Assessing systematic error in forcefields using a Bayesian approach to parameterization

Patrick B. Grinaway,[1,*] Kyle A. Beauchamp[+],[2,†] Michael R. Shirts,[3,‡] and John D. Chodera[2,§]

[1]*Graduate Program in Physiology, Biophysics, and Systems Biology, Weill Cornell Medical College, New York, NY*
[2]*Computational Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY*
[3]*Department of Chemical Engineering, University of Virginia, Charlottesville, VA*

(Dated: April 1, 2015)

This is an abstract.

Keywords: molecular mechanics forcefields; forcefield parameterization; forcefield accuracy; biomolecular simulation

I. INTRODUCTION

Predictive simulations of many molecular scale phenomena require accurate force fields. There is little dispute that atomistic molecular simulation has had an enormous impact across a wide variety of fields, from chemistry to biophysics to materials science. Molecular simulations—generally variations of molecular dynamics or Metropolis Monte Carlo simulation techniques—utilize a force field to describe the behavior of material under equilibrium or nonequilibrium conditions within the realm of statistical mechanics. For simplicity, we confine our discussion to atomistic models of molecular systems obeying classical statistical mechanics, and concentrate on equilibrium thermodynamic properties, though these concepts could readily be extended to kinetic properties or coarse-grained potentials as well. Our proposal illustrates the parameterization concepts developed in this proposal in the context of mixtures of small organic molecules in the liquid phase across a range of compositions and temperatures at ambient pressure.

Force fields specify how to construct the potential energy function to model a molecular system. Molecular mechanics force fields define how the potential energy function $U(\mathbf{x};\theta)$ and corresponding forces $F(\mathbf{x};\theta) \equiv -\nabla_{\mathbf{x}} U(\mathbf{x};\theta)$ are constructed for a given system of interest, where $\mathbf{x}$ denotes atomic coordinates and θ force field parameters. These force fields typically consist of four essential components: (1) A **functional form** specifies the potential $U(\mathbf{x};\theta)$, generally inspired by known physical behavior but with free parameters that can be fit to reproduce experimental or quantum chemical data; (2) a set of N **atom types** that describe how atoms in similar chemical environments are grouped together and assigned identical parameters, reducing the total size of the parameter space; (3) a set of **parameters** θ associated with one or more atom types for each of the kinds of interactions in the system, where interactions typically include valence terms (bond stretching, angle bending, torsions) and nonbonded terms (atomic repulsion and dispersive attraction, electrostatic interactions); and (4) a set of **nonbonded combining rules** that can be used to determine how parameters for pairs of atom types are combined, avoiding the need for $O(N^2)$ distinct sets of nonbonded interaction parameters.

Current force field parameterization approaches have a number of significant limitations. Traditionally, force fields have been constructed through a manually laborious process guided by a combination of experimental data, quantum chemical calculations, and physical insight. The functional forms in use by many modern force fields—for example, Lennard-Jones potentials for describing the dispersive and repulsive interactions between nonbonded atoms—were chosen some decades ago for describing simple liquids with forms chosen as a compromise between physical insight and computational convenience. While a variety of functional forms have been elaborated, the forms—and indeed many parameters—in use by a multitude of modern biomolecular force fields (e.g. AMBER, CHARMM, OPLS) remain largely unchanged [1]. While quantum chemical calculations have been very useful for determining many of the valence terms and charge models, they have not been as useful as experimental data in the parameterization of nonbonded interactions, which we focus on in this proposal.

The procedure by which force fields have been parameterized has gradually become more sophisticated over the decades as computational power has increased and the systems modeled have become more complex. Early models of water, such as TIP3P and TIP4P [2], were essentially parameterized by iterative manual selection of geometry and parameters given a fixed functional form. Early biomolecular force fields, such as AMBER *parm94* [3], used human insight into the nature of chemical environments to select a variety of distinct atom types to which individual parameters were assigned, with quantum chemical calculations providing a great deal of aid in selecting valence parameters and partial atomic charges. Later attempts to parameterize the enormous space of small organic molecules with a general small molecule force field utilized semi-automated optimization approaches to select parameters, such as genetic algorithms [4] or derived extrapolation approaches [5–9]. Gradient-based optimization approaches, such as least-squares optimization of an objective function, were later in-

* patrick.grinaway@choderalab.org
† kyle.beauchamp@choderalab.org
‡ michael.shirts@virginia.edu
§ Corresponding author; john.chodera@choderalab.org

troduced, as in the parameterization of the TIP4P-Ew water model [10].

