AIP | APL Materials

Research Update: Computational materials discovery in soft matter

Tristan Bereau, Denis Andrienko, and Kurt Kremer

Citation: APL Mater. **4**, 053101 (2016); doi: 10.1063/1.4943287
View online: http://dx.doi.org/10.1063/1.4943287
View Table of Contents: http://scitation.aip.org/content/aip/journal/aplmater/4/5?ver=pdfcov
Published by the AIP Publishing

Research Update: Computational materials discovery in soft matter

Tristan Bereau, Denis Andrienko, and Kurt Kremer
Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany

(Received 17 December 2015; accepted 10 February 2016; published online 15 March 2016)

Soft matter embodies a wide range of materials, which all share the common characteristics of weak interaction energies determining their supramolecular structure. This complicates structure-property predictions and hampers the direct application of data-driven approaches to their modeling. We present several aspects in which these methods play a role in designing soft-matter materials: drug design as well as information-driven computer simulations, e.g., histogram reweighting. We also discuss recent examples of rational design of soft-matter materials fostered by physical insight and assisted by data-driven approaches. We foresee the combination of data-driven and physical approaches a promising strategy to move the field forward. [http://dx.doi.org/10.1063/1.4943287]

I. DATA-DRIVEN MATERIALS DESIGN

The fundamental laws of physics and chemistry provide us with constitutive laws and equations which can be used to predict material properties and to link them to the chemical composition and processing conditions. In materials design, predicting properties from chemical structures is termed as a *forward* problem (see Fig. 1). We routinely rely on experimental measurements, computer simulations, analytical theory, and statistical modeling to solve the forward problem, i.e., to map structure to function.

The design of materials, on the other hand, amounts to identifying the adequate structure *given* properties of interest.[1,2] Unfortunately, we do not have physical laws to tackle the backward problem. One could of course attempt a brute-force solution and try to solve the forward problem for as many compounds as possible. An exhaustive account, however, bears no hope: the number of drug-like small (less than about 500 atomic mass units) organic molecules alone has been estimated in the range of 10^{60}.[3,4] As a result, the empirical backward solution is in many cases driven by our chemical and physical understanding of the system.

More recently, researchers have started exploring an alternative strategy to tackle this many-compound problem. Since similar chemical structures often share a number of properties, the number of structural signatures—or chemical features—leading to significantly different properties ought to be much smaller than the number of compounds.[5–7] If so, one can solve the time-consuming forward problem for a small subset of compounds and predict the properties of the others by effectively *interpolating* across the training subset[8] (see Fig. 1). Interpolation is performed by first identifying correlations between simple chemical features, or *descriptors* (e.g., molecular mass, size, or number of hydrogen bonds) and said properties,[9] resting on the assumption that simple descriptors *can* be extracted. The correlations drawn then call for a qualitative validation based on thorough physical and chemical understanding, thereby avoiding artifacts due to noisy or erroneous data. Ultimately, this yields an *empirical* backward optimization route, in which the desired property can be achieved by tuning the chemistry along the descriptor (see magenta in Fig. 1).

Given that the correlations between structural descriptors and chemical properties are often nonlinear,[10] interpolation schemes call for advanced statistical methodologies, e.g., *machine learning*. Machine learning consists of a set of classification and regression schemes whose predictive performance *improves* with increasing training set size.[11,12] This motivates the development of solutions that can faster explore portions of chemical space by reducing the computational investment associated

