A method for redesigning molecular mechanics force field parameterization by use of a Bayesian statistical framework

Bryce C. Manubay[1,*]

[1]*University of Colorado - Department of Chemical and Biological Engineering*

(Dated: October 9, 2016)

I. Objectives $(0.5pages)$

- Molecular dynamics (MD) simulation is fast becoming a more useful tool in many scientific studies.
- However, some limitations remain in the ability of MD force fields to accurately and transferably describe molecular environments.
- Currently, force fields are parameterized with fixed functional forms with, often, poor physical motivation and require the chemical intuition of experts to manually correct parameters, leading to a more suitable product. Additionally, the creation of a transferable method to update existing force fields based on new experimental data is limited due to lack of understanding and lack of consistency in how the original parameterization was done
- A possible solution to these problems is by recasting the force field parameterization process as a bayesian inference problem.
- The objective of this paper is introduce a framework for using high quality experimental data in order to automatically generate families of MD force fields consistent with the data used.
- In this paper I will describe the overall parameterization framework and my roles in the project, first, collecting and organizing large amounts of high quality thermochemical data and, currently, investigating use of the Multistate Bennett Acceptance Ratio (MBAR) as a means to improve throughput by reducing simulation requirements during the parameterization process.

II. Significance $(0.5pages)$

- A broad variety of research has been greatly impacted by the advent and improvement of MD simulation tools.
  - Observing physical phenomena at a molecular scale (phase changes, ligand docking, etc.)[1]

* bryce.manubay@colorado.edu

  - Drug discovery and deisgn of new molecules[2]

- The fundamental part of molecular simulation for describing the energetic interactions of a system is referred to as a force field, hence transferable and quantitatively accurate force fields are imperative for the use of molecular simulation tools to continue to proliferate.
  - Transferability of MD force fields and particularly sets of parameters is an extremely popular topic (and current limitation) in the molecular simulation field.[3–6] Transferability encourages use by providing convenience for scientists with wide arrays of research interests and simplifying the mystery that most observe force fields with.
  - Inaccurate and poorly parameterized force fields have been shown to grossly misrepresent molecular systems[7,8]

- A few notable attempts, such as GAAMP and ForceBalance, have been made in recent years towards more automated and systematic force field parameterization methods.[9–12] 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. However none provided the ability for the computer to automatically and systematically explore choices of fitting data, optimization algorithm and functional forms in order to objectively find families of force fields consistent with fitting data and reward those with the least model complexity.

III. Background and related literature $(1.5pages \pm 0.5pages)$

- Molecular dynamics force fields define how we construct the potential energy functions (and thereby the forces) of an atomistic system under study. The potential is constructed such that it is a function of solely the atomic coordinates and a set of parameters associated with the force field. Transferable force fields generally have three major parts:

  1 The **functional forms** of the potential, i.e. the mathematical equations for the energy equation. A classic example of a non-bonded interaction form is the 12-6 Lennard-Jones (LJ) potential.

  2 **Atom types** which describe similar chemical environments such that we can assign different atoms identical parameters, thereby shrinking the parameter space and helping to avoid overfitting.

  3 **Parameters** that are associated with one or many atom types which determine the magnitude of the interactions in the system

- As alluded to earlier, there are severe limitations in current methods for force field parameterization. Until very recently, force fields have primarily been made manually guided by experimental and quantum chemical simulation data as well as the intuition of expert computational chemists.[13–17] Some functional forms used in modern force fields, like the 12-6 LJ potential, have poor physical basis. While the attractive term of the LJ potential has physical basis on the true behavior of dispersion forces, the repulsive term loosely approximates Pauli repulsion and is used for computational convenience. Despite attempts at improvement, many of the functional forms and parameters of these force fields also remain mostly unchanged due to the lack of clear, systematic methods for updating them.[18]

- As stated previously, parameterization methods have slowly become more sophisticated over the last decade and a half. Many early biomolecule force fields were parameterized manually guided by chemical intuition, but force fields like AMBER *parm94* showed intuitive departure by shrinking parameter space with clever atom typing.[19] The parameterization of GAFF used a semi-automated genetic algorithm approach to select parameters.[17] Even more sophisticated optimization approaches such as least-squares optimization of an objective function have been utilized in creation of the TIP4P-Ew water model[20] and in the ForceBalance parameterization scheme[10–12]. Even with these more sophisticated optimization schemes there are still issues in needing to assign weights to different kinds of data (i.e. different properties) when they are included in the same objective. There are also possibilities of multiple optima in parameter space (i.e. different sets of parameters that all are consistent with data used during parameterization) and least-squares optimization does not discriminate the global optima from the other possibilities.