Despite this progress, critical deficiencies in the force field parameterization process remain:

- **Atom types are imposed by fiat**, and are products of chemical intuition, without statistical clarity on whether available experimental data is being under- or overfit.
- **Least-squares optimization techniques are vulnerable to getting trapped in local optima** in parameter space, with no clear way to tell if global optima have been found.
- **Objective functions require human-assigned weights** to individual classes of experimental data in order to include them in the same objective. [11]
- **Functional forms and combining rules are often chosen for convenience or through historical inertia**, rather than a data-driven approach that penalizes unnecessary complexity.
- **There is currently no clear way to quantify the systematic error** induced by uncertainty in the appropriate choice of atom types, functional forms, parameters, and combining rules.

Bayesian inference provides a statistical framework for data-driven parameter selection. The fundamental concepts behind Bayesian inference are straightforward. Given a model $\mathcal{M}$ with unknown parameters θ_* and observed data $\mathcal{D}$ generated from the model, we can write the conditional probability that a particular choice of parameters θ was responsible for data $\mathcal{D}$ as $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)\, p(\theta)$, where $p(\mathcal{D}|\theta)$ is the probability that the data $\mathcal{D}$ were observed given true model parameters θ, and $p(\theta)$ denotes the prior probability of parameter choice θ known *a priori* before any data was observed. Note that the result of a Bayesian inference step is not a single parameter set $\hat{\theta}$, but an entire *posterior distribution* over parameters given data, $p(\theta|\mathcal{D})$. When this distribution is not analytically tractable, it can be efficiently sampled using standard Markov chain Monte Carlo techniques [12] essentially identical to standard techniques used to sample equilibrium distributions from molecular mechanics force fields.

The fundamental concept behind this proposal is to recast the force field parameterization problem as a Bayesian inference problem. In this framework, the appropriate underlying model (atom types, functional forms, combining rules) and associated parameters (force field parameters θ) are jointly *inferred* from a set of experimental data $\mathcal{D}$. Critical to this approach is the ability to construct the likelihood function $p(\mathcal{D}|\theta, \mathcal{M})$ based on an understanding of the experimental measurement process and a dataset for which the measurement uncertainties are well characterized.

Parameterizing a force field for small organic molecular liquids is a good model system for studying force field parameterization approaches. We will test these Bayesian techniques in the context of parameterizing a force field for mixtures of small organic molecular liquids at ambient pressure over a range of compositions and biologically relevant temperatures (10–60°C). While a liquid force field will have significant utility on its own, it primarily serves as a stepping stone to larger efforts once these methodologies have been validated, retaining many of the same challenges as more complex force field parameterization challenges across a range of disciplines.

Notably, both data collection and molecular simulation of these systems are tractable and inexpensive. Molecular mechanics force fields have not typically been parameterized for mixture properties such as excess heats of mixture and excess density (Figure **??**)[13, 14]. Exceptions to this cover only a few selected mixtures [15, 16], leaving only fitted analytical models generally available for such modeling complex mixtures. [17, 18] And yet, the mixing properties of molecular fluids and polymers are the driving forces which control the behavior of complex systems, from polymer self-assembly to biomolecular interactions.

This proposal also re-evaluates current choice of non-bonded potential functions and associated combining rules. These terms have not received significant attention in biomolecular or small organic molecule force fields in quite some time (with notable but rare exceptions [1]), despite the fact that significant improvement is still possible (e.g. Figure **??**). Lessons learned here will be directly applicable in the parameterization of molecular mechanics force fields for biomolecules, arbitrary small organic and inorganic molecules, and large ranges of temperatures and pressures associated with chemical engineering problems.

II. METHODS

III. CONCLUSIONS

Locavore biodiesel gentrify 90's small batch skateboard. Bicycle rights gentrify pop-up normcore, Thundercats single-origin coffee tofu American Apparel pug tattooed post-ironic. Shabby chic fanny pack biodiesel, cornhole Pinterest pug selvage forage beard literally four dollar toast roof party hella fingerstache master cleanse. Stumptown American Apparel locavore listicle. Cold-pressed hashtag Neutra kale chips, ugh occupy deep v slow-carb pug roof party Bushwick Tumblr shabby chic Austin. Pug selfies mustache umami, asymmetrical DIY mlkshk wayfarers Williamsburg farm-to-table Marfa single-origin coffee. Irony blog Marfa butcher, tousled selvage forage kale chips master cleanse single-origin coffee asymmetrical Williamsburg Neutra.

Banjo actually organic, salvia umami Odd Future pickled whatever brunch. Actually freegan kale chips cronut jean shorts, heirloom four loko organic fingerstache fap Bushwick biodiesel Thundercats asymmetrical deep v. +1 ethical umami distillery bitters, Odd Future mumblecore. Polaroid occupy vegan dreamcatcher, 90's stumptown tilde

Marfa butcher Schlitz retro. Butcher Brooklyn seitan, American Apparel gastropub cred Austin small batch. YOLO aesthetic Williamsburg selfies, try-hard mustache occupy cardigan sriracha meggings flannel wayfarers. Mumblecore kitsch deep v fixie McSweeney's Truffaut, pop-up readymade salvia skateboard hoodie dreamcatcher polaroid Helvetica.