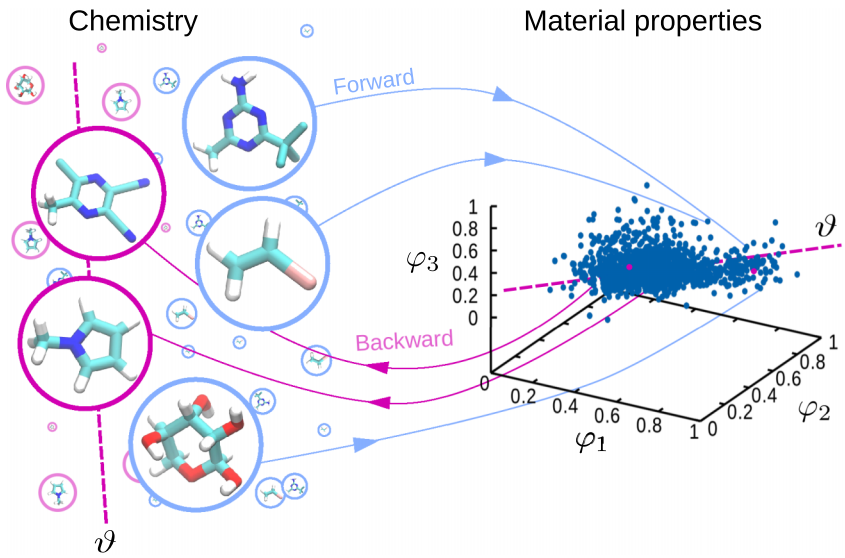

FIG. 1. Schematic link between chemical structure (left) and material properties (right). The forward process (colored in blue) consists of evaluating—via experiments, simulations, analytical theory, or statistical modeling—a number of material properties φ_i for a chemical structure. The identification of simple descriptors, ϑ, which link the chemistry to material properties, enables an empirical backward optimization, highlighted here in magenta.

with each compound. Two main avenues have traditionally prevailed: technological improvements of computer hardware[13] and the development of more efficient algorithms.[14,15]

In contrast to the early successes of machine learning to predict electronic properties of molecules,[16–18] phenomena where thermal fluctuations play an essential role have proven much more challenging. Soft matter systems, for which the characteristic interaction energies are comparable to thermal energy, can lead to fluctuations over a wide range of length (~nm to μm) and time (~ps to s) scales.[19] As a result, solving the forward problem is strongly limited by computational resources: the ability to compute or simulate large systems over long time scales mirrors experimental challenges to probe increasingly small and fast processes.[20–25] The present perspective discusses the challenges associated with designing soft-matter materials using computational tools. We highlight two sources of difficulties: (i) the accurate prediction of thermodynamic properties and (ii) the identification of clear descriptors for backward optimization. These difficulties are so severe that computational materials design has made relatively little progress in soft matter compared to other systems, e.g., thermoelectrics,[26] crystal-structure prediction,[27,28] or electrocatalytic materials.[29] The present review aims at bridging two largely disconnected fields, namely, computational materials design and soft matter, expected to rapidly grow closer with hardware and methodological developments. We illustrate the issues associated with predicting soft-matter materials by means of examples in the field of computational drug design as well as provide examples of rational design guided by our understanding of the physics and chemistry, assisted by data-mining. We finally conclude by proposing an outlook on data-driven methods in soft matter.

II. WHY IS SOFT MATTER DIFFICULT TO PREDICT?

A. Statistical mechanics and computer simulations

To better understand the challenges associated with predicting properties of soft-matter systems, we first recall basic aspects of statistical mechanics and computer simulations of molecular systems.

While quantum and classical physics provide laws to describe the evolution of all degrees of freedom in the system, large systems—on the order of 10^{23} and more molecules—call for coarser approaches. Statistical mechanics provides a framework to link macroscopic parameters with the

microscopic details, or ensemble of states—the probability distribution of this collection of states is called statistical ensemble. In what follows, we only consider *equilibrium* ensembles, i.e., those with no explicit time dependence in the phase-space distribution.

Fixing different macroscopic parameters will give rise to specific statistical ensembles. For instance, the canonical ensemble describes a mechanical system at fixed temperature, volume, and number of particles. The probability of sampling microstate i at temperature T is then proportional to the Boltzmann distribution

$$p_i = \frac{\mathrm{e}^{-\beta E_i}}{Z}, \tag{1}$$

where $\beta^{-1} \equiv k_\mathrm{B}T$, k_B is Boltzmann's constant, E_i is the energy of microstate i, and Z—the partition function—normalizes the probability. The canonical ensemble thus weighs each microstate according to the corresponding Boltzmann factor. Rather than studying a single microstate, one typically probes a (large) subset $\mathcal{S}$ that represents specific values of coarse-grained coordinates of the system, e.g., the folded state of a protein. In this case, the *stability* of this subset is often described in terms of its (projected) free energy

$$F(\mathcal{S}) = -k_\mathrm{B}T \ln p_\mathcal{S} = -k_\mathrm{B}T \ln \sum_{j\in\mathcal{S}} p_j, \tag{2}$$

where $\mathcal{S}$ encapsulates the collection of microstates corresponding to the subset of interest. The free-energy difference between different subsets, $\mathcal{S}_1$ and $\mathcal{S}_2$, provides a measure of their relative stability: $\Delta F = F(\mathcal{S}_1) - F(\mathcal{S}_2)$.

Practically, evaluating Eq. (2) requires the energy of every microstate of the system, which then permits a direct evaluation of observables at a particular state point. Unfortunately, the sheer number of microstates is often colossal, making these configurational integrals intractable for all but the simplest of systems.

