Outline: Bayesian sampling driven classical mechanics force field parameterization

Bryce C. Manubay

July 13, 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 [**?**, **?**, **?**, **?**, **?**]
      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. [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]

      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][**?**, **?**, **?**]
      3. Find more examples

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

      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.**

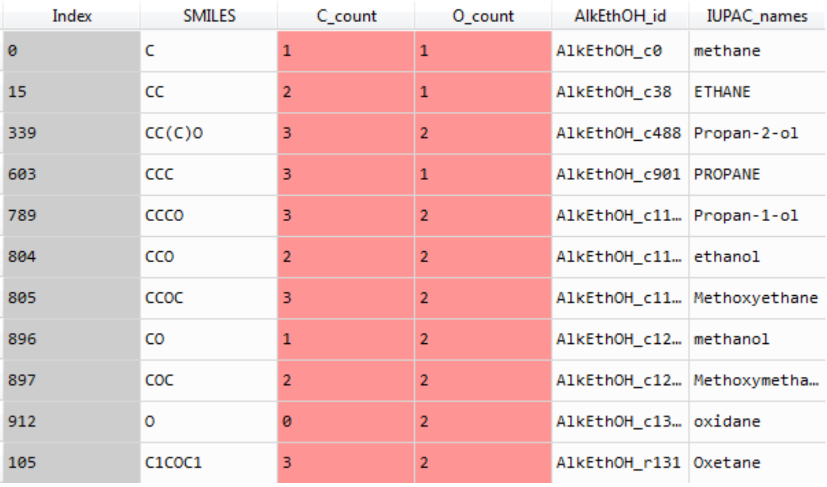
D. Bayesian statistics has been applied to a large number of big data and optimization problems [?, ?, ?, ?, ?, ?, ?, ?]
1. add citations from the recent lit stuff on OpenFF Slack

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.[?]
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 ‘smirnoff99Frosst’ 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) $f_N\left(x\right) = \sum_{i=0}^{N} A_i \sin^2\left(\frac{2i\pi x}{\tau_i} + \psi_i\right)$

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

2. Bayes' Theorem

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

b. Where $Pr\left(\right)$ 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) $r_{i,j} = \sqrt{\left(\bar{x}_i - \bar{x}_j\right)^2}$

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

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

(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)

(1) Deciding on surrogate models to use

(2) Parameters

- Equilibrium bond length: x_0 (units of)
- Bonded force constant: k_{bond}
- k^*_{bond}(Transformed bonded force constant): $\frac{1}{\sqrt{k_{bond} \times \beta}}$ (units of)
- Equilibrium bond angle: θ_0 (units of radians)
- Angular force constant: k_{angle}
- k^*_{angle}(Transformed angular force constant): $(\frac{\pi}{180}) \times \frac{1}{\sqrt{k_{angle} \times \beta}}$ (units of radians)

(3) Model forms

(4) Bond Observables

- $O_{bond,av}(x_0, k^*_{bond})$ (Bond length average)

- $O_{bond,var}(x_0, k^*_{bond})$ (Bond length variance)

(5) Model equations (listed in increasing order of complexity)

(6) Simple theoretical

- $O_{bond,av}(x_0) = x_0$
- $O_{bond,var}(k^*_{bond}) = {k^*_{bond}}^2$

(7) Linear (no intercept or cross term)

- $O_{bond,av}(x_0) = a_{0,av} \times x_0 + a_{1,av} \times k^*_{bond}$
- $O_{bond,var}(k^*_{bond}) = a_{0,var} \times x_0 + a_{1,var} \times k^*_{bond}$

(8) Linear (w/ intercept)

- $O_{bond,av}(x_0) = a_{0,av} + a_{1,av} \times x_0 + a_{2,av} \times k^*_{bond}$
- $O_{bond,var}(k^*_{bond}) = a_{0,var} a_{1,var} \times x_0 + a_{2,var} \times k^*_{bond}$

