
# Data-Driven Methods in Multiscale Modeling of Soft Matter

67

Tristan Bereau

## Contents



**Abstract**

As in many other scientific fields, data-driven methods are rapidly impacting multiscale modeling. This chapter will illustrate some of the many ways advanced statistical models and a data-centric perspective help augmenting computer simulations in soft matter. A specific focus on force fields, sampling, and simulation analysis is presented, taking advantage of machine learning, high-throughput schemes, and Bayesian inference.

## 1 Introduction

Advanced statistical models are rapidly impregnating many technological and scientific fields, from the automobile industry to robotics to particle physics. Not only do novel data-driven methods offer new perspectives on approaching long-

T. Bereau (✉)
Theory Group, Max Planck Institute for Polymer Research, Mainz, Germany
e-mail: bereau@mpip-mainz.mpg.de



W. Andreoni, S. Yip (eds.), *Handbook of Materials Modeling*,
https://doi.org/10.1007/978-3-319-44677-6_40

standing problems, they hold the promise of accelerating the pace of research. Materials science is one such field, where data is likely to accelerate computational rational design. The decisive impact of materials design in various aspects of our society has led to large-scale strategies – among others the Materials Genome Initiative (Jain et al. 2013). These recent efforts are already bearing fruit in various disciplines of hard condensed matter, inorganic chemistry, and also semiconductor physics (Curtarolo et al. 2013). Interestingly, little has happened in soft matter.

The slow development of computational materials design in soft matter likely precisely arises from what makes these systems unique: the prominent role of thermal fluctuations. Soft matter systems display weak characteristic energies on par with thermal energy, $k_\mathrm{B}T$, leading to fascinating phenomena, such as self-assembly. On the other hand, thermal fluctuations obscure the link between the chemistry and materials properties, because of the complex interplay of a system with its environment. This makes computational materials discovery for soft matter all the more challenging (Bereau et al. 2016).

Modeling soft matter systems is traditionally approached using multiscale simulations. They bridge the relevant length and time scales of the system: from quantum, to classical atomistic, to coarse-grained (CG), and to continuum resolutions. These methods are all entrenched within certain physical laws and symmetries. They stand at odds with purely data-driven methods, which typically contain little physics a priori but are instead mostly empirical. Can we benefit by combining these two paradigms?

This chapter discusses recent examples that apply data-driven methods to *augment* multiscale modeling in soft matter. Here, I will emphasize how advanced statistical models can help improve existing methodologies or offer new perspectives. The chapter describes efforts in building better force fields, tackling sampling challenges, but also efficiently analyzing computer simulations. In each case, significant progress is achieved by a variety of methods, such as machine learning (ML), high-throughput schemes, and Bayesian inference. This chapter will assume prior exposure to computer simulations – it is intended to help the simulator better grasp the benefits of introducing data-driven methods in their research.

## 2 Force Fields

Force fields lie at the heart of classical particle-based modeling. When numerically integrating Newton's equations of motion, the force field dictates how particles interact over time. As such, the force field encodes all the physics and chemistry of the model, no less. Accuracy here is critical because it determines the aggregate behavior of the system after heaps of integration steps. Emergent complexity arises from countless evaluations of $\mathbf{F} = m\mathbf{a}$. In this sense, the force field links the system's chemical composition to its long-time properties, such as free energies or kinetic properties. The corollary to this critical role is the attention force fields have received in the last three to four decades (Maple et al. 1988; Halgren 1992; Halgren and Damm 2001; Wang et al. 2001; Ponder and Case 2003; Mackerell 2004).

Force fields map a particle configuration to interaction energies and forces, leading to the coveted *potential energy surface*. The mapping ought to hit an appropriate balance between accuracy and computational investment: the physics should be described appropriately at small numerical cost. For instance, a simple spring will capture the limited range of a covalent bond but will evidently fail to describe anharmonic effects. Identifying the sweet spot depends critically on the problem at hand. The other facet of a force field development project entails transferability: given a parametrization among certain configurations, compounds, and environments, to what extent can the resulting model extrapolate to scenarios absent from the training set? In the following, we highlight recent strategies where ML has helped improve force field accuracy and transferability.