Instead of exhaustively enumerating microstates, computer simulations tackle the problem by *sampling* the most relevant regions of the statistical ensemble. How can one sample efficiently out of an enormous pool of microstates? A simple but extremely efficient procedure draws microstates according to the Boltzmann distribution itself (Eq. (1)). This importance sampling scheme has given rise to the two main types of molecular simulations: Markov chain Monte Carlo with Metropolis update and molecular dynamics.[30,31]

Both algorithms rely on an evaluation of the system's energies for each microstate. In soft matter, the system is often modeled by a multi-dimensional potential energy surface, expressed in terms of interatomic interaction potentials that need to be parametrized (i.e., the so-called force field). A Monte Carlo simulation samples states by proposing a new trial system configuration at each step. It then evaluates the energy difference between subsequent configurations and finally accepts or rejects the trial configuration based on the Boltzmann factor of this energy difference. Molecular dynamics, on the other hand, numerically integrates the classical equations of motions of the system according to the interatomic forces (i.e., spatial derivatives of the interaction energies).

The predictability of (classical) computer simulations depends on three main aspects: the modeling of the relevant physics, the accuracy of the force fields, and the simulation time. We avoid discussions of the first topic—entirely system-dependent—to focus on the other two. The ability of force fields to accurately represent the energy landscape is of critical importance. They are thus subject to constant refinement.[32–34] Collaborative open-data approaches to improvements, testing, and validation of force fields provide a novel framework to refine force fields[35] and could later take advantage of a data-driven approach. Beyond accuracy, force-field-based methods rely on their ability to quickly evaluate the energy or forces of a configuration, thus sampling a representative ensemble of configurations. Failure to thoroughly sample the statistical ensemble easily leads to drastic errors and artifacts in the derived thermodynamic quantities.[36] The simulation time achievable for a given system is inherently limited by the computational power at hand. The last three decades have yielded a roughly exponential increase available in simulation time for fixed system size.[15] Though the most straightforward approaches to molecular dynamics would scale computational power with the square of the number of atoms, N^2 (i.e., commensurate with

the number of interactions), the use of interaction cutoffs and neighbor lists can reduce the scaling to $N \ln N$, typical of long-range electrostatics algorithms based on fast-Fourier transforms.[30] While the performance achieved strongly depends on many parameters (e.g., hardware, algorithms, or system), several benchmarks report values in the 10 to 100 ns/day for solvated proteins with $N \sim 10^4$–10^5 using roughly 10^2–10^3-CPU cores.[37,38] Dedicated hardware, which is designed specifically to run molecular dynamics, has shown significant improvements over more common high-performance-computing solutions.[34,39,40] The recent development of graphics processing units (GPUs) for scientific computing has provided wider access to significant computational performance, as a modern GPU can emulate the performance of 10–100 CPU cores.[38,41] Boosting computational performance of a single simulation using parallelization inherently requires large systems, as the system is divided spatially in smaller units. A recent alternative consists of following swarms of independent simulation trajectories and gathering statistics, which has helped analyze complex molecular processes both in terms of the equilibrium populations and kinetics.[42–45]

B. Example: Protein-ligand binding in computational drug design

One of the most important applications of soft-matter materials prediction is drug design.[46–48] Drugs are (small) molecules that alter certain biochemical pathways by interacting with specific macromolecules, e.g., proteins. The binding of a ligand to a protein pocket will stabilize certain conformational states of the macromolecule, thereby affecting its biological function. Strong binding is thus primordial to ensure significant interactions between the two entities at reasonable ligand concentrations. We note that although other physiological aspects play a major role—e.g., specificity to a single protein receptor, permeability through cell membranes, or time before the drug gets metabolized[49]—we specifically focus on protein-ligand binding.

Stable binding, in terms of the abovementioned statistical mechanics formalism, corresponds to a significant free-energy difference between the bound complex and the weakly (or non-)interacting dissociated pair, see Fig. 2(a). What does this imply at the microscopic level? Remember that free energies average over the individual Boltzmann-weighted energies of many microstates (Eq. (2)). In terms of sheer numbers of microstates, the subset of conformations that describe the bound state will be extremely small compared to the vast number of geometries describing two unbound molecules. Given the form of the Boltzmann factor, the free-energy difference can only favor the bound state if its few microstates have significantly lower energies compared to the others. The competition between the weight of each microstate and their number—energy and entropy, respectively—completely dictates the performance of a ligand.

Practically, optimizing protein-ligand binding is a formidable challenge: not only is the number of putative drug candidates tremendous (10^{60} or so drug-like small molecules[3]), but a careful characterization of any one of them also requires the evaluation of extensive sums (Eq. (2)). Classical all-atom computer simulations can provide accurate estimates of the binding free energy (i.e., down to $\approx 1\ k_B T$ error),[50] at the expense of significant computational investment: up to 10–100 compounds in a single study, at most.[50,51] This strongly limits the number of trial compounds that can be simulated, such that computer simulations are more typically applied to optimize an already-identified promising molecular scaffold. Identifying the said scaffold instead requires alternative methods.

