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: June 12, 2018)

* 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: $(k_BT)^{-1}$
- 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
  - 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.[14–24]
    - 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.[25, 26]
    - Force fields like AMBER *parm94* showed intuitive departure by shrinking parameter space with clever atom type defined by expert computational chemists.[14]
  - Second Gen
    - The parameterization of GAFF used a semi-automated genetic algorithm approach to select parameters.[27]
    - The parameterization of the rigid Tip4p-Ew model utilized a unique gradient assisted method. [28]
    - An incredible amount of work over a long period of time has still been necessary to get biomolecular force fields somewhat correct.[29]
  - 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.[30–33]
    - 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. [34–41]
    - Bayesian inference methods have also been applied for uncertainty quantification in MD as well as limited parameterization problems on simple Lennard-Jones systems. [42–44]
  - Surrogate models/metamodels
    - Parameterization on purely expensive simulator calculations (either MD or QM) is extremely expensive
    - Metamodeling has been critical in accelerating sampling driven processes which involve expensive calculations [45]
    - Some previous work has utilized efficient metamodels to accelerate Bayesian inference driven parameterization of LJ models with mixed results. Statistical and modeling methods were great, but the physical intuition to constrain the problem correctly was lacking. [44]
    - COFFE papers –> innovative metamodeling and sparse grid optimization techniques [46]
  - 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 consistent with the experimental data it was trained on.**
        - How can we combine different techniques, to find reasonable force fields in medium dimensionality in a computationally efficient manner
    - Additional ideas for motivating what we're doing:
        - How many person-years does it typically take to create new force fields and how much of that is limited by the expense of the optimization process?

III. Methods

- Simulation protocol
- What parameters?
    - Non-bonded for cyclohexane and ethanol
    - 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
    - For this paper we will be optimizing our parameters on two thermophysical properties; molar volume, $\hat{V}$, and heat of vaporization, ΔH_{vap}. This section details the methods we will use to calculate each.
        - Molar Volume
          System volume, V, can easily be calculated as:

$$V = x \times y \times z \tag{1}$$

where x, y and z are the edge lengths of the simulation. This can be converted to a molar volume by dividing by the number of moles in the periodic box. We can write molar volume as:

$$\hat{V} = \frac{V}{N_{mol}} = \frac{V \times N_{Av}}{N_{part}} \tag{2}$$

Where N_{mol} are the number of moles per box, N_{part} are the number of particles per box and N_{Av} is Avogadro's number.

1. *Heat of Vaporization*

The definition of the enthalpy of vaporization is:

$$\Delta H_{vap} = H_{gas} - H_{liq} = E_{gas} - E_{liq} + P(V_{gas} - V_{liq}) \tag{3}$$

The uncertainty in this calculation can be computed by bootstrapping or analytical estimation using MBAR. We will compare both results in order to determine whether the cheaper analytical estimate is accurate enough to be used.

- Methods for metamodeling
    - MBAR
    - Surrogate models
        - 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 $i = 1, ..., N$
            -

$$\hat{Z}\left(x_0\right) = \sum_{i=1}^{N} w_i\left(x_0\right) \times Z\left(x_i\right) \tag{4}$$

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

$$\begin{aligned} \underset{W}{\text{minimize}} \quad & W^T \cdot \mathrm{Var}_{x_i} \cdot W - \mathrm{Cov}^T_{x_i x_0} \cdot W - W^T \cdot \mathrm{Cov}_{x_i x_0} + \mathrm{Var}_{x_0} && (5) \\ \text{subject to} \quad & \mathbf{1}^T \cdot W = 1 && (6) \end{aligned}$$

