Outline: Bayesian sampling driven classical mechanics force field parameterization

Bryce C. Manubay

May 17, 2017

I. Objective: Test and validate use of Bayesian sampling driven parameterization procedure using inexpensively generated single molecule simulation data as experimental evidence

II. Intro:

   A. Classical mechanics simulations have been useful in the study of chemistry, biology and materials ranging from the simple to very complex. However the force fields that scientists use in their studies aren't always consistent and an element of uncertainty to any study done.

      1. Produce quantitatively different results depending on the force field [1, 2, 3, 4, 5]
      2. Can make choice of force field more important than it should be

   B. Current force field parameterization efforts are heuristic and often guided by the physical intuition of scientists rather than systematically. [6, 7, 8, 9, 10, 11, 12, 13]

      1. Go through cited examples explicitly
      2. ForceBalance: tries to be somewhat more quantitative by minimizing an objective function. However, weights are still chosen by hand. [more discussion][14, 15, 16]
      3. Find more examples

   C. Big reason behind this being how force fields are optimized to reproduce very specific types of "target data"[17, 18]

      1. There is no single right way to optimize a force field to reproduce all kinds of observables

    a. Usually, when you optimize for certain observables, estimates of other observables can become less accurate
2. Additionally, it is difficult to update existing force fields with new kinds of data (either more molecular diversity or new properties) without changing the entire force field
    a. Forcefield parameters are inherently coupled by nature of how they are designed (sometimes weakly, sometimes strongly)
    b. Chemically/physically, interactions and geometries at an atomic scale are coupled
    c. You often need to adjust all (or many) parameters if you adjust one
    d. Could be a lot, could be a little. Situations vary depending on how the changing interactions affect the overall system
3. **We would like to probabilistically determine all FAMILIES of force fields consistent with given data in a systematic manner. Also, given more data (new data), we would like a method to update extant force fields, so that new force fields could be determined to given this new data.**

D. Bayesian statistics has been applied to a large number of big data and optimization problems [19, 20, 21, 22, 23, 24, 25, 26]
1. add citations from the recent lit stuff on OpenFF Slack

E. With this in mind the authors of this paper have developed a novel process for parameterizing classical mechanics force fields using experimental data as evidence for Bayesian inference driven parameterization. For this particular paper we have set up a toy case in order to test and validate the Bayesian inference parameterization.
1. The experimental data used as evidence will be trajectory data produced from a force field developed by members of our force field parameterization team.[27]
2. For simplicity and time considerations the simulated data is only of single molecules and hence the parameters being changed will be limited to those involved in bonded interactions. Specifically:
    a. bonded force constants
    b. equilibrium bond lengths
    c. angular force constants
    d. equilibrium bond angles
    e. torsional force constants
3. We will investigate if the Bayesian inference approach will recover the original force field parameters using the simulated data under the original force field as evidence if the force field is perturbed.

III. Methods

A. Molecules?

1. ≤ 3 carbons
2. ≤ 2 oxygens

Index	SMILES	C_count	O_count	AlkEthOH_id	IUPAC_names
0	C	1	1	AlkEthOH_c0	methane
15	CC	2	1	AlkEthOH_c38	ETHANE
339	CC(C)O	3	2	AlkEthOH_c488	Propan-2-ol
603	CCC	3	1	AlkEthOH_c901	PROPANE
789	CCCO	3	2	AlkEthOH_c11…	Propan-1-ol
804	CCO	2	2	AlkEthOH_c11…	ethanol
805	CCOC	3	2	AlkEthOH_c11…	Methoxyethane
896	CO	1	2	AlkEthOH_c12…	methanol
897	COC	2	2	AlkEthOH_c12…	Methoxymetha…
912	O	0	2	AlkEthOH_c13…	oxidane
105	C1COC1	3	2	AlkEthOH_r131	Oxetane

Figure 1: The molecules being used as the test set for this initial parameterization. Each row entry includes the SMILES string, C and O composition, ID from the AlkEthOH set and the IUPAC name for a given molecule.