Given difficulties in accurately simulating binding events for many compounds, virtual high-throughput screening efforts (i.e., which test many compounds) have so far primarily employed *statistical modeling*.[52] The prediction of protein-ligand binding affinity given a protein receptor and the chemical structure of a ligand relies on correlations between physical descriptors and binding stability.[7] Establishing correlations requires large datasets to train the statistical models. Given the scarcity of computer simulations, most models are trained against experimental measurements. The predictability of these models has remained extremely limited: typical benchmarks report linear correlation coefficients between predicted and experimentally measured binding affinities around 50%–60%.[53] As such, statistical scoring fails to efficiently filter out most ligands to focus on key promising candidates. Ideally, high predictability would ensure a reliable identification of few compounds that would be worth either simulating using accurate force fields or synthesizing in the

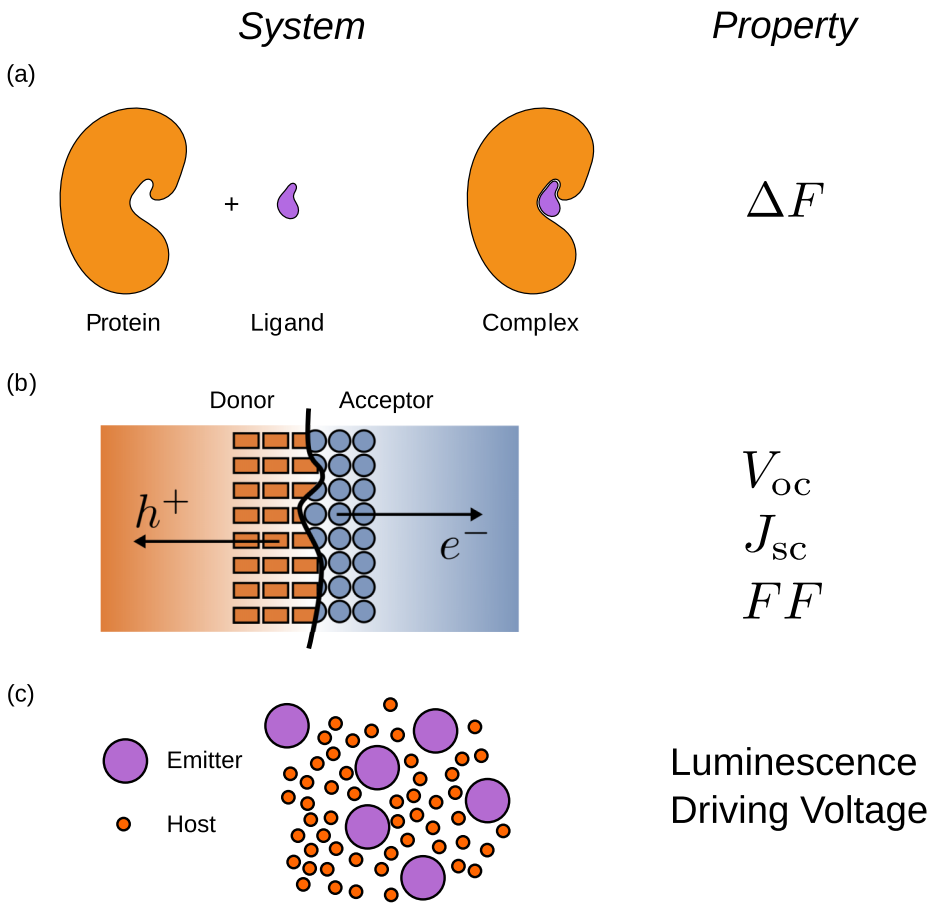

FIG. 2. Cartoon representations of several examples presented in this review: (a) the competition between the unbound protein-ligand system and the complex; (b) a donor/acceptor interface for solar-cell design; and (c) emitter-host interactions in an organic light-emitting diode. Common properties of interest are denoted to the right.

laboratory. A number of reasons can be brought forward for the limited success of statistical scoring methods:

- Experimental measurements carry errors that may strongly vary between different protocols and can be significant compared to the binding affinity itself.
- Statistical models rely on examples to learn why certain ligands will bind to a protein. On the other hand, very few compounds will bind significantly to any protein receptor, significantly skewing the amount of information toward poor binders. These low hit rates hinder the training process.
- The ligands used in the training set amount to synthesized compounds available in the laboratory. These libraries are not only extremely sparse compared to the size of chemical space (i.e., 10^3–10^5 vs. 10^{60}); available compounds tend to be extremely correlated and fail to represent the overall diversity of compound chemistry.[54] As a result, scoring methods that are trained against these datasets typically show poor transferability between protein receptors and classes of ligands.
- Statistical scoring approaches, though often employing advanced methodologies (e.g., machine learning), aim at directly correlating ligand and receptor geometry with binding free energy. Critically, the inclusion of explicit thermodynamics—and most importantly entropic contributions—appears most challenging.

