*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 27, 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
- $\beta$: $(k_B T)^{-1}$
- $M$: Molar mass
- $\rho$: 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
  - Force fields that are transferable and quantitatively accurate are necessary for molecular simulation to be useful. [1–3]
- Transferability and inaccuracy issues
  - Transferability of MD force fields, and particularly sets of force field parameters, is a 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 for the same or similar systems 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]
    - Force fields like AMBER *parm94* showed intuitive departure by shrinking parameter space with clever atom typing defined by expert computational chemists.[14]
  - Second Gen
    - The parameterization of GAFF used a semi-automated genetic algorithm approach to select parameters.[28]
    - The parameterization of the rigid Tip4p-Ew model utilized a unique gradient assisted method. [33]
    - An incredible amount of work over a long period of time has still been necessary to get biomolecular force fields somewhat correct.[**?** ]
  - 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.[29–32]
    - 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. [34–36, 38–42]
    - Bayesian inference methods have also been applied for uncertainty quantification in MD as well as limited parameterization problems on simple Lennard-Jones systems. [37**? ?** ]
  - 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 [**?** ]
    - COFFE papers [**?** ]
  - 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.**
    - How can we combine different techniques, to find reasonable force fields in medium dimensionality in a computationally efficient manner

## 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 $\beta$ instead of $T$ yields:

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

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. Speed of Sound*

The definition of the speed of sound is[**?** ]:

$$c^2 = \left( \frac{\partial P}{\partial \rho} \right)_S = -\frac{V^2}{M} \left( \frac{\partial P}{\partial V} \right)_S \tag{11}$$

$$c^2 = \frac{V^2}{\beta M} \left[ \frac{\left( \frac{\gamma_V}{k_B} \right)^2}{\frac{C_V}{k_B}} + \frac{\beta}{V \kappa_T} \right] \tag{12}$$

Where:

$$\gamma_V = \left( \frac{\partial P}{\partial T} \right)_V \tag{13}$$

$\gamma_V$ 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.

$$\left( \frac{\partial P}{\partial V} \right)_S = \frac{\left(\frac{\partial S}{\partial V}\right)_P}{\left(\frac{\partial S}{\partial P}\right)_V} \tag{14}$$

$$\left( \frac{\partial S}{\partial V} \right)_P = \left( \frac{\partial S}{\partial T} \right)_P \left( \frac{\partial T}{\partial V} \right)_P = \frac{C_P}{T} \left( \frac{\partial T}{\partial V} \right)_P = \frac{C_P}{TV\alpha} \tag{15}$$

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:

$$\left(\frac{\partial S}{\partial P}\right)_V = \left(\frac{\partial S}{\partial T}\right)_V \left(\frac{\partial T}{\partial P}\right)_V = \frac{C_V}{T}\left(\frac{\partial T}{\partial P}\right)_V = \frac{C_V}{T\gamma_V} \tag{16}$$

Thus our derivation yields:

$$\left(\frac{\partial P}{\partial V}\right)_S = \frac{C_P \gamma_V}{C_V V \alpha} \tag{17}$$

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

*a. Analytical derivative of density with respect to temperature*

$$\alpha = -\frac{d\ln\langle\rho\rangle}{dT} \tag{18}$$

*b. Numerical derivative of density over range of T of interest* The same finite differnces approach as shown for isothermal compressibility can be applied here, thus:

$$\alpha = -\frac{d\ln\langle\rho\rangle}{dT} = - \frac{1}{2\rho(T,P)}\left(\ln\langle\rho(P,T+\Delta T)\rangle - \ln\langle V(P,T-\Delta T)\rangle\right) \tag{19}$$

*c. Using the enthalpy-volume fluctuation formula*

$$\alpha = \frac{\langle VH\rangle - \langle V\rangle\langle H\rangle}{k_B\langle T\rangle^2\langle V\rangle} \tag{20}$$

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

$$\gamma_V = -\frac{\alpha}{\kappa_T} \tag{21}$$

- 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

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

* 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 $i = 1, ..., N$
  .

$$\hat{Z}\left(x_0\right) = \sum_{i=1}^{N} w_i\left(x_0\right) \times Z\left(x_i\right) \tag{24}$$

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

$$\underset{W}{\text{minimize}} \qquad W^T \cdot \mathrm{Var}_{x_i} \cdot W - \mathrm{Cov}_{x_i x_0}^T \cdot W - W^T \cdot \mathrm{Cov}_{x_i x_0} + \mathrm{Var}_{x_0} \tag{25}$$

$$\text{subject to} \qquad \mathbf{1}^T \cdot W = 1 \tag{26}$$

  - where the literals

$$\{\mathrm{Var}_{x_i}, \mathrm{Var}_{x_0}, \mathrm{Cov}_{x_i x_0}\} \tag{27}$$

    stand for

$$\left\{\mathrm{Var}\left(\left[Z(x_1) \ \cdots \ Z(x_N)\right]^T\right), \mathrm{Var}(Z(x_0)), \mathrm{Cov}\left(\left[Z(x_1) \ \cdots \ Z(x_N)\right]^T, Z(x_0)\right)\right\} \tag{28}$$

  - 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

– **Hypothesis: With a multi-fidelity hierarchical observable calculation scheme, we can quickly approach the true forward model produced by MD simulation**

– Explanation of potential multi-fidelity posterior sampling algorithms:

* 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

* 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

## 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 inference approach with purely simulation used in the likelihood estimate, but will also result in a final force field with accuracy rivaling that of the expensive approach.**
- Experiments for testing sampling workflow
- **Hypothesis: Different sampling algorithms will affect the speed of convergence to our final force field as well final distribution of parameters sampled.**
- Ideas for comparing force fields resultant from different sampling methods
  - KL divergence
  - Speed of convergence
  - Stability of solution (from different starting points do we get same answer?)
  - 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.