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

Bryce C. Manubay

University of Colorado - Department of Chemical and Biological Engineering

(Dated: October 14, 2016)

I. Objectives

Molecula 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. Many popular and currently used force fields were parameterized with fixed functional forms which, often, have poor physical motivation. The chemical intuition of experts is also often required 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 parameterizations were 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 generally describe the overall parameterization framework and my roles in the project thus far. First, collecting and curating large amounts of high quality experimental thermochemical data and, currently, investigating use of the Multistate Bennett Acceptance Ratio (MBAR) as a means to improve parameterization throughput by reducing computational expense while making updates to the posterior distribution of parameter sets consistent with experimental data provided.

II. Significance

A broad variety of research from drug discovery to metallurgy has been greatly impacted by the advent and improvement of MD simulation tools. Observing physical phenomena such as protein folding dynamics and ligand docking at a molecular scale is widely studied using MD tools.[1;2] Drug discovery and deisgn of new pharmaceutical leads has also been made more efficient.[3] The fundamental part in molecular simulation for describing the energetic interactions of a system is referred to as a force field. Hence, the development of force fields which are readily transferable between dissimilar physical systems and are quantitatively accurate is imperative for the use of molecular simulation tools to continue to proliferate.

Transferability of MD force fields, and particularly sets of force field parameters, is an extremely popular topic (and current limitation) in the molecular simulation field.[4;5;6;7] Transferability of force fields encourages use by providing convenience for scientists with wide arrays of research interests and by making parameter space less complex through generalization by chemical similarity. Inaccurate and poorly parameterized force fields have been shown to grossly misrepresent molecular systems.[8;9;10]

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.[11;12;13;14] 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. The bayesian inference scheme described in this paper will provide a workflow for discovering families of force field parameters consistent with experimental data and a variety of functional forms.

Additionally, as I will demonstrate later in my discussion of data mining and curation of the NIST ThermoML database, the chemical diversity in readily available thermochemical databases is lacking. Not only that, but the distribution of data amongst commonly measured properties is heavily skewed towards certain properties. Having learned this since beginning the project, one of the potential uses of the parameterization scheme is to fill in the many gaps in experimental thermochemical data. With updated general descriptions of chemical space and property data on

chemically similar compounds, this parameterization process should provide force fields to accurately simulate property data for which no experimental data exists.

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

Molecular dynamics force fields define how to 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 one can assign different atoms (or series of 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

Rolled into functional forms, **combining rules** are also sometimes considered. **Combining rules** describe how to combine parameters when an interaction contains multiple atom types.

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.[15;16;17;18;19] 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 popular force fields remain mostly unchanged due to the lack of clear, systematic methods for updating them.[20]

Parameterization methods have slowly become more sophisticated over the last decade and a half with advances in computational power and to accommodate modeling increasingly more complex systems. Many early force fields were parameterized manually for narrow classes of molecules with large redundant parameter spaces.[21] Force fields like AMBER *parm94* showed intuitive departure by shrinking parameter space with clever atom typing defined by expert computational chemists.[22] The parameterization of GAFF used a semi-automated genetic algorithm approach to select parameters.[19] Even more sophisticated optimization approaches such as least-squares optimization of an objective function have been utilized in the creation of the TIP4P-Ew water model[23] and in the ForceBalance parameterization scheme[12;13;14]. Even with these more sophisticated optimization schemes there are still issues in needing for the user to assign weights to different kinds of data (i.e. different properties) when they are included in the same objective function. Molecular systems aren't necessarily uniquely defined by a single parameter set. There are 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 to a 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[24] to recalibrating initial force estimates in coarse grained MD models to target atomistic MD and experimental data[25]. 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 **Equation (1)**, consider a model M (including functional forms and atom types) with some unknown set of parameters which produced data D. θ is a choice of parameters consistent with data D. What **equation (1)** states is that the probability of θ given D (the *posterior*) can be determined from the probability of observing D given θ (the *likelihood function*) and the probability of θ (the *prior*). The *prior* is imposed by physical constraint or by the previous round of inference. Note that in iterative bayesian inference, the posterior of the previous round becomes the prior in the new iteration. This bayesian inference produces not just a single parameter set, but an entire posterior distribution of parameters given data. This is advantageous given that many different parameter

sets can be consistent with the data used and the distribution of these consistent sets of parameters can inform what new data could help narrow the distribution and improve the parameter estimates.