These points highlight the difficulties in correlating geometrical and chemical aspects of soft-matter systems to their thermodynamics. Recently, efforts have turned to deep-learning algorithms to

leverage information from many distinct biological sources: multitask networks that simultaneously predict binding against different protein receptors improve the performance, as compared to independent models.[55]

Finally, we note that beyond mere binding constants, computer simulations provide an atomic description of the binding modes and pathways,[45,56–58] providing the means to a rational design of drugs without tens of thousands of training samples, but rather using chemical insight.[39,40] Unlike traditional statistical scoring methods, computer simulations also provide receptor flexibility to fully take into account the diversity of binding poses and associated entropic contributions.[59] For a more detailed review on the statistical mechanics of protein-ligand binding, see the work of Wereszczynski and McCammon.[60]

III. ADVANCED STATISTICAL TECHNIQUES IN SOFT-MATTER SIMULATIONS

The challenges associated with accurately evaluating thermodynamic quantities (Eq. (2)) of soft-matter systems have limited the advancement of data-driven predictions of new materials. Instead, other types of advanced statistical methods have already provided enormous help in extracting information from computer simulations. Rather than interpolating across many instances of the forward process (Fig. 1), these methods focus on improving the quality of individual samples. In the following, we briefly outline a few examples of data-driven methods that augment computer simulations through enhanced analysis or sampling.

A. Analysis of multiple thermodynamic states

Molecular dynamics or Metropolis Monte Carlo simulations that generate a canonical ensemble will primarily sample microstates that are within a few k_BT of the local energetic minimum. The system may hop between distinct local minima, as long as there are no significant free-energy barriers in between, otherwise the system may remain trapped for long times. Enhanced-sampling methods address this problem by altering the distribution function used in generating microstates.[61,62] At its simplest, increasing the temperature $T^{(0)} \rightarrow T^{(1)}$ will provide a wider exploration of conformational space, thanks to more thermal energy. On the other hand, analyzing the higher-temperature simulation $T^{(1)}$ will yield thermodynamic quantities for that particular thermodynamic state, which may not be of interest. Instead, one seeks the evaluation of thermodynamic quantities at temperature $T^{(0)}$ using simulations run at elevated temperatures—systematically disentangling the thermodynamic states used for sampling and analysis.

It helps to consider the canonical probability distribution (Eq. (1)) in terms of energies rather than states,

$$p(E) = \frac{\Omega(E)\mathrm{e}^{-\beta E}}{Z}, \tag{3}$$

which highlights the density of states, $\Omega(E)$. $\Omega(E)$ is a material property of the system—it does not depend on its thermodynamic state. Determination of the density of states provides access to all thermodynamic quantities. Given an infinitely long simulation, one could simply divide out the Boltzmann factor to isolate $\Omega(E)$. In practice, the exponentially suppressed tails of a canonical distribution will limit a reliable estimation of the density of states to a narrow energy interval. On the other hand, a simulation run at different temperatures will provide partial—but complementary—information about $\Omega(E)$: low (high) temperatures will preferentially sample the low (high) energy regions of the density of states. A series of simulations run at different temperatures will each contribute to estimating the density of states in different energy ranges. Given sampling difficulties, leveraging the information contained in all simulations can dramatically help accurately estimating $\Omega(E)$.

Effectively, we seek a statistical model that best reproduces the data provided S canonical simulation trajectories at different inverse temperatures $\beta^{(j)} = (k_BT^{(j)})^{-1}$. Each simulation j records the distribution of energies sampled by means of histograms, $H^{(j)}$, each of size N. Since each histogram represents a canonical distribution, the associated (unknown) probability of sampling energy