Fanny pack photo booth crucifix, PBR trust fund pickled sustainable. Williamsburg disrupt before they sold out irony pug banjo. Kale chips lomo hella food truck, mixtape literally Blue Bottle Marfa Odd Future. Keffiyeh flexitarian normcore pickled flannel, irony tattooed. Fingerstache selfies Vice semiotics, High Life vegan Kickstarter trust fund twee bespoke literally bitters Portland. Wes Anderson taxidermy swag Austin disrupt. Pitchfork Banksy pickled, vegan cray irony drinking vinegar health goth.

Literally hashtag master cleanse, organic tofu quinoa food truck banjo chillwave drinking vinegar Etsy Williamsburg wayfarers Marfa Carles. Brunch photo booth next level, kitsch church-key bitters lomo Banksy cold-pressed gastropub 8-bit blog chambray. Carles sriracha synth, tattooed hella four loko Etsy typewriter try-hard Intelligentsia VHS irony gastropub chambray vegan. Roof party Austin small batch, sriracha tofu cronut church-key try-hard gentrify tilde. Kogi squid fanny pack cliche, mustache Carles 3 wolf moon iPhone. Austin kitsch deep v raw denim. Pork belly art party crucifix ennui, pop-up chia organic.

IV. ACKNOWLEDGEMENTS

We thank Vijay S. Pande (Stanford University), Lee-Ping Wang (Stanford University), Peter Eastman (Stanford University), Robert McGibbon (Stanford University), Jason Swails (Rutgers University), David L. Mobley (University of California, Irvine), Christopher I. Bayly (OpenEye Software), and members of Chodera lab for helpful discussions.

[1] J. W. Ponder and D. A. Case, Adv. Prot. Chem. **66**, 27 (2003).

[2] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926âĂŞ935 (1983).

[3] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. L. Ferguson, D. C. Spellmayer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).

[4] J. Wang and P. A. Kollman, J. Comput. Chem. **22**, 1219 (2001).

[5] S. K. Burger and G. A. Cisneros, J. Comput. Chem. pp. 1–7 (2013), ISSN 1096-987X, URL http://www.ncbi.nlm.nih.gov/pubmed/23828265.

[6] M. M. Law and J. M. Hutson, Comput. Phys. Comm. **102**, 252 (1997), ISSN 0010-4655, URL http://www.sciencedirect.com/science/article/pii/S0010465597000131.

[7] I. J. Chen, D. Yin, and A. D. MacKerell, J. Comput. Chem. **23**, 199 (2002), ISSN 1096-987X, URL http://dx.doi.org/10.1002/jcc.1166.

[8] M. Hernandes and R. Longo, J. Mol. Model. **11**, 61 (2005), ISSN 1610-2940, URL http://dx.doi.org/10.1007/s00894-004-0222-9.

[9] D. Horinek, S. I. Mamatkulov, and R. R. Netz, J. Chem. Phys. **130**, 124507 (2009), URL http://scitation.aip.org/content/aip/journal/jcp/130/12/10.1063/1.3081142.

[10] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, J. Chem. Phys. **120**, 9665 (2004).

[11] H. Paliwal, Ph.D. thesis, University of Virgina School of Engineering and Applied Science (2014).

[12] J. S. Liu, *Monte Carlo strategies in scientific computing* (Springer-Verlag, New York, 2002), 2nd ed.

[13] D. González-Salgado and I. Nezbeda, Fluid Phase Equil. **240**, 161 (2006).

[14] E. J. W. Wensink, A. C. Hoffmann, P. J. van Maaren, and D. van der Spoel, J. Chem. Phys. **119**, 7308 (2003).

[15] B. Chen, J. J. Potoff, and J. I. Siepmann, J. Phys. Chem. B **105**, 3093 (2001), URL http://dx.doi.org/10.1021/jp003882x.

[16] J. M. Stubbs, J. J. Potoff, and J. I. Siepmann, J. Phys. Chem. B **108**, 17596âĂŞ17605 (2004).

[17] B. E. Poling, J. M. Prausnitz, and J. P. O'Connell, *The Properties of Gases and Liquids* (McGraw-Hill, New York, 2001), 5th ed.

[18] V. Diky, R. D. Chirico, A. F. Kazakov, C. D. Muzny, and M. Frenkel, J. Chem. Info. Model. **49**, 503 (2009).