## 2.1 Accuracy: Beyond Traditional Basis Sets

Traditionally, most of the functional forms commonly used in molecular mechanics have largely been constrained by computational considerations. Among others, a pairwise decomposition is an appealing treatment of intermolecular interactions but fails to capture some of the many-body physics, as found, for instance, when modeling dispersion (Tkatchenko et al. 2012). Mathematically, this is a basis set problem: the vector space used to construct the force field fails to accurately reproduce all aspects of the underlying potential energy surface.

A striking illustration of the basis set problem arises upon coarse-graining. Coarse-graining reduces the representation of a molecular system by grouping atoms into larger particles or beads. Structure-based coarse-graining aims at a systematic derivation of CG potentials from reference atomistic simulations (Voth 2008; Peter and Kremer 2010; Noid 2013). Several methods exist to derive CG potentials that aim at best reproducing the underlying forces or distribution functions. Examples of these strategies include force matching and iterative Boltzmann inversion. The averaging performed over the degrees of freedom that have been coarse-grained away effectively leads to a potential of mean force (PMF). This PMF is typically a many-body quantity. The many-body aspect holds even when the reference simulation only relies on pairwise interactions, because of correlations owing to the missing degrees of freedom (Rühle et al. 2009). This situation makes the pairwise assumption even more critical in CG models, limiting an accurate description of the structure and thermodynamics.

Unlike standard regression schemes, a machine learning (ML) algorithm does not aim at optimally fitting parameters on a predefined basis set. It instead looks for similarities between training points to interpolate the target property in a high-dimensional feature space. Because the interpolation can always improve with added training points, a specific attribute of ML is its ability to improve its accuracy with added data. We illustrate the concept with kernel ridge regression (Rasmussen and Williams 2006), though neural network-type architectures share a number of aspects (see Behler 2016). Consider the regression of property $p$ of sample $\mathbf{x}$. A kernel machine will consist of the prediction:

$$p(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}), \tag{1}$$

where the sum runs over training samples, $\alpha_i$ is the weight of sample $i$, and $K(\mathbf{x}, \mathbf{x}')$ is the kernel between samples $\mathbf{x}$ and $\mathbf{x}'$. The kernel consists of a similarity measure, or covariance function, between two samples:

$$K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \exp\left( -\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2} \right), \tag{2}$$

where here we chose a Gaussian kernel with Euclidean distance as a measure between two samples. This metric implies that the distance is 1 and 0 for samples that are identical and very different, respectively. The middle part of the equation expresses the kernel as an inner product between the two samples. While the samples are expressed in their input space, also called "representations," the so-called *kernel trick* implicitly maps the samples via $\phi$ into an infinite-dimensional feature space, where the interpolation between samples takes place (Schiilkopf 2001). The optimization of the weights $\boldsymbol{\alpha}$ consists of solving Eq. 1 for the samples in the training set, adding a regularization term $\lambda$:

$$\boldsymbol{\alpha} = (K + \lambda \mathbb{1})^{-1} \mathbf{p}, \tag{3}$$

where $\lambda$ is set by the amount of noise in the reference data. Interestingly, while the basis set limits the possible accuracy of a regression problem, it is largely the relevance of the representation that determines the accuracy of an ML model (Huang and von Lilienfeld 2016).

There is a rapidly growing literature of studies applying machine learning to learning chemical properties, such as atomization energies (Rupp et al. 2012; Ramakrishnan and von Lilienfeld 2017). Adequate training can yield remarkably accurate predictions (Faber et al. 2017). Recent studies have aimed at using ML methods to help optimize a potential energy surface or force field (Li et al. 2017; Huan et al. 2017). When it comes to learning forces, the intrinsic orientation of the vector must be reproduced. Unlike scalar quantities, vectors contain three independent components. Different strategies have been devised to tackle this issue: the derivative of the kernel with respect to particle coordinates (Bartók et al. 2010; Chmiela et al. 2017), local axis systems (Bereau et al. 2015), or covariant kernels (Glielmo et al. 2017).