B. Simulations
    1. Generate simulation data by simulating a set of AlkEthOH molecules using the 'smirff99Frosst' forcefield
    2. Simulation parameters:
        a. Thermostated to 300 K
        b. 0.5 ns time steps
        c. Friction coefficient of 1 ps^{-1}
        d. 4 ns simulations
        e. Frame recored every 1000 steps
    3. Currently making O-H LJ parameters small and finite in order to keep hydrogens from floating into other atoms

C. Sampling the Bayesian Posterior
    1. Observables:
        a. Bonds and Angles:
            (1) It is well known that simulated bond and angle frequency distributions can be well described by simple gaussian distributions
            (2) We therefore choose observables to be simple parameters of the gaussian distribution
                (a) Mean bond lengths and angles

(b) Variance of bond lengths and angles

b. Torsions:

(1) Torsion distributions are generally more difficult to describe than bonds and angles

(2) Additionally, distributions are often multimodal and may not have constant period between modes

(3) Hence, we choose to describe torsion distributions as fourier series expansions of the form:

(a) $f_N(x) = \sum_{i=0}^{N} A_i \sin^2\left(\frac{2i\pi x}{\tau_i} + \psi_i\right)$

(b) Fourier series coefficients, phase angles, and periodicities will be the observables used to describe the distributions

2. Bayes' Theorem

a. Simple version: $Pr(\Theta|O) = \frac{Pr(O|\Theta)Pr(\Theta)}{Pr(B)}$

b. Where $Pr()$ is a probability function and Θ are parameters for a model estimating our observed data O

3. Prior probability models

a. Simple uniform priors for all parameters

b. Limits on uniforms dependent on parameter

(1) Equilibriun bond lengths and angles: $\pm 20\%$ of the true parameter value

(2) Force constants: 0 as the floor and twice the highest value in the force field as a ceiling

(a) Bonded force constant: 0 to 4000 $\left(\frac{kcal}{mol \cdot ^2}\right)$

(b) Angular force constant: 0 to 1400 $\left(\frac{kcal}{mol \cdot rad^2}\right)$

(c) Torsional force constant: 0 to 21 $\left(\frac{kcal}{mol}\right)$

4. Liklihood function

a. Forward models have already been determined and are simple (how do we calculate bond lengths and angles from a time series of coordinates?)

(1) Bond Length

(a) $r_{i,j} = \sqrt{(\bar{x}_i - \bar{x}_j)^2}$

(b) where $r_{i,j}$ is the bond length between atoms i and j and $\bar{x}_i$ are the atomic coordinates of atom i

(2) Bond Angle

(a) $\dot{x}_{i,j,k} = (\bar{x}_i - \bar{x}_j) \cdot (\bar{x}_j - \bar{x}_k)$

(b) $\theta_{i,j,k} = \arccos \frac{\dot{x}_{i,j,k}}{r_{i,j} \cdot r_{j,k}}$

(c) where $\theta_{i,j,k}$ is the angle between bonded atoms i,j and k

(3) Torsion Angle

(a) $\hat{x}_{i,j} = \frac{(\bar{x}_i - \bar{x}_j)}{\sqrt{(\bar{x}_i - \bar{x}_j)^2}}$

(b) $n_{i,j,k} = \hat{x}_{i,j} \times \hat{x}_{j,k}$
(c) $m = n_k \times \hat{x}_{i,j}$
(d) $y = n_{i,j,k} \cdot n_{j,k,l}$
(e) $z = m \cdot n_{j,k,l}$
(f) $\phi_{i,j,k,l} = \arctan \frac{y}{z}$
(g) where $\phi_{i,j,k,l}$ is the torsion angle formed by atoms i,j,k and l

b. There are a few possible options for choices of likelihood functions

(1) $L\left(O|\Theta\right) = \frac{1}{\sqrt{2\pi\sigma_O^2}} \cdot \exp\left(-\frac{\left(\bar{O}-O_{mod}(\Theta)\right)^2}{2\sigma_O^2}\right)$
(2) $L\left(O|\Theta\right) = \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi\sigma_O^2}} \cdot \exp\left(-\frac{(O_j-O_{mod}(\Theta))^2}{2\sigma_O^2}\right)$
(3) $L\left(O|\Theta\right) = \exp\left(-\frac{\frac{1}{M}\sum_{j=1}^{M}(O_{mod}(\Theta)-O_j)^2}{2\sigma_O^2}\right)$