(9) Linear (w/ intercept and cross term)

- $O_{bond,av}(x_0) = a_{0,av} + a_{1,av} \times x_0 + a_{2,av} \times k^*_{bond} + a_{3,av} \times x_0 \times k^*_{bond}$
- $O_{bond,var}(k^*_{bond}) = a_{0,var} a_{1,var} \times x_0 + a_{2,var} \times k^*_{bond} + a_{3,var} \times x_0 \times k^*_{bond}$

(10) Quadratic

- $O_{bond,av}(x_0) = a_{0,av} + a_{1,av} \times x_0 + a_{2,av} \times k^*_{bond} + a_{3,av} \times x_0 \times k^*_{bond} + a_{4,av} \times x_0^2 + a_{5,av} \times {k^*_{bond}}^2$
- $O_{bond,var}(k^*_{bond}) = a_{0,var} a_{1,var} \times x_0 + a_{2,var} \times k^*_{bond} + a_{3,var} \times x_0 \times k^*_{bond} + a_{4,var} \times x_0^2 + a_{5,var} \times {k^*_{bond}}^2$

(11) Angle Observables

- $O_{angle,av}(x_0, k^*_{angle})$ (Bond length average)
- $O_{angle,var}(x_0, k^*_{angle})$ (Bond length variance)

(12) Model equations

(a) Same models replacing the x_0 with θ_0 and k^*_{bond} with k^*_{angle}

b. Justification of surrogate models for $O(\theta)$ for bond and angle observables

(1) Main criteria for choosing a model were based on calculated RMSD between simulated observables and surrogate models values evaluated at the same parameters. We want models that are "good enough" in order to rapidly sample across parameter space that are as simple as possible. We chose the simplest models with RMSD less than an order of magnitude greater than the next most complex model. The surrogates tested were simple theoretical models, linear surfaces of varying complexity and up to a nine parameter quadratic polynomial surface.

(2) Bonds

(a) For the bond length average observable it was determined that the simple theoretical model, which in this case is just the value of the equilibrium bond length parameter (x_0) is sufficient to describe the bond length average. The next most complex model was a two parameter planar surface (no intercept or cross term) and the decrease in RMSD between model and simulation was less than 2.6 fold (from 0.0037 to 0.0014).

(b) For bond length variance we determined that the simple theoretical model ($\frac{1}{k_{bond} \times \beta}$ units of ^{2}) was sufficient. With an RMSD of 8.65e-5 between model and simulation, only only the quadratic surface model performed slightly better.

(3) Angles

(a) For bond angle average the two parameter planar model was chosen which had an RMSD of 0.439 between model and simulation, 2 times greater than the next most complex model (linear model with a third intercept parameter). It is of note that the simple theoretical model (equilibrium bond angle parameter, θ_0) was highly inadequate here due to the great variability in observed bond angle average across chemistries.

(b) For bond angle variance the simple theoretical model was chosen. The RMSD between model and simulation was 0.00891, less than 1.06 times greater than the next most complex model (two parameter linear model).

D. Multistate reweighting using MBAR

1. Not too sure how useful this is going to be if we can use surrogate modeling to help assess likelihood instead
2. Maybe just demonstrate surrogate modeling vs. using MBAR to help speed sampling on smaller scale?
3. Utility of MBAR
   a. Way more efficient than direct simulation
   b. Able to describe parameters of estimated distribution when there is significant configurational overlap with the original
4. Utilized in the same way surrogate modeling would be, but less efficient (we think)
5. 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.

E. Proposing new parameter states and MCMC acceptance criteria

1. 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

2. For each new proposal of parameters during iterative sampling, 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 $Uniform\left(0,1\right)$ then we accept the proposed state and update our position (i.e. the proposed state becomes the current state). A new iteration now begins.
3. If the acceptance criteria are not met then the current state is retained, a new move is proposed and the cycle is repeated.