The use of ML potentials has mostly been applied to replace expensive ab initio MD simulations (Li et al. 2015; Morawietz et al. 2016; Deringer and Csányi 2017), where the computational cost difference between a single-point electronic structure calculation and ML prediction is significant. When aiming at predicting classical reference models however, the gain is smaller. Recent work on coarse-graining two benzene molecules in water indicates a better reproduction of the PMF compared to force matching (John 2016). The cost of the prediction remains significant compared to traditional pairwise potentials, but these results provide potential avenues to break the glass ceiling of pairwise interactions in CG potentials.

## 2.2 Transferability: Across Conformations, Phases, and Compositions

Molecular simulations often exhibit a complex relationship between model parameters and the resulting emergent properties. This obscures the role and impact of force field parameters. For instance, how does the tuning of a Lennard-Jones parameter affect a compound's hydration free energy or a folding timescale? This complex relationship can make force field parametrization a tedious, long, and rather unsystematic process. Systematically understanding the relationship between force field parameter and thermodynamic properties can help automate parametrization methods (Stroet et al. 2017). More often than not, coarser models tend to be more difficult to parametrize, because the missing physics require ad hoc compensations. For instance, most biomolecular atomistic models are additive – they do not explicitly model induction/polarization. Instead, mean-field polarization effects are incorporated effectively by tuning the other force field terms, most importantly Coulomb and Lennard-Jones. This typically comes at the cost of limited phase transferability: not only will they not transfer from the gas to the condensed phase, these models are typically state-point dependent. In other words, they are bound to a limited range of thermodynamic parameters, such as temperature and pressure.

Enhancing the phase transferability of these models is subject to ongoing research – a field where ML can help (Deringer and Csányi 2017) – but not the only strategy. The obscure link between model parameters and emergent properties leads to an unsystematic, largely empirical approach to force field parametrization – the craftsmanship of a biomolecular modeler. As such, developing more automated parametrization schemes offers extremely valuable perspectives: reduced parametrization efforts would speed up and enhance the pace of research in molecular simulations. In the following, two examples from atomistic and coarse-grained modeling illustrate this emerging trend. They both leverage the link between chemical properties and specific force field parameters.

High-resolution models offer a closer, more straightforward link from chemistry to force field parameters. This has motivated the development of force fields from first principles: Electronic structure calculations provide molecular and atomic properties, such as atomic polarizabilities or electrostatic coefficients, used as parameters for classical models (Van Vleet et al. 2016). This framework still requires reference calculations for every new compound considered. One can instead envision relying on the abovementioned use of ML to *predict* these chemical properties. Such a scheme was recently introduced (Bereau et al. 2018) to construct classical intermolecular potentials from atomic polarizabilities, multipole electrostatic coefficients (Bereau et al. 2015), and atomic density parameters. These parameters are fed into a physics-based model based on perturbation theory and an overlap model at long and short ranges, respectively. They lead to a remarkably small number of global parameters that only need tuning across organic compounds once and for all.

Switching to a coarse-grained resolution, the transferable Martini biomolecular force field offers a set of bead types, from which one constructs biomolecules, from proteins to lipids to sugars (Marrink and Tieleman 2013). Charged groups

interact through integer-charge Coulomb interactions. Otherwise, beads interact by means of Lennard-Jones interactions, with a predefined interaction matrix that determines the cross interactions between beads (Marrink et al. 2007). The model aims at capturing the essential thermodynamics of partitioning of chemical groups in different environments. In particular, it relies heavily on the water/octanol partitioning to assign a measure of hydrophobicity to the bead. Though not readily accessible, the water/octanol partitioning can also be predicted: ML models exist to do just that (Tetko et al. 2001). This enables a completely automated parametrization of Martini for small molecules, which both optimizes the mapping from atoms to beads and assigns the most appropriate bead type to every chemical group (Bereau and Kremer 2015). The parametrization scheme was applied to the calculation of solvation free energies for more than 1,000 compounds, clearly illustrating the potential benefits beyond manual parametrizations.