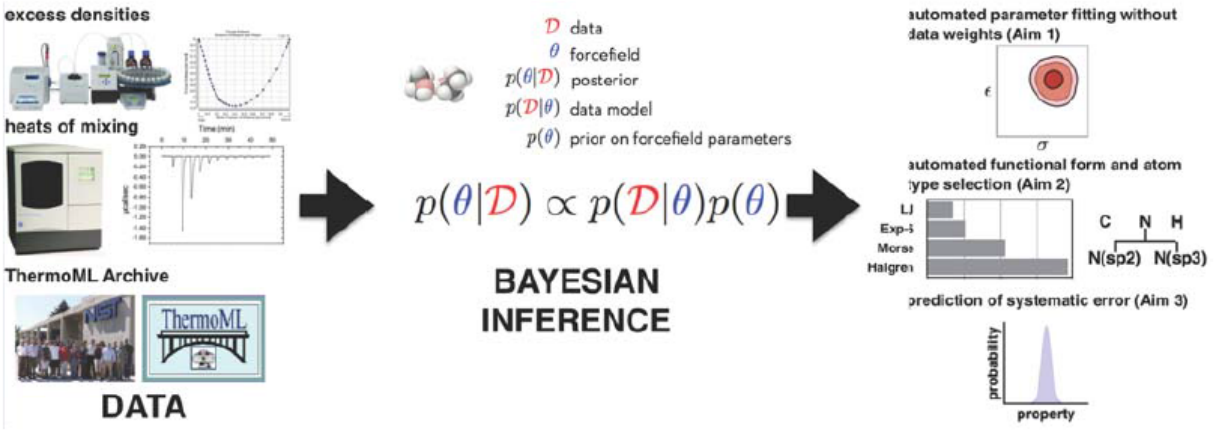

FIG. 1: A schematic overview of the bayesian inference workflow where force field parameters are inferred given experimental data and a model

The remaining **Methods** and **Progress** sections will be split it into two parts each. The first will address the problem of gathering, categorizing and curating large amounts of experimental thermochemical data for use as evidence in the bayesian inference scheme. The second part will discuss how to address decreasing the computational expense of the simulations required for the parameterization process using MBAR to estimate simulated properties at new parameter states.

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

1. Data mining and curation of thermochemical data in ThermoML

The experimental thermochemical data to be used for the parameterization process was collected from the NIST ThermoML database managed by the Thermodynamics Research Center at the NIST Boulder campus. In order to explore the composition of the data in ThermoML I used the *ThermoPyL* Python tool developed by the Chodera lab at the Memorial Sloan Kettering Cancer Center.[26] The *ThermoPyL* tool parses the standard ThermoML XML format into a Pandas dataframe format. Further filters can be applied to the dataframes to filter by chemical composition, properties of thermodynamic state such as temperature or pressure, as well as thermochemical properties available.

The first planned novel use of the bayesian parameterization procedure is to parameterize a general force field for simulating small organic liquids and their mixtures. Other than being a concrete test case with readily available experimental data, this choice is motivated by shortcomings of current force fields to accurately describe organic liquid mixtures (and particularly excess properties).[27] Given this apparent problem the properties selected to be used in the pool of potential evidence were chosen for the purpose of more fully constraining parameter space with the goal of being able to accurately simulate properties of organic liquid mixtures. The properties chosen from neat liquid data were mass density, isobaric heat capacity, speed of sound and static dielectric constant. The properties chosen from the binary mixture data were mass density, speed of sound, static dielectric constant, excess molar volume, excess molar isobaric heat capacity, excess molar enthalpy and infinite dilution activity coefficients. To consider how these properties might affect the constraint of parameter space one must simply think intuitively about the physics. First, consider mass density. Ultimately, the value of mass density for a bulk liquid is determined by the volume of the system and therefore the space between molecules. Therefore, the most important parameters for this simulated quantity would be for those describing the non-bonded interactions between molecules. As a counterexample consider the static dielectric constant, which is a function of the system dipole moment or more generally electrostatics. Clearly, accurately simulating this property would require constraints placed on some electrostatic potential and potentially parameters controlling harmonic bond potential.

A flow diagram representing the filtering process is shown in **figure 2**. The process is started by organizing a locally stored version of the ThermoML database into a Pandas dataframe. First, a filter is applied to discard journal articles with known erroneous data. Next, we filter all data if it doesn't fall within our previous properties of interest list. Filters for chemical composition and bond order are next applied, specifically it was decided that for initial testing to only look at organics containing C, O and/or H atoms with single or aromatic bonding. Additionally, the molecules had another filter that they must appear in a diverse list of alkanes, ethers and alcohols (coined AlkEthOH) that was constructed by Chris Bayly of OpenEye Software. AlkEthOH represents a limited test set to validate the machinery for parameterization. Finally, only data with temperatures 250 - 400 K and pressures 1 - 1000 atm are kept in the liquid phase are kept. The data is then saved in an easily machine-readable format such as JSON or PKL. Additionally, potential data for use as evidence was sent to the the TRC for validation of quality. The likelihood function described earlier will be a function of uncertainties associated with the evidence, thus accurate estimates are imperative. The TRC group led by Ken Kroenlein have internally kept estimates for uncertainties of all data points in ThermoML. Thus, their screening involves checking their uncertainty estimates against what authors published and noting any outliers.