F. Proposed MCMC sampling algorithm and experiments

1. *Hypothesis:* The speed of convergence of the sampling is affected by the order and frequency by which surrogate modelling and MBAR are utilized to predict observable quantities in place of classical mechanics simulations.

   a. **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.

   b. **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.)

   c. **Step 3:** Several options are possible in step 3, in order to test one of our secondary hypotheses

   (1) **3a:** Use MBAR to conservatively extend our knowledge of the observables in local parameter space. From this local knowledge of $O\left(\Theta\right)$ we can construct surrogate models for the observables and carry out MCMC sampling to perform bayesian inference using the surrogate models as forward data models. Once the sampling has converged (which can be heuristically determined by examining the previous N MCMC samples), carry out a new simulation at that final state.

   (2) **3b:** Repeat steps 1 and 2 N (can test various values of N) more times (prescribed parameter states determined by MCMC acceptance criteria described in the previous section) and then extend our local knowledge of $O\left(\Theta\right)$ using MBAR. From this local knowledge of $O\left(\Theta\right)$ we can construct surrogate models for the observables and carry out MCMC sampling to perform bayesian inference using the surrogate models as forward data models. Once the sampling has converged, carry out a new simulation at that final state.

   (3) **3c:** Repeat steps 1 and 2 N (can test various values of N) more times (prescribed parameter states determined by MCMC acceptance criteria described in the previous section). From this

local knowledge of $O(\Theta)$ we can construct surrogate models for the observables and carry out MCMC sampling to perform bayesian inference using the surrogate models as forward data models. Once the sampling has converged, carry out a new simulation at that final state.This is the same as 3b, but without the use of MBAR.

d. **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.

Results and Analyses

I. Ideas for deliverables/presentation of results:

A. 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.

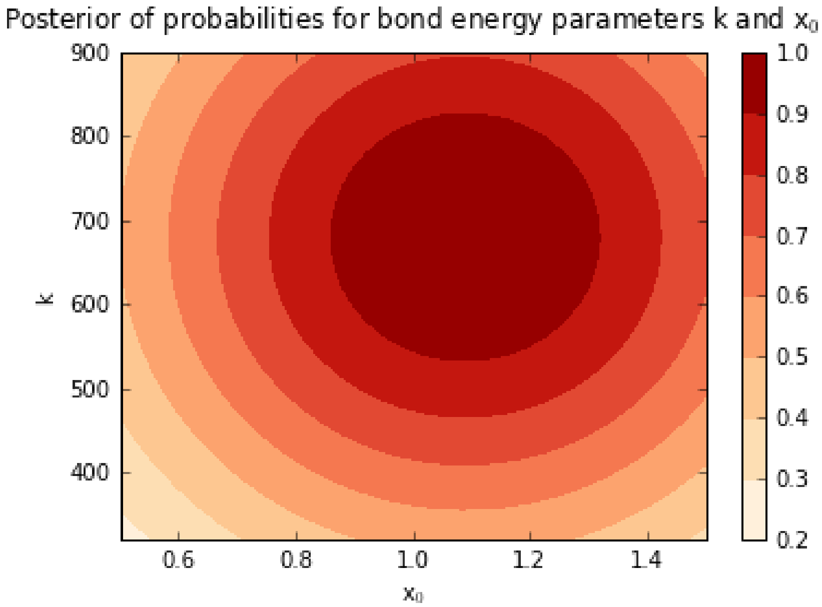


Figure 2: Final iteration of 2D heatmap representation of posterior probability distribution marginalized over all others.

B. Show any non-gaussian posteriors

C. Want to compare our final Bayesian sampled forcefield to the original

   1. How different are they?
   2. Compare observables
   3. Error
   4. Confidence intervals on final parameter values pulled from posterior (need to investigate best way to go about this)

D. Efficiency of process with and without surrogate modeling (just simulation vs with MBAR vs with surrogate modeling)

Conclusions

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

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

III. 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).

IV. 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.