5. Sampling and choice of surrogate model
   a. What is a surrogate model?
      (1) Surrogate models allow us to inexpensively determine outcomes that cannot be easily measured directly
      (2) In the context of our force field parameterization, we use surrogate models to estimate simulated observables cheaply in order to inexpensively sample our posterior
      (3) Surrogate models can allow us to accelerate posterior sampling and by updating the model frequently (as more data is available) provide an accurate final answer
      (4) **Currently looking up best practices for making accurate surrogate models with sparse high dimensional data
      (5) Will use lowest level of complexity possible here (like splining for example)
6. Multistate reweighting using MBAR
   a. Not too sure how useful this is going to be if we can use surrogate modeling to help assess likelihood instead
   b. Maybe just demonstrate surrogate modeling vs. using MBAR to help speed sampling on smaller scale?
   c. Utility of MBAR
      (1) Way more efficient than direct simulation
      (2) Able to describe parameters of estimated distribution when there is significant configurational overlap with the original
   d. Utilized in the same way surrogate modeling would be, but less efficient (we think)
   e. Preliminary results have shown that reweighting with MBAR is accurate within conservative changes of 3% in bonded force constant and 2% change in minimum bond length. Thus, each move during posterior construction will end in a new simulation on that cusp to generate new evidence before the next reweighting move.

7. Proposing new parameter states and MCMC acceptance criteria
    a. Every iteration of the MCMC sampling procedure a new set of parameters are proposed based on the current state and a prescribed proposal width. Right now that is done by randomly sampling from a normal distribution where the current parameter value is the average and the proposal width is the standard deviation of the normal distribution
    b. During a single iteration of the sampling loop we calculate the likelihood and prior probabilities of both the current and the proposed parameter states. The acceptance probability (p_{accept}) is the ratio of the proposal probability to the current probability ($\frac{L_{current}Pr_{current}}{L_{proposal}Pr_{proposal}}$). If the acceptance probability is greater than a random sample generated on $Uniform\,(0,1)$ then we accept the proposed state and update our position (i.e. the proposed state becomes the current state). A new iteration now begins.
    c. If the acceptance criteria are not met then the current state is retained, a new move is proposed and the cycle is repeated.
8. Proposed MCMC sampling algorithm and experiments
    a. *Hypothesis:* The speed of convergence of the sampling is affected by the order and frequency by which surrogate modelling and MBAR are utiilized to predict observable quantities in place of classical mechanics simulations.
        (1) **Step 1:** Simulate with an altered force field prescribed by the initial parameter states which are an input to the sampler function or those proposed by the last step of the sampler.
        (2) **Step 2:** Calculate the observables associated with the parameters being changed (bond length average and variance for bonded parameter types, bond angle average and variance for angle parameter types, etc.)
        (3) **Step 3:** Several options are possible in step 3, in order to test our hypothesis
            (a) **3a:** Use MBAR to conservatively extend our knowledge of the observables in local parameter space. From this we can construct surrogate models for the observables and carry out MCMC sampling on the surrogate models. Once the sampling has converged (which can be heuristically determined by examining the previous few MCMC samples), carry out a new simulation at that final state.
            (b) **3b:** Repeat steps 1 and 2 N more times (prescribed parameter states determined by MCMC acceptance criteria) and then use those N+1 simulations to construct surrogate models for our observables. Carry out MCMC sampling on the surrogate models. Once the sampling has converged, carry out a new simulation at that final state.

(c) **3c:** Same as 3b, but without the use of MBAR.

(4) **Step 4:** Carry out a new simulation with parameters prescribed by a new proposed MCMC move and calculate the observables associated with the parameters being changed. If the calculated observables are statistically the same as those from the previous parameter state then the sampling has fully converged and the final parameters have been determined. Else, go back to step 1.

IV. Results and Analyses

A. Ideas for deliverables/presentation of results:

1. I think it would be worthwhile to show the evolution of the posterior distribution over iterations, so maybe show a few snapshots of the posterior heat map for select SMIRKS (k vs x_0, or the equivalent for the angle parameters) over iterations, i.e.