E_i is given by $p_i^{(j)} = \Omega_i \exp(-\beta^{(j)} E_i)/Z^{(j)}$, where $\Omega_i \equiv \Omega(E_i)$. Finding the set of probabilities $p_i^{(j)}$ (i.e., optimizing $\Omega(E)$) that best reproduce the data $H^{(j)}$ corresponds to maximizing the conditional probability $P(p^{(j)} \mid H^{(j)})$. From Bayes' theorem, we obtain

$$P\left(p^{(j)} \mid H^{(j)}\right) = \frac{P(H^{(j)} \mid p^{(j)}) P(p^{(j)})}{P(H^{(j)})}, \tag{4}$$

where the three terms in the numerators are denoted posterior, likelihood, and prior. We drop the denominator, as it simply contributes a normalizing factor. In the following, we further consider all models equally likely—i.e., we apply a constant prior—such that $P(p^{(j)} \mid H^{(j)}) \propto P(H^{(j)} \mid p^{(j)})$. Though the posterior is, in general, difficult to evaluate directly, we can express the likelihood as a multinomial distribution (assuming no correlation between histogram counts)

$$P\left(H^{(1)} \cdots H^{(S)} \mid p^{(1)} \cdots p^{(S)}\right) \propto \prod_{k=1}^{S} \prod_{i=1}^{N} \left(p_i^{(k)}\right)^{H_i^{(k)}}. \tag{5}$$

In other words, the model that best reproduces the data would maximize the likelihood given by the set of probabilities applied to the content of the sampled histograms. Maximizing the likelihood function is more easily performed by working with its logarithm. Further, we add Lagrange multipliers, $\lambda^{(k)}$, to constrain the resulting models to yield normalized probabilities. The resulting log-likelihood function reads

$$\mathcal{L} = \sum_{k=1}^{S} \sum_{i=1}^{N} H_i^{(k)} \ln\left(\Omega_i \mathrm{e}^{-\beta^{(k)} E_i - f^{(k)}}\right) + \sum_{k=1}^{S} \lambda^{(k)} \left(1 - \sum_{i=1}^{N} \Omega_i \mathrm{e}^{-\beta^{(k)} E_i - f^{(k)}}\right), \tag{6}$$

where $f^{(j)} = \ln Z^{(j)}$. Maximizing $\mathcal{L}$ is obtained by setting its derivatives against $f^{(j)}$, $\lambda^{(j)}$, and Ω_i to 0, which yields the set of self-consistent equations

$$\mathrm{e}^{f^{(k)}} = \sum_{i=1}^{N} \Omega_i \mathrm{e}^{-\beta^{(k)} E_i}, \tag{7}$$

$$\Omega_i = \frac{\sum_{k=1}^{S} H_i^{(k)}}{\sum_{j=1}^{S} N^{(j)} \mathrm{e}^{-\beta^{(j)} E_i - f^{(j)}}}, \tag{8}$$

and $N^{(k)} = \sum_i H_i^{(k)}$. Eq. (8) provides a minimum variance estimator for the density of states from the S simulation trajectories. This weighted histogram analysis method[63,64] inscribes itself within a series of methods known as extended bridge sampling estimators, reviewed by Shirts and Chodera.[65] These methods have provided valuable assistance to help analyze the equilibrium thermodynamics of complex (bio)molecular systems.[66–68]

B. Markov State models

Beyond static equilibrium properties, computer simulations aim at a kinetic understanding of molecular processes. Solute-solvent reorganization, protein-ligand binding, and protein folding all operate at different time scales.[69] Among the many ways one can analyze the kinetics of a molecular system, Markov state models systematically link microscopic transitions with the slow processes of the system (e.g., the folding of a protein or the flip-flop rate of a lipid in a bilayer).[44,70] These kinetic models describe the evolution of a system in terms of microstate transitions under a Markovian (i.e., memoryless) assumption. The construction of these Markov state models typically corresponds to a Bayesian framework similar to the one presented above: a probabilistic model of the transition probability matrix is optimized to best reproduce the underlying simulation trajectory, i.e., a count matrix of state-to-state transitions. Though normalizing this count matrix should formally yield the transition probability matrix, finite sampling often leads to inconsistencies between forward and backward microscopic transitions—violating detailed balance. Enforcing detailed balance in the transition probability matrix greatly helps alleviating finite-sampling limitations. The ability to

consistently reconstruct robust kinetic models from several simulations has opened a new paradigm in computer simulations: extremely long time scales (up to ms time scale) may be reconstructed from many short simulations that collectively sample the relevant part of conformational space.[24,43]

C. Enhanced sampling from Bayesian inference

Finally, we mention a recently proposed enhanced sampling scheme based on Bayesian inference that balances contributions from a computer simulation and structural biology data. MacCallum *et al.*[71] sample the system from a biased energy function, $E_{\text{bias}} = -\beta^{-1} \ln P(\mathbf{x} \mid \mathbf{D})$, which corresponds to a probabilistic model of configuration $\mathbf{x}$ that best reproduces external data $\mathbf{D}$. From Bayes' theorem (Eq. (4)), they express E_{bias} in terms of a prior, which corresponds to the Boltzmann-weighted energy as obtained from the force field, while the likelihood incorporates ambiguous or uncertain experimental information, $\mathbf{D}$, by means of restraints. The combination of physical models with experimental information efficiently leads to the correct structure of several proteins. We emphasize here that while experimental information can help drive the simulation toward relevant conformations, their selection relies on the accuracy of the (physics-based) force field.

