{D_{old}} = \frac{\sum_i \sum_{j \neq i} N_i N_j \left(\sigma_i \sigma_j\right)^{\frac{3}{2}}{}_{new}}{\sum_i \sum_{j \neq i} N_i N_j \left(\sigma_i \sigma_j\right)^{\frac{3}{2}}{}_{old}} \tag{22}$$

      · Enthalpy

$$\frac{H_{new}}{H_{old}} = \frac{\sum_i \sum_{j \neq i} N_i N_j \left(\epsilon_i \epsilon_j\right)^{\frac{1}{2}}{}_{new}}{\sum_i \sum_{j \neq i} N_i N_j \left(\epsilon_i \epsilon_j\right)^{\frac{1}{2}}{}_{old}} \tag{23}$$

    * Response surface polynomials

$$y = f\left(x\right)\beta + E \tag{24}$$

      · Questions/concerns for these models:
        · How many data points to fit?
        · Over what span of parameter space?
        · Will need to be fitting over all parameters of mixture
          · Depending on order of model, given like 10 force field parameters, could have A LOT of model parameters
          · Not sure if this method is viable
    * GP models
      · Formalism for estimating some quantity Z at unknown location x_0 ($Z\left(x_0\right)$) from N pairs of observed values $w_i\left(x_0\right)$ and $Z\left(x_i\right)$ where $i = 1, ..., N$
      ·

$$\hat{Z}\left(x_0\right) = \sum_{i=1}^{N} w_i\left(x_0\right) \times Z\left(x_i\right) \tag{25}$$

      · We find our weight matrix, **W**, by minimizing **W** subject to the following system of equations:
        ·

$$\underset{W}{\text{minimize}} \qquad W^T \cdot \mathrm{Var}_{x_i} \cdot W - \mathrm{Cov}^T_{x_i x_0} \cdot W - W^T \cdot \mathrm{Cov}_{x_i x_0} + \mathrm{Var}_{x_0} \tag{26}$$

$$\text{subject to} \qquad \mathbf{1}^T \cdot W = 1 \tag{27}$$

      · where the literals

$$\left\{\mathrm{Var}_{x_i}, \mathrm{Var}_{x_0}, \mathrm{Cov}_{x_i x_0}\right\} \tag{28}$$

        stand for

$$\left\{\mathrm{Var}\left(\left[Z(x_1) \ \cdots \ Z(x_N)\right]^T\right), \mathrm{Var}(Z(x_0)), \mathrm{Cov}\left(\left[Z(x_1) \ \cdots \ Z(x_N)\right]^T, Z(x_0)\right)\right\} \tag{29}$$

      · The weights summarize important procedures of the inference process:
        · They reflect the structural closeness of samples to the estimation location, x_0
        · They have a desegregating effect, to avoid bias caused by sample clustering
  - Explanation of potential multi-fidelity posterior sampling algorithm
    * 3 Levels of property calculation
      · High fidelity: Full MD simulation at a single point in parameter space

        - Medium fidelity: Use MBAR to estimate properties over a conservative range of parameter space in order to create a hypervolume of data over which we can construct a model
        - Low fidelity: Use data from medium fidelity calculations in order to fit a regression model over a hypervolume of parameter space
            - For right now, most plausible technique is GP regression, but could brainstorm some others
    * Simple physical surrogate models (2 levels of property estimation)
        - High fidelity: Full MD simulation at a single point in parameter space
        - Low fidelity: Using simple, physically motivated models from previous section we can estimate properties at new parameter states solely based on how we expect the change in parameters to affect the potential energy
    * Using MBAR as a look up table (2 levels of property estimation)
        - High fidelity: Full MD simulation at a single point in parameter space
        - Medium fidelity: Use MBAR to estimate properties over a conservative range of parameter space in order to create a hypervolume of data
            - Rather than attempting to fit a model we can use MBAR discrete MBAR calculated observables in order to evaluate our likelihood
            - Using a discrete set of calculations, we can iterate over those values in order to find the highest point of probability and then perform a new simulation and repeat the process
            - As we narrow in on the region of highest probability density, we can refine the grid over which we're searching in order to more accurately represent the final posterior
- Construction of "gold standard" posterior distribution of force field parameters
    * Will need to brainstorm best way to go about doing this
    * In theory, we could construct a posterior with purely simulation, but it would take so much resource time
    * Might be a better idea to compare differences in the cheap posteriors we produce

IV. Experiments + Results and Analyses

- Hypothesis: Using a multi-fidelity likelihood calculation scheme described in the previous section will provide not only a substantial speed up over a traditional "vanilla" Bayesian inference approach with purely simulation used in the likelihood estimate, but will also allow for accuracy in the final force field rivaling that of the expensive approach.
- Experiments for testing sampling workflow
    - Interchanging surrogate models
        * All combinations of response surface models?
    - Do we include MBAR?
        * Use MBAR like a look up table. Don't fit points produced with MBAR. Just MC sample on a grid of points produced with MBAR to see where the higher probability region is. If on edge, move to edge and simulate again, repeat.
        * Note: Physical surrogates won't make use of MBAR since there won't be fitting data
- Ideas for comparing sampling methods
    - KL divergence
    - Speed of convergence
    - Simulation of properties, using final parameters, that were not in training set
        * 3-fold verification
            - Different properties not in training set
            - Extrapolation to thermodynamic state outside of training set (T, P)
            - Other molecules outside of training set that have the same SMIRKS types