### 2.3 Example: Learning of Atomic Polarizabilities

As an illustrative example, we consider the learning of atomic polarizabilities across small organic molecules, following Bereau et al. (2018). Atomic polarizabilities are estimated using the Hirshfeld ratio, which consists of a spatial integral over the electron density of an atom in a molecule, compared to the corresponding free atom. Reference data consist of quantum chemistry calculations for thousands of isolated small molecules. We refer the interested reader to Bereau et al. (2018) for further technical details. The code to generate the data below can be found in a repository online (Bereau 2018).

We build an ML model using kernel ridge regression (Eq. 1) and encode atomic environments – the representation – using the Coulomb matrix: a pairwise matrix of inverse distances scaled by the product of atomic numbers (Rupp et al. 2012). The dataset is split between training and test sets to ensure out-of-sample predictions, thereby limiting overfitting. Figure 1a shows a learning curve: the mean absolute error (MAE) as a function of the number of atoms incorporated in the training set. The error systematically decreases with added data. Note the power law behavior. Figure 1b displays the correlation between predicted and reference Hirshfeld ratios for the rightmost ML model shown in panel a. The color-coding distinguishes between chemical elements.

## 3 Sampling

Sampling is the second corner stone of particle-based modeling in soft matter: teasing out a representative subset of conformational space is essential to extract reliable condensed-phase properties, from free energies to kinetics. The difficulty lies in assessing how much sampling is good enough. Umbrella sampling simulations are notoriously challenging, as they often hide slow conformational changes happening on degrees of freedom orthogonal to the reaction coordinate(s) (Neale et al. 2011).

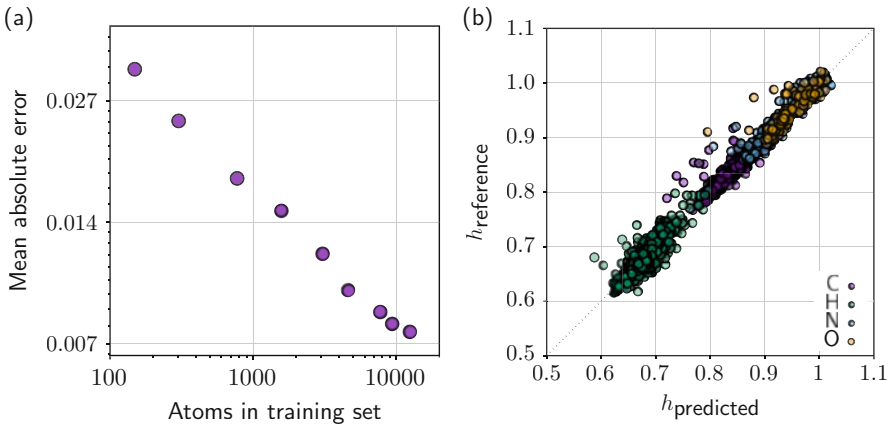


**Fig. 1** (**a**) Saturation curve of the mean absolute error (MAE) as a function of training set size for Hirshfeld ratios. Note that both axes are on a log scale. (**b**) Correlation plot of out-of-sample predictions of the Hirshfeld ratio, $h$, for the ML model with largest training set size. The different chemical elements considered are color-coded

Unsupervised machine learning techniques may help systematically improve the sampling of conformational space within an umbrella sampling protocol (Ferguson et al. 2011).