- Bayesian inference provides a robust statistical framework for force field parameterization. It has been shown that bayesian approaches can be applied toa wide variety of data driven sciences. It's been used for balancing data to help minimize influence of oversampled populations and generate more robust predictive models[21] to recalibrating initial force estimates in coarse grained MD models to target atomistic MD and experimental data[22]. Baye's theorem clearly provides a framework for the problem at hand thusly:

$$P\left(\theta|D\right) \propto P\left(D|\theta\right) P\left(\theta\right) \tag{1}$$

  In the above, θ is a family of parameters consistent with data D. What this bayesian inference produces is not just a single parameter set, but an entire posterior distribution of over parameters given data. This is advantageous given that many highly correlated sets of parameters can equally well fit the dataand the distribution of these consistent families of parameters can inform what new data could help narrow the distribution and improve the parameter estimates.

A. **Methods** $(1.5pages \pm 0.5pages)$

B. **Progress** $(1.5pages \pm 0.5pages)$

C. **Research plan** $(0.5pages)$

---

[1] G. Jayachandran, V. Vishal, and V. S. Pande, "Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece," *J Chem Phys*, vol. 124, no. 16, p. 164902, Apr. 2006. [Online]. Available: http://scitation.aip.org/content/aip/journal/jcp/124/16/10.1063/1.2186317

[2] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, “Role of Molecular Dynamics and Related Methods in Drug Discovery,” *J. Med. Chem.*, vol. 59, no. 9, pp. 4035–4061, May 2016. [Online]. Available: http://dx.doi.org/10.1021/acs.jmedchem.5b01684

[3] N. A. Vellore, J. A. Yancey, G. Collier, R. A. Latour, and S. J. Stuart, “Assessment of the Transferability of a Protein Force Field for the Simulation of Peptide-Surface Interactions,” *Langmuir*, vol. 26, no. 10, pp. 7396–7404, May 2010. [Online]. Available: http://dx.doi.org/10.1021/la904415d

[4] D. A. Puleo and R. Bizios, *Biological Interactions on Materials Surfaces: Understanding and Controlling Protein, Cell, and Tissue Responses*. Springer Science & Business Media, Jun. 2009.

[5] F. Sato, S. Hojo, and H. Sun, “On the Transferability of Force Field ParametersWith an ab Initio Force Field Developed for Sulfonamides,” *J. Phys. Chem. A*, vol. 107, no. 2, pp. 248–257, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1021/jp026612i

[6] A. Martin-Calvo, J. J. Gutiérrez-Sevillano, J. B. Parra, C. O. Ania, and S. Calero, “Transferable force fields for adsorption of small gases in zeolites,” *Phys Chem Chem Phys*, vol. 17, no. 37, pp. 24 048–24 055, Oct. 2015.

[7] 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, no. 2, pp. 647–655, Jul. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2905107/

[8] 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, no. 3, Mar. 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374674/

[9] L. Huang and B. Roux, “Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data,” *J. Chem. Theory Comput.*, vol. 9, no. 8, pp. 3543–3556, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1021/ct4003477

[10] 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, no. 11, pp. 1885–1891, Jun. 2014. [Online]. Available: http://dx.doi.org/10.1021/jz500737m

[11] 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, no. 34, pp. 9956–9972, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1021/jp403802c

[12] 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, no. 1, pp. 452–460, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1021/ct300826t

[13] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, “CHARMM: A program for macromolecular energy, minimization, and dynamics calculations,” *J. Comput. Chem.*, vol. 4, no. 2, pp. 187–217, Jun. 1983. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/jcc.540040211/abstract

[14] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. WiÃşrkiewicz-Kuczera, D. Yin, and M. Karplus, “All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins,” *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, Apr. 1998. [Online]. Available: http://dx.doi.org/10.1021/jp973084f

[15] G. C. Soo, F. K. Cartledge, R. J. Unwalla, and S. Profeta, “Development of a molecular mechanics (MM2) force field for Îś-chlorosilanes,” *Tetrahedron*, vol. 46, no. 24, pp. 8005–8018, Jan. 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0040402001814575

[16] T. A. Halgren, “Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94,” *J. Comput. Chem.*, vol. 17, no. 5-6, pp. 490–519, Apr. 1996. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P/abstract

[17] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general amber force field,” *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/jcc.20035/abstract

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

[19] 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, no. 19, pp. 5179–5197, May 1995. [Online]. Available: http://dx.doi.org/10.1021/ja00124a002

[20] 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, no. 20, pp. 9665–9678, May 2004. [Online]. Available: http://scitation.aip.org/content/aip/journal/jcp/120/20/10.1063/1.1683075

[21] K. Klein, S. Hennig, and S. K. Paul, “A Bayesian Modelling Approach with Balancing Informative Prior for Analysing Imbalanced Data,” *PLOS ONE*, vol. 11, no. 4, p. e0152700, Apr. 2016. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152700

[22] P. N. Patrone, T. W. Rosch, and F. R. P. Jr, “Bayesian calibration of coarse-grained forces: Efficiently addressing transferability,” *The Journal of Chemical Physics*, vol. 144, no. 15, p. 154101, Apr. 2016. [Online]. Available: http://scitation.aip.org/content/aip/journal/jcp/144/15/10.1063/1.4945380