[1] Jayachandran, G.; Vishal, V.; Pande, V. S. *J Chem Phys* **2006**, *124*, 164902.
[2] Beauchamp, K. A.; Ensign, D. L.; Das, R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12734–12739.
[3] De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. *J. Med. Chem.* **2016**, *59*, 4035–4061.
[4] Vellore, N. A.; Yancey, J. A.; Collier, G.; Latour, R. A.; Stuart, S. J. *Langmuir* **2010**, *26*, 7396–7404.
[5] Puleo, D. A.; Bizios, R. *Biological Interactions on Materials Surfaces: Understanding and Controlling Protein, Cell, and Tissue Responses*; Springer Science & Business Media, 2009; pp 77–78.
[6] Sato, F.; Hojo, S.; Sun, H. *J. Phys. Chem. A* **2003**, *107*, 248–257.
[7] Martin-Calvo, A.; Gutiérrez-Sevillano, J. J.; Parra, J. B.; Ania, C. O.; Calero, S. *Phys Chem Chem Phys* **2015**, *17*, 24048–24055.
[8] Lange, O. F.; van der Spoel, D.; de Groot, B. L. *Biophys J* **2010**, *99*, 647–655.
[9] Martín-García, F.; Papaleo, E.; Gomez-Puertas, P.; Boomsma, W.; Lindorff-Larsen, K. *PLoS One* **2015**, *10*.
[10] Vanommeslaeghe, K.; Yang, M.; MacKerell, A. D. *J. Comput. Chem.* **2015**, *36*, 1083–1101.
[11] Ewen, J. P.; Gattinoni, C.; Thakkar, F. M.; Morgan, N.; Spikes, H. A.; Dini, D. *Materials* **2016**, *9*, 651.
[12] Petrov, D.; Zagrovic, B. *PLOS Computational Biology* **2014**, *10*, e1003638.
[13] Guvench, O.; MacKerell, A. D. *Methods Mol. Biol.* **2008**, *443*, 63–88.
[14] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
[15] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D. *The Journal of Chemical Physics* **1983**, *79*, 926–935.
[16] Burger, S. K.; Cisneros, G. A. *J Comput Chem* **2013**, *34*, 2313–2319.
[17] Law, M. M.; Hutson, J. M. *Computer Physics Communications* **1997**, *102*, 252 – 268.
[18] Chen, I. J.; Yin, D.; MacKerell, A. D. *J Comput Chem* **2002**, *23*, 199–213.
[19] Horinek, D.; Mamatkulov, S. I.; Netz, R. R. *The Journal of Chemical Physics* **2009**, *130*, 124507.
[20] Hernandes, M. Z.; Longo, R. L. *J Mol Model* **2005**, *11*, 61–68.
[21] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
[22] MacKerell, A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
[23] Allinger, N. L.; Tribble, M. T.; Miller, M. A.; Wertz, D. H. *J. Am. Chem. Soc.* **1971**, *93*, 1637–1648.
[24] Soo, G. C.; Cartledge, F. K.; J. Unwalla, R.; Profeta, S. *Tetrahedron* **1990**, *46*, 8005–8018.
[25] Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.
[26] Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.
[27] Monticelli, L.; Tieleman, D. P. *Methods Mol. Biol.* **2013**, *924*, 197–213.
[28] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
[29] Huang, L.; Roux, B. *J. Chem. Theory Comput.* **2013**, *9*, 3543–3556.
[30] Wang, L.-P.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.
[31] Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. B* **2013**, *117*, 9956–9972.
[32] Wang, L.-P.; Chen, J.; Van Voorhis, T. *J. Chem. Theory Comput.* **2013**, *9*, 452–460.
[33] Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *The Journal of Chemical Physics* **2004**, *120*, 9665–9678.
[34] Farrell, K.; Oden, J. T.; Faghihi, D. *Journal of Computational Physics* **2015**, *295*, 189–208.
[35] Klein, K.; Hennig, S.; Paul, S. K. *PLOS ONE* **2016**, *11*, e0152700.
[36] Wu, S.; Angelikopoulos, P.; Papadimitriou, C.; Moser, R.; Koumoutsakos, P. *Phil. Trans. R. Soc. A* **2016**, *374*, 20150032.
[37] Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. *The Journal of Chemical Physics* **2012**, *137*, 144103.
[38] Zhu, J.; Chen, J.; Hu, W.; Zhang, B. *arXiv:1411.6370 [cs, stat]* **2014**, arXiv: 1411.6370.
[39] Cailliez, F.; Bourasseau, A.; Pernot, P. *J. Comput. Chem.* **2014**, *35*, 130–149.
[40] Liu, J. S. *Monte Carlo Strategies in Scientific Computing*; Springer, 2001.
[41] Box, G. E.; Tiao, G. C. *Bayesian Inference in Statistical Analysis*; John Wiley Sons, Inc., 1992.
[42] Patrone, P. N.; Rosch, T. W.; Jr, F. R. P. *The Journal of Chemical Physics* **2016**, *144*, 154101.