The last decade has seen a significant leap forward in timescales accessible to computer simulations. Naively, reaching longer timescales means running ever-longer single trajectories. This bodes poorly with high-performance computer clusters, because one cannot parallelize a trajectory in time. A recent paradigm shift consisted in better leveraging the statistics contained in swarms of short trajectories covering the relevant parts of conformational space. For instance, Markov state models (MSMs) discretize the simulation trajectory in conformational space and in time to analyze its long-time kinetics (Noé 2008; Bowman et al. 2013). This framework has been shown to be extremely efficient in leveraging computational resources available – from distributed computing to high-performance clusters – even at a time when dedicated hardware has significantly pushed the state of the art for long trajectories (Shaw et al. 2014). The surge in high-throughput short simulations has helped approach the sampling problem more systematically: an adaptive sampling strategy spawns new simulations from poorly populated regions of conformational space, until convergence is found. Examples include protein-protein interactions (Plattner et al. 2017) and intrinsically disordered proteins (Kukharenko et al. 2016). These ideas rely on a simple concept: it is often easier to locally equilibrate highly diverse seed conformations than waiting for a single trajectory to cross all relevant barriers. This can be extended to a multiscale approach, in which relevant snapshots from computationally efficient CG trajectories are backmapped to provide these seed conformations at the atomistic level. This strategy can help cut down the computational investment of free-energy calculations

by more than tenfold (Menichetti et al. 2017b). Analogously, a more data-driven alternative has been proposed that tries to *extrapolate* possibly interesting new seed conformations from unsupervised machine learning techniques (Chiavazzo et al. 2017).

The discussion so far has focused on sampling conformational space. Recent developments in materials discovery are aiming at exploring chemical compound space – the diversity of chemical compounds – to extract thermodynamic properties. From a simulation perspective, this poses significant challenges due to the compounded issue of sampling both across conformational and chemical compound space. While unattainable at an atomistic resolution for the foreseeable future, coarse-graining can help address this: high-throughput coarse-grained simulations provide an ensemble study of the PMF for the insertion of solute molecules in a lipid bilayer. The study both predicted PMFs for more than 450,000 compounds and identified novel linear relationships between bulk measurements and features of the PMF (Menichetti et al. 2017a).

## 4 Analysis

Everything mentioned so far has focused on improving the quality of computer simulations by improving the force field or the sampling. This section instead consists of extracting insight or information from an *existing* simulation. Advanced data-driven and statistical methodologies have helped develop more robust methods to analyze computer simulations.

Some of the most interesting developments in the analysis of computer simulations have come from approaching the very concept of probability in a new way. The traditional approach to probability theory – the one taught most often at an elementary level – is so-called *frequentist*. It interprets probability from the frequency or propensity of an event to occur. Complementary to this is the Bayesian perspective: how can one infer a reasonable expectation given limited data and/or prior belief? It offers an elegant framework to evaluate the probability of a model $M$, when dealing with limited data $D$, as illustrated by Bayes' theorem:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \tag{4}$$

where $P(M|D)$ and $P(D|M)$ are coined the posterior and the likelihood, respectively. Bayes' theorem has shown extremely useful because while the posterior is typically difficult to evaluate directly, the likelihood is often easier. In addition, it highlights the concept of prior information by means of $P(M)$, which encodes external information we may already hold on the validity of a model. For instance, physical laws and symmetries can naturally be enforced into the prior, effectively biasing the distribution of models to those that satisfy these constraints. Several examples illustrate the conceptual benefits of approaching a problem in a Bayesian framework:

- The weighted histogram method (WHAM) provides a minimum-variance estimator to best estimate the density of states of a system from different simulations (Ferrenberg and Swendsen 1989). These simulations provide complementary information to the system by encompassing a range of temperatures or different values of a collective variable in enhanced sampling. The likelihood incorporates the different Boltzmann distributions, while the prior ensures the normalization of probabilities. A derivation can be found elsewhere (Bereau et al. 2016; Ferguson 2017).
- MSMs build a discrete propagator for the time evolution of a simulation or single-molecule experiment (Noé 2008; Bowman et al. 2013). The simulation trajectory or experimental time series feeds into the likelihood, while the prior incorporates several constraints, most notably detailed balance.
- The MSM of a simulation trajectory can be further tuned to best incorporate external kinetic information. This is useful when a model is known to yield inconsistent kinetics, such as most coarse-grained models. So-called biased MSMs incorporate the coarse reference kinetic information (e.g., folding timescale or mean first passage time) as a prior, thereby selecting more consistent probabilistic models (Rudzinski et al. 2016). This conceptual framework was recently applied to incorporate experimental information to atomistic simulations (Olsson et al. 2017). More generally, the blending of physics-based models with experimental information has recently been subject to increasing interest (Perez et al. 2015, 2017).

Hidden Markov models (HMMs) add to MSMs the possibility of handling unobserved/hidden states (Rabiner and Juang 1986). While these states are not directly visible, the output, which is dependent on the state, is visible. One illustrative analogy consists of a hermit stuck inside a cave: he is attempting to forecast the weather but cannot see the sky outside. His best strategy is then to collect indirect evidence by analyzing the state of a seaweed – probabilistically related to the state of the weather – and thereby to *infer* the hidden state of the weather. HMMs can be thought as a nonlinear filtering process and have been shown to be useful in several studies, from the identification of liquid-ordered and liquid-disordered domains in lipid membrane simulations (Sodt et al. 2014) to the kinetics of protein-protein association (Plattner et al. 2017).

## 5 Outlook

Advanced data-driven methods and data-centric simulation protocols are rapidly impacting the field of soft matter and are here to stay: (i) Supervised machine learning techniques – primarily kernel methods and neural network – will likely contribute to more accurate and transferable force fields; (ii) high-throughput methods have already pushed the boundaries of conformational sampling and are likely to affect the systematic screening of compounds and materials; and (iii) Bayesian inference provides a conceptually appealing framework to combine

simulation data and physical laws and symmetries. Another notable method of interest is unsupervised machine learning, which looks for features/structure in "unlabeled" datasets, such as clustering or dimensionality reduction techniques (Fisher et al. 2014). The rapid ongoing developments of unsupervised machine learning are likely to significantly affect computer simulations in the years to come.

**Acknowledgments** Various discussions have helped shape some of the views developed in this chapter. I am especially grateful to Denis Andrienko, Kurt Kremer, Joseph F. Rudzinski, Omar Valsson, and Anatole von Lilienfeld.

This work was supported in part by the Emmy Noether Programme of the Deutsche Forschungsgemeinschaft (DFG).

## References

Bartók AP, Payne MC, Kondor R, Csányi G (2010) Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett 104(13):136403

Behler J (2016) Perspective: machine learning potentials for atomistic simulations. J Chem Phys 145(17):170901

Bereau T (2018) Example: ML model of Hirshfeld ratios. https://gitlab.mpcdf.mpg.de/trisb/handbook_example. Accessed 28 Feb 2018

Bereau T, Kremer K (2015) Automated parametrization of the coarse-grained martini force field for small organic molecules. J Chem Theory Comput 11(6):2783–2791

Bereau T, Andrienko D, von Lilienfeld OA (2015) Transferable atomic multipole machine learning models for small organic molecules. J Chem Theory Comput 11(7):3225–3233

Bereau T, Andrienko D, Kremer K (2016) Research update: computational materials discovery in soft matter. APL Mater 4(5):053101

Bereau T, DiStasio RA Jr, Tkatchenko A, von Lilienfeld OA (2018) Non-covalent interactions across organic and biological subsets of chemical space: physics-based potentials parametrized from machine learning. J Chem Phys 147(24):241706

Bowman GR, Pande VS, Noé F (Eds) (2013) An introduction to Markov state models and their application to long timescale molecular simulation, Advances in Experimental Medicine and Biology 797. Springer, Dordrecht (NL)