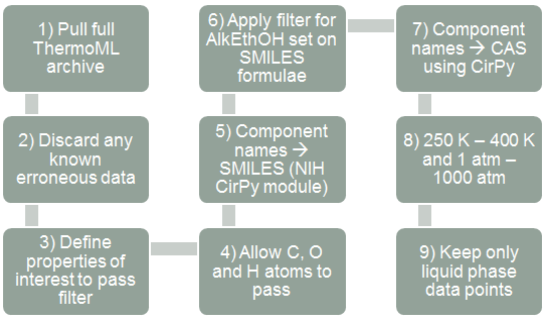

FIG. 2: A simplified diagram of the filtering algorithm used to find experimental data for use as evidence

2. Exploring use of multistate reweighting to reduce computational expense

During each update of iterative bayesian inference, evaluation of the likelihood function will require new simulated evidence given a perturbed parameter set. However, with phase space overlap, new evidence at adjacent states can be estimated using multistate reweighting tools such as MBAR.[28] Given MBAR has a much higher computational efficiency that fully simulating a new state, this could greatly accelerate full construction of a posterior distribution of parameters. *Equation 3* shown below is the formulation that allows for reduced free energies to be found at other states using a simulated state as reference, i.e. it allows for the solution of free energy differences.

$$\hat{f}_i = \sum_{j=1}^{k} \sum_{n=1}^{N_j} \frac{\exp\left[-u\left(x_{jn}\right)\right]}{\sum_{k=1}^{K} N_k \exp\left[\hat{f}_k - u_k\left(x_{jn}\right)\right]} \tag{2}$$

Where u is a reduced potential energy, x is a configuration, K is the number of states and N is the number of configurations from the state. These free energy calculations also allow for estimating expectation values of certain obserables at the other thermodynamic states for which relative free energies were calculated. This is done by calculating relative weights of the sampled state to the unsampled states using the formulations shown below in **equations 4,5 and 6**.

Where $A(x_n)$ is some mechanical observable, W_{na} is a weight and $\hat{A}$ is the estimated expectation of $A(x)$.

The goal of this exercise is to see how large of parameter perturbations can be made and still make accurate estimates of some observable at that new perturbed state. For this testing phase we've chosen a toy problem of using single molecule simulation observables, such as bond lengths and angles, as the evidence for verifying the validity of the bayesian inference process in the least computationally expensive manner. For testing the safe extent of parameter perturbation I have devised a simple scheme for comparison of reweighted observable estimates to a true sampled value.

- Data Mining and curation from ThermoML
  - ThermoPyL
  - Filtering algorithm(s)
  - Some stats on total database
  - Some of the different tests/sets we searched for to check size of chemical environment

  - AlkEthOH
- Reweighting workflow to determine length of jumps allowable in parameter space
  - Purpose of reweighting workflow
  - Brief overview of MBAR maybe
  - Criteria for safe jump and motivation behind that
  - Simulations ran

All the stats from the data mining shit

Potential figure from results of jump tests

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

C. Research plan $(0.5pages)$

References

[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] K. A. Beauchamp, D. L. Ensign, R. Das, and V. S. Pande, "Quantitative comparison of villin headpiece subdomain simulations and triplet-triplet energy transfer experiments," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108, no. 31, pp. 12 734–12 739, Aug. 2011.

[3] 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

[4] 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

[5] 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.

[6] 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

[7] 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.

[8] 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/

[9] 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/

[10] K. Vanommeslaeghe, M. Yang, and A. D. MacKerell, "Robustness in the fitting of molecular mechanics parameters," *J. Comput. Chem.*, vol. 36, no. 14, pp. 1083–1101, May 2015. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/jcc.23897/abstract

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

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

[21] N. L. Allinger, M. T. Tribble, M. A. Miller, and D. H. Wertz, "Conformational analysis. LXIX. Improved force field for the calculation of the structures and energies of hydrocarbons," *J. Am. Chem. Soc.*, vol. 93, no. 7, pp. 1637–1648, Apr. 1971. [Online]. Available: http://dx.doi.org/10.1021/ja00736a012

[22] 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

[23] 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

[24] 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

[25] 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

[26] K. A. Beauchamp, J. M. Behr, A. S. Rustenburg, C. I. Bayly, K. Kroenlein, and J. D. Chodera, "Toward Automated Benchmarking of Atomistic Force Fields: Neat Liquid Densities and Static Dielectric Constants from the ThermoML Data Archive," *J. Phys. Chem. B*, vol. 119, no. 40, pp. 12 912–12 920, Oct. 2015. [Online]. Available: http://dx.doi.org/10.1021/acs.jpcb.5b06703

[27] E. J. W. Wensink, A. C. Hoffmann, P. J. v. Maaren, and D. v. d. Spoel, "Dynamic properties of water/alcohol mixtures studied by computer simulation," *The Journal of Chemical Physics*, vol. 119, no. 14, pp. 7308–7317, Oct. 2003. [Online]. Available: http://scitation.aip.org/content/aip/journal/jcp/119/14/10.1063/1.1607918

[28] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," *J Chem Phys*, vol. 129, no. 12, Sep. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2671659/