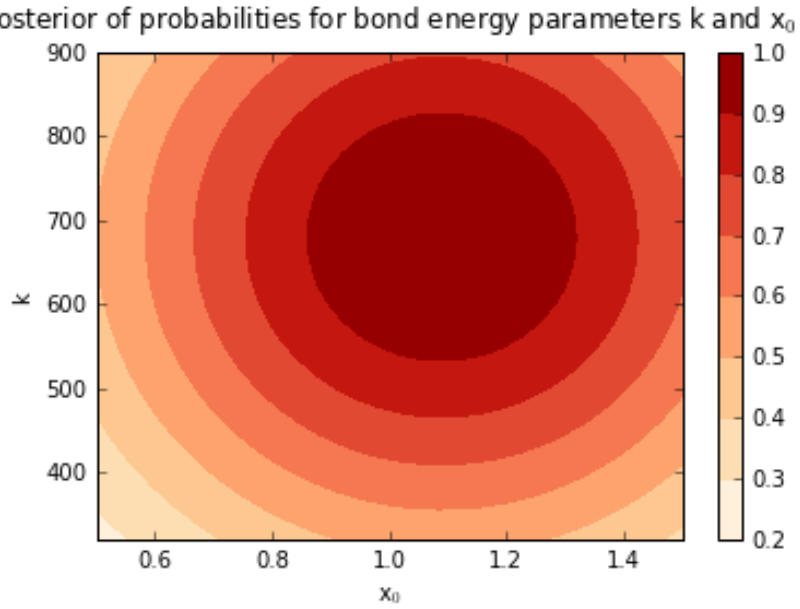

Figure 2: Final iteration of 2D heatmap representation of posterior probability distribution marginalized over all others.

2. Show any non-gaussian posteriors
3. Want to compare our final Bayesian sampled forcefield to the original
   a. How different are they?
   b. Compare observables
   c. Error

   d. Confidence intervals on final parameter values pulled from posterior (need to investigate best way to go about this)
4. Efficiency of process with and without surrogate modeling (just simulation vs with MBAR vs with surrogate modeling)

V. Conclusions

A. Impacts of study and implication for uses of classical force field parameterization

B. Highlight that, despite practical complexity, the parameterization process can and will get much more complex

C. We have presented a novel and fully automated process for parameterization of classical force fields driven by Bayesian inference given some experimental data. Not only does the process provide fully automated parameter optimization and selection based on probability, but also a means to update extant classical force fields with new experimental data. The original force field parameters were all recovered within the uncertainties that we determined from their posterior distributions (let's assume).

D. While we have only presented a toy problem to test the validity and efficiency of the process; we have well demonstrated the challenges presented by force field parameterization. Moreover, we have shown that using simple surrogate models in order to cheaply calculate observables a function of parameter greatly decreases the costs of assessing the likelihood function. While this was not imperative to the success of this initial parameterization problem, it will be as we move towards using bulk phase properties as evidence where large scale simulations will become a prohibitive computational expense when assessing the likelihood function.

Bibliography

[1] O. F. Lange, D. van der Spoel, and B. L. de Groot, "Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data," *Biophys J*, vol. 99, pp. 647–655, July 2010.

[2] F. Martín-García, E. Papaleo, P. Gomez-Puertas, W. Boomsma, and K. Lindorff-Larsen, "Comparing Molecular Dynamics Force Fields in the Essential Subspace," *PLoS One*, vol. 10, Mar. 2015.

[3] J. P. Ewen, C. Gattinoni, F. M. Thakkar, N. Morgan, H. A. Spikes, and D. Dini, "A Comparison of Classical Force-Fields for Molecular Dynamics Simulations of Lubricants," *Materials*, vol. 9, p. 651, Aug. 2016.

[4] D. Petrov and B. Zagrovic, "Are Current Atomistic Force Fields Accurate Enough to Study Proteins in Crowded Environments?," *PLOS Computational Biology*, vol. 10, p. e1003638, May 2014.

[5] O. Guvench and A. D. MacKerell, "Comparison of protein force fields for molecular dynamics simulations," *Methods Mol. Biol.*, vol. 443, pp. 63–88, 2008.