                - where the literals

$$\left\{\mathrm{Var}_{x_i}, \mathrm{Var}_{x_0}, \mathrm{Cov}_{x_i x_0}\right\} \tag{7}$$

stand for

$$\left\{\mathrm{Var}\left(\left[Z(x_1) \ \ \cdots \ \ Z(x_N)\right]^T\right), \mathrm{Var}(Z(x_0)), \mathrm{Cov}\left(\left[Z(x_1) \ \ \cdots \ \ Z(x_N)\right]^T, Z(x_0)\right)\right\} \tag{8}$$

            - 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
            - Formalizing expressions for my specific problem (which variables are which and what exactly do the covariances refer to).
    - **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
        * Sampling from our posterior using GP regression will rapidly lead us to a new optima in the local parameter space that we have modeled.
        * We can directly sample configurations at this new optima using MD and repeat the hierarchical property calculation and sampling using GP models in order to "learn" the functional form of the properties across parameter space.
        * We arrive in the high probability region of our parameters cheaply and can finish sampling using the high fidelity calculation method.

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
- Results from 2D experiments with $[\#6X4:1]$
    - Stability of solution (multiple starting points)

        - Maybe also consider does starting out of liquid phase break the process
    - How similar do all of the final posteriors end up being
    - Simulation of properties using final distribution of parameters
    - How does sparsity of MBAR calculations in GP models affect final posterior?
        - Start with gradient information (2 $\times$ dimensionality + 1) and move up from there.
        - What is the most dense the calculations should be and how many factor levels should be tested (this may require some trial-and-error to sort)?
    - Checking across 3 general small molecule force fields (*GAFF2*, *parm99* and *CGenFF*) the standard deviation in ϵ was 0.0034 and in r$_{min,half}$ was 0.0638, which are about 21.7 and 4.3 % of the *smirnoff99Frosst* values, respectively.
    - Given standard deviations of parameters from other general organic forcefields, the stability tests will be, at max, 20 % change in ϵ and 5 % in r$_{min,half}$.

- Same results from 4D experiments with $[\#1:1]-[\#6X4]$ added
    - Checking across 3 general small molecule force fields (*GAFF2*, *parm99* and *CGenFF*) the standard deviation in ϵ was 0.0144 and in $r_{min,half}$ was 0.0672, which are about 13.2 and 3.5 % of the *smirnoff99Frosst* values, respectively.
    - Given standard deviations of parameters from other general organic forcefields, the stability tests will be, at max, 15 % change in ϵ and 5 % in r$_{min,half}$.

- Scale up to all 5 SMIRKS (10-D parameter space)
    - Similar variance analyses for ϵ and r$_{min,half}$ of $[\#8X2H1+0:1]$, $[\#1:1]-[\#8]$, $[\#1:1]-[\#6X4]-[\#7,\#8,\#9,\#16,\#17,\#35]$ yield the following:
        - $[\#8X2H1+0:1]$: $Var\left(\epsilon\right)^{0.5}=0.0516$ and $Var\left(r_{min,half}\right)^{0.5}=0.0405$, which are about 24.5 and 2.4 % of the *smirnoff99Frosst* values, respectively.
        - $[\#1:1]-[\#8]$: $Var\left(\epsilon\right)^{0.5}=0.1202$ and $Var\left(r_{min,half}\right)^{0.5}=0.1047$, which are about 40.1 and 20000 % of the *smirnoff99Frosst* values, respectively. Given that these values are pretty ridiculous, I'll keep with the trend of 20 % changes in ϵ and 5 % changes in r$_{min,half}$.
        - $[\#1:1]-[\#6X4]-[\#7,\#8,\#9,\#16,\#17,\#35]$: $Var\left(\epsilon\right)^{0.5}=0.00505$ and $Var\left(r_{min,half}\right)^{0.5}=0.0275$, which are about 32.1 and 2.0 % of the *smirnoff99Frosst* values, respectively.
    - Given results from variance analyses over all 5 SMIRKS/atom types, I think maximum changes of 20 % of the ϵ and 5 % of the r$_{min,half}$ in *smirnoff99Frosst* would be appropriate for the stability tests.