D. Machine learning in molecular simulations

Though not yet widely used, machine learning has made a number of contributions to assist the construction, execution, and analysis of molecular simulations. The following exclusively focuses on the link between the two fields. Machine learning models have helped parametrize molecular force fields for small molecules of both coarse-grained models[72] and static electrostatic multipole coefficients.[73] Forces computed from a simulation model can be augmented by machine-learned quantum-mechanical calculations, though such an application remains to be reported for soft-matter systems.[74–76] Finally, data-driven methods have been proposed to automatically detect molecular patterns[77] and transition-state dividing surfaces.[78]

IV. RATIONAL DESIGN BASED ON PHYSICAL/CHEMICAL UNDERSTANDING

We now illustrate how cheminformatic concepts can be applied to design materials. Our goal here is to use a particular case study in order to exemplify soft-matter-related issues: the complexity of the forward problem, the limited accuracy of *in silico* predictions, and small training set sizes available from experimental sources.

The specific case study we are going to discuss is an organic solar cell. In these optoelectronic devices, light absorption leads to the generation of excited, strongly bound electron-hole pairs, or excitons. Excitons dissociate into free charge carriers at interfaces between two materials, one of which is the electron donor and the other one is the acceptor. Free charges then diffuse towards the electrodes and are injected into the external circuit. The main task of the compound design algorithm is to identify pairs of donor/acceptor molecules maximizing solar cell efficiency, as depicted in Fig. 2(b).

The straightforward cheminformatic approach would be to collect available experimental data and to correlate donor-acceptor pairs with solar cell efficiencies. In practice, such a brute-force strategy never works. First, a single figure of merit, such as solar cell efficiency in our case, is too complex to correlate directly to the chemical structure. Second, the number of experimental samples is normally limited to a few hundreds, which is not enough to train the model or refine the descriptors.

A thorough physical understanding of the problem is key to overcome the first issue: by being familiar with all elementary processes occurring in an organic solar cell, we can factorize its efficiency on several experimentally accessible quantities, such as the open circuit voltage V_{oc}, short circuit current J_{sc}, and fill factor FF. These quantities have been used to train a model of 33

descriptors on approximately 50 compounds.[79] It has been found that V_{oc} correlates well with the gas-phase ionization potential of the donor.

Since the provided descriptors were trained only on a small set of donors, the model could not be used to predict efficiencies of compounds outside the training set. In soft matter systems, synthesis, device optimization, and characterization are often time consuming; hence, substantially extending the training set is not feasible to do experimentally. This motivates using computer simulations and to solve the forward problem *in silico*.

For organic solar cells, the simulation strategy for solving the forward problem is, in principle, well-defined: One needs to predict the materials morphology, calculate the energetic landscape for charges and excitons, evaluate rates of charge/exciton transfer, electron-hole, and exciton-electron recombination, solve the time-dependent master equation, and analyze distributions of currents and occupation probabilities to extract/optimize measurable properties such as open-circuit voltage, short circuit current for solar cells.[80,81] As formulated, this roadmap is practically impossible to implement. Already the prediction of the materials morphology of partially crystalline systems is not feasible from first principles. Moreover, one is forced to use multiscale approaches: On a molecular scale, every molecule has its own unique environment created by its neighbors with local electric fields leading to level shifts, broadening, and spatial correlations of charge/exciton energies. On a mesoscopic scale, the size of phase-separated domains of the donor and acceptor governs the efficiency of exciton splitting and charge percolation. On a macroscopic scale, light in-coupling needs to be accounted for to maximize absorption. Likewise, the typical time scales of dynamic processes such as charge and energy transfer span several orders of magnitude. Hence, charge/exciton kinetics cannot be treated via numerical methods with a fixed time step, but rate-based descriptions must be employed instead. Finally, to be predictive, the methods need to be quantitative. That is, not only each particular method is required to provide accurate values for the quantities of interest but also the entire procedure of "bottom-up scaling" from the micro- to the macroscopic world should be robust.