[6] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *J. Am. Chem. Soc.*, vol. 117, pp. 5179–5197, May 1995.

[7] W. L. Jorgensen, J. Chandrasekhar, and J. D. Madura, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, pp. 926–935, July 1983.

[8] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, "Development of an improved four-site water model for biomolecular simulations: TIP4p-Ew," *The Journal of Chemical Physics*, vol. 120, pp. 9665–9678, May 2004.

[9] S. K. Burger and G. A. Cisneros, "Efficient optimization of van der Waals parameters from bulk properties," *J Comput Chem*, vol. 34, pp. 2313–2319, Oct. 2013.

[10] M. M. Law and J. M. Hutson, "I-nolls: A program for interactive nonlinear least-squares fitting of the parameters of physical models," *Computer Physics Communications*, vol. 102, no. 1, pp. 252 – 268, 1997.

[11] I. J. Chen, D. Yin, and A. D. MacKerell, "Combined ab initio/empirical approach for optimization of Lennard-Jones parameters for polar-neutral compounds," *J Comput Chem*, vol. 23, pp. 199–213, Jan. 2002.

[12] D. Horinek, S. I. Mamatkulov, and R. R. Netz, "Rational design of ion force fields based on thermodynamic solvation properties," *The Journal of Chemical Physics*, vol. 130, p. 124507, Mar. 2009.

[13] M. Z. Hernandes and R. L. Longo, "AIPAR: ab initio parametrization of intermolecular potentials for computer simulations," *J Mol Model*, vol. 11, pp. 61–68, Feb. 2005.

[14] L.-P. Wang, T. J. Martinez, and V. S. Pande, "Building Force Fields: An Automatic, Systematic, and Reproducible Approach," *J. Phys. Chem. Lett.*, vol. 5, pp. 1885–1891, June 2014.

[15] L.-P. Wang, T. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martinez, and V. S. Pande, "Systematic Improvement of a Classical Molecular Model of Water," *J. Phys. Chem. B*, vol. 117, pp. 9956–9972, Aug. 2013.

[16] L.-P. Wang, J. Chen, and T. Van Voorhis, "Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data," *J. Chem. Theory Comput.*, vol. 9, pp. 452–460, Jan. 2013.

[17] J. W. Ponder and D. A. Case, "Force fields for protein simulations," *Adv. Protein Chem.*, vol. 66, pp. 27–85, 2003.

[18] L. Monticelli and D. P. Tieleman, "Force fields for classical molecular dynamics," *Methods Mol. Biol.*, vol. 924, pp. 197–213, 2013.

[19] K. Farrell, J. T. Oden, and D. Faghihi, "A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems," *Journal of Computational Physics*, vol. 295, pp. 189–208, Aug. 2015.

[20] K. Klein, S. Hennig, and S. K. Paul, "A Bayesian Modelling Approach with Balancing Informative Prior for Analysing Imbalanced Data," *PLOS ONE*, vol. 11, p. e0152700, Apr. 2016.

[21] S. Wu, P. Angelikopoulos, C. Papadimitriou, R. Moser, and P. Koumoutsakos, "A hierarchical Bayesian framework for force field selection in molecular dynamics simulations," *Phil. Trans. R. Soc. A*, vol. 374, p. 20150032, Feb. 2016.

[22] P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos, "Bayesian uncertainty quantification and propagation in molecular dynamics simulations: A high performance computing framework," *The Journal of Chemical Physics*, vol. 137, p. 144103, Oct. 2012.

[23] J. Zhu, J. Chen, W. Hu, and B. Zhang, "Big Learning with Bayesian Methods," *arXiv:1411.6370 [cs, stat]*, Nov. 2014. arXiv: 1411.6370.

[24] F. Cailliez, A. Bourasseau, and P. Pernot, "Calibration of forcefields for molecular simulation: Sequential design of computer experiments for building cost-efficient kriging metamodels," *J. Comput. Chem.*, vol. 35, pp. 130–149, Jan. 2014.

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

[26] G. E. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. John Wiley Sons, Inc., 1992.

[27] "SMIRFF forcefield github repository." https://github.com/open-forcefield-group/smarty. Accessed: 2017-03-09.