- Ideas for comparing force fields (different start points)
    - KL divergence/some other probability distribution convergence test
        - Stability of solution (from different starting points do we get same answer?)
        - Maximum likelihood calculation of GP
            - More detail. Is this possible?
        - Discrete KL divergence
        - Maximum Mean Discrepancy estimators
        - Some theory: http://papers.nips.cc/paper/3417-estimation-of-information-theoretic-measures-for-continuous-random-variables
        - Practically: https://github.com/gregversteeg/NPEET
    - Speed of convergence
        - Number of iterations to convergence
        - Total wall time
    - 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] Burger, S. K.; Cisneros, G. A. *J Comput Chem* **2013**, *34*, 2313–2319.

[16] Law, M. M.; Hutson, J. M. *Computer Physics Communications* **1997**, *102*, 252 – 268.

[17] Chen, I. J.; Yin, D.; MacKerell, A. D. *J Comput Chem* **2002**, *23*, 199–213.

[18] Horinek, D.; Mamatkulov, S. I.; Netz, R. R. *The Journal of Chemical Physics* **2009**, *130*, 124507.

[19] Hernandes, M. Z.; Longo, R. L. *J Mol Model* **2005**, *11*, 61–68.

[20] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

[21] MacKerell, A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[22] Allinger, N. L.; Tribble, M. T.; Miller, M. A.; Wertz, D. H. *J. Am. Chem. Soc.* **1971**, *93*, 1637–1648.

[23] Soo, G. C.; Cartledge, F. K.; J. Unwalla, R.; Profeta, S. *Tetrahedron* **1990**, *46*, 8005–8018.

[24] Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490–519.

[25] Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.

[26] Monticelli, L.; Tieleman, D. P. *Methods Mol. Biol.* **2013**, *924*, 197–213.

[27] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[28] 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.

[29] Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

[30] Huang, L.; Roux, B. *J. Chem. Theory Comput.* **2013**, *9*, 3543–3556.

[31] Wang, L.-P.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.

[32] 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.

[33] Wang, L.-P.; Chen, J.; Van Voorhis, T. *J. Chem. Theory Comput.* **2013**, *9*, 452–460.

[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] Zhu, J.; Chen, J.; Hu, W.; Zhang, B. *arXiv:1411.6370 [cs, stat]* **2014**, arXiv: 1411.6370.

[38] Cailliez, F.; Bourasseau, A.; Pernot, P. *J. Comput. Chem.* **2014**, *35*, 130–149.

[39] Liu, J. S. *Monte Carlo Strategies in Scientific Computing*; Springer, 2001.

[40] Box, G. E.; Tiao, G. C. **1992**,

[41] Patrone, P. N.; Rosch, T. W.; Jr, F. R. P. *The Journal of Chemical Physics* **2016**, *144*, 154101.

[42] Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. *The Journal of Chemical Physics* **2012**, *137*, 144103.

[43] Rizzi, F.; Najm, H.; Debusschere, B.; Sargsyan, K.; Salloum, M.; Adalsteinsson, H.; Knio, O. *Multiscale Model. Simul.* **2012**, *10*, 1428–1459.

[44] Kulakova, L.; Arampatzis, G.; Angelikopoulos, P.; Chatzidoukas, P.; Papadimitriou, C.; Koumoutsakos, P. *arXiv:1705.08533 [physics, stat]* **2017**, arXiv: 1705.08533.

[45] Shirts, M. R.; Chodera, J. D. *J Chem Phys* **2008**, *129*.

[46] Hülsmann, M.; Kirschner, K. N.; Krämer, A.; Heinrich, D. D.; Krämer-Fuhrmann, O.; Reith, D. In *Foundations of Molecular Modeling and Simulation: Select Papers from FOMMS 2015*; Snurr, R. Q., Adjiman, C. S., Kofke, D. A., Eds.; Springer Singapore: Singapore, 2016; pp 53–77.