Chiavazzo E, Covino R, Coifman RR, Gear CW, Georgiou AS, Hummer G, Kevrekidis IG (2017) Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. Proc Natl Acad Sci 114(28):E5494–E5503

Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller KR (2017) Machine learning of accurate energy-conserving molecular force fields. Sci Adv 3(5):e1603015

Curtarolo S, Hart GL, Nardelli MB, Mingo N, Sanvito S, Levy O (2013) The high-throughput highway to computational materials design. Nat Mater 12(3):191–201

Deringer VL, Csányi G (2017) Machine learning based interatomic potential for amorphous carbon. Phys Rev B 95(9):094203

Faber FA, Hutchison L, Huang B, Gilmer J, Schoenholz SS, Dahl GE, Vinyals O, Kearnes S, Riley PF, von Lilienfeld OA (2017) Machine learning prediction errors better than DFT accuracy. arXiv e-preprints arXiv:170205532

Ferguson AL (2017) Bayeswham: a Bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method. J Comput Chem 38(18):1583–1605

Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, Kevrekidis IG (2011) Integrating diffusion maps with umbrella sampling: application to alanine dipeptide. J Chem Phys 134(13):04B606

Ferrenberg AM, Swendsen RH (1989) Optimized Monte Carlo data analysis. Phys Rev Lett 63(12):1195

Fisher DH, Pazzani MJ, Langley P (eds) (2014) Concept formation: knowledge and experience in unsupervised learning. Morgan Kaufmann Series in Machine Learning, San Mateo (CA)
Glielmo A, Sollich P, De Vita A (2017) Accurate interatomic force fields via machine learning with covariant kernels. Phys Rev B 95(21):214302
Halgren TA (1992) The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. J Am Chem Soc 114(20):7827–7843
Halgren TA, Damm W (2001) Polarizable force fields. Curr Opin Struct Biol 11(2):236–242
Huan TD, Batra R, Chapman J, Krishnan S, Chen L, Ramprasad R (2017) A universal strategy for the creation of machine learning-based atomistic force fields. npj Comput Mater 3(1):37
Huang B, von Lilienfeld O (2016) Communication: understanding molecular representations in machine learning: the role of uniqueness and target similarity. J Chem Phys 145(16):161102–161102
Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA (2013) Commentary: the materials project: a materials genome approach to accelerating materials innovation. Apl Mater 1(1):011002
John S (2016) Many-body coarse-grained interactions using gaussian approximation potentials. arXiv preprint arXiv:161109123
Kukharenko O, Sawade K, Steuer J, Peter C (2016) Using dimensionality reduction to systematically expand conformational sampling of intrinsically disordered peptides. J Chem Theory Comput 12(10):4726–4734
Li Y, Li H, Pickard FC IV, Narayanan B, Sen FG, Chan MK, Sankaranarayanan SK, Brooks BR, Roux B (2017) Machine learning force field parameters from ab initio data. J Chem Theory Comput 13(9):4492–4503
Li Z, Kermode JR, De Vita A (2015) Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. Phys Rev Lett 114(9):096405
Mackerell AD (2004) Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 25(13):1584–1604
Maple JR, Dinur U, Hagler AT (1988) Derivation of force fields for molecular mechanics and dynamics from ab initio energy surfaces. Proc Natl Acad Sci 85(15):5350–5354
Marrink SJ, Tieleman DP (2013) Perspective on the MARTINI model. Chem Soc Rev 42(16):6801–6822
Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, De Vries AH (2007) The martini force field: coarse grained model for biomolecular simulations. J Phys Chem B 111(27):7812–7824
Menichetti R, Kanekal KH, Kremer K, Bereau T (2017a) In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force. J Chem Phys 147(12):125101
Menichetti R, Kremer K, Bereau T (2017b) Efficient potential of mean force calculation from multiscale simulations: solute insertion in a lipid membrane. Biochem Biophys Res Commun. 498:282–287. https://doi.org/10.1016/j.bbrc.2017.08.095
Morawietz T, Singraber A, Dellago C, Behler J (2016) How Van der Waals interactions determine the unique properties of water. Proc Natl Acad Sci 113:8368–8373
Neale C, Bennett WD, Tieleman DP, Pomès R (2011) Statistical convergence of equilibrium properties in simulations of molecular solutes embedded in lipid bilayers. J Chem Theory Comput 7(12):4175–4188
Noé F (2008) Probability distributions of molecular observables computed from Markov models. J Chem Phys 128(24):244103
Noid W (2013) Perspective: coarse-grained models for biomolecular systems. J Chem Phys 139(9):09B201_1
Olsson S, Wu H, Paul F, Clementi C, Noé F (2017) Combining experimental and simulation data of molecular processes via augmented Markov models. Proc Natl Acad Sci 114(31):8265–8270
Perez A, MacCallum JL, Dill KA (2015) Accelerating molecular simulations of proteins using Bayesian inference on weak information. Proc Natl Acad Sci 112(38):11846–11851

Perez A, Morrone JA, Dill KA (2017) Accelerating physical simulations of proteins by leveraging external knowledge. Wiley Interdiscip Rev Comput Mol Sci 7:e1309
Peter C, Kremer K (2010) Multiscale simulation of soft matter systems. Faraday Discuss 144:9–24
Plattner N, Doerr S, De Fabritiis G, Noe F (2017) Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. Nat Chem 9:1005–1011
Ponder JW, Case DA (2003) Force fields for protein simulations. Adv Protein Chem 66:27–85
Rabiner L, Juang B (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3(1):4–16
Ramakrishnan R, von Lilienfeld OA (2017) Machine learning, quantum chemistry, and chemical space. Rev Comput Chem 30:225–256
Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning, vol 1. MIT Press, Cambridge (MA)
Rudzinski JF, Kremer K, Bereau T (2016) Communication: consistent interpretation of molecular simulation kinetics using Markov state models biased with external information. J Chem Phys 144(5):051102
Rühle V, Junghans C, Lukyanov A, Kremer K, Andrienko D (2009) Versatile object-oriented toolkit for coarse-graining applications. J Chem Theory Comput 5(12):3211–3223
Rupp M, Tkatchenko A, Müller KR, Von Lilienfeld OA (2012) Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett 108(5):058301
Schiilkopf B (2001) The kernel trick for distances. In: *Advances in neural information processing systems. Proceedings of the 2000 conference*, vol 13. MIT Press, Cambridge (MA), p 301
Shaw DE, Grossman J, Bank JA, Batson B, Butts JA, Chao JC, Deneroff MM, Dror RO, Even A, Fenton CH et al (2014) Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In: Proceedings of the international conference for high performance computing, networking, storage and analysis. IEEE Press, New Orleans, pp 41–53
Sodt AJ, Sandar ML, Gawrisch K, Pastor RW, Lyman E (2014) The molecular structure of the liquid ordered phase of lipid bilayers. J Am Chem Soc 136(2):725
Stroet M, Koziara KB, Malde AK, Mark AE (2017) Optimization of empirical force fields by parameter space mapping: a single-step perturbation approach. J Chem Theory Comput 13:6201–6212
Tetko IV, Tanchuk VY, Villa AEP (2001) Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. J Chem Inf Comput Sci 41(5):1407–1421
Tkatchenko A, DiStasio RA Jr, Car R, Scheffler M (2012) Accurate and efficient method for many-body van der Waals interactions. Phys Rev Lett 108(23):236402
Van Vleet MJ, Misquitta AJ, Stone AJ, Schmidt JR (2016) Beyond Born–Mayer: improved models for short-range repulsion in ab initio force fields. J Chem Theory Comput 12(8):3851–3870
Voth GA (2008) Coarse-graining of condensed phase and biomolecular systems. CRC Press, Boca Raton
Wang W, Donini O, Reyes CM, Kollman PA (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 30(1):211–243