At this level of complexity, simulations become as demanding as experiments. The benefit is of course that we learn a lot more details, especially at a molecular scale. This, however, is possible only for a few selected systems, e.g., P3HT:PCBM or DCV:C60 bulk heterojunction solar cells[82,83] and does not help us to extend the training set or to train a better model. As a compromise, one often adopts simplified solutions to the forward problem. Simplified schemes are, however, prone to systematic errors, since they are normally constructed phenomenologically or employ assumptions which are not valid for all systems.[84] These models can be used to formalize our understanding of experimental systems but again, they are incapable of predicting properties of new systems.

Hence, cheminformatics seems to be not a very useful tool when it comes to problems with a complex forward problem, such as soft-matter systems. There is, however, a hidden bonus: data-driven approaches help to gain better physical and chemical understanding of the system. Indeed, in this particular example, they helped to establish a link between the optical gap in a solid state, which determines the amount of the harvested solar spectrum, and the gas-phase ionization potential of the donor. This correlation is of course rather trivial: for chemically similar compounds, the optical gap and the ionization energy minus electron affinity do correlate, at least in the gas phase. Analyzing this correlation further, we can conclude that it ignores the electron-hole interaction and environmental effects. These can then be added by using additional descriptors, e.g., molecular quadrupole moment[83,85] or the packing motif of the crystal structure. Hence, we have gained valuable information as to what (supra)molecular quantities are relevant to refine or simplify the solution of the forward problem.

Design strategy for donor/acceptor combinations for solar cells is of course not an isolated example: pre-screening of host-guest systems in organic light emitting diodes (OLEDs, see Fig. 2(c)) has evolved in a similar manner.[86] Here, simple molecular descriptors are indeed used to rank the compounds of interest,[87] but the main simulation effort is dedicated to solving the forward problem (in this case predicting current-voltage-luminescence characteristics of an OLED) for a handful of systems.[88]

Thus, by identifying the set of relevant descriptors, cheminformatics has helped us to better understand the physics of the system, to link it to molecular properties (chemistry), and to build and

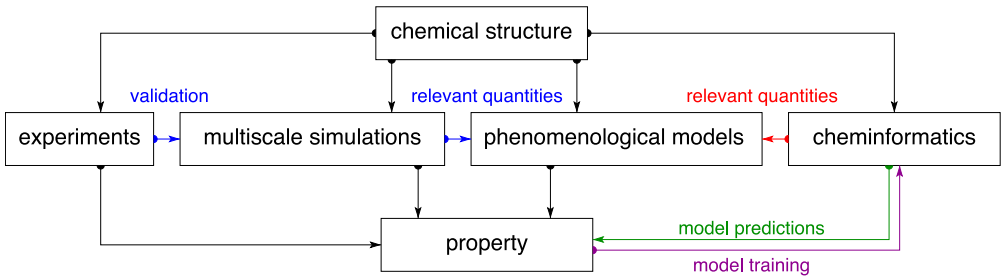

FIG. 3. Role of cheminformatics in the soft matter compound design: identification of relevant descriptors for phenomenological models and (eventually) prediction of material properties for large training sets.

design a better model for solving the forward problem, as schematically illustrated in Fig. 3. Even though this does not necessarily lead to a prediction of new compounds, it does contribute to our physical and chemical understanding of the system and hence generates new ideas and models.

V. OUTLOOK

Computational materials design of soft-matter systems faces two problems: First, the computational investment necessary to accurately estimate thermodynamic properties from simulations remains significant. This hinders the number of putative materials one can study and has so far prevented the development of datasets of simulation results to build machine-learning models. Second, the complexity of soft-matter systems prevents the identification of few, relevant descriptors. Instead, these materials typically exhibit complex behavior at many length and time scales, due to their small characteristic energy, as well as experimental characterization issues, e.g., polydispersity or stereoregular defects. Rational design is inherently challenging due to the many variables soft-matter systems depend on.

We foresee that the onset of high-throughput computer simulations—enabled by both hardware (e.g., GPUs) and methodological (e.g., coarse-graining) advancements—will lead to data-driven approaches that better balance energetic and entropic contributions to thermodynamics. In the meantime, we showed that data-driven approaches already augment computer simulations by improving different aspects (e.g., parametrization, sampling, or analysis). This combination of data and physics-based models makes for more than the sum of its parts. Setting aside our understanding of physics and chemistry to make ways for statistics alone would be wasteful: machine learning models merely interpolate training sets—learning the complexity of these systems would require colossal amounts of data to correlate behavior our laws of physics can anyway predict. Probabilistic models of simulation trajectories (e.g., histogram reweighting, Markov state models) help us enforce laws of physics we already know, e.g., detailed balance in equilibrium systems, to help alleviate finite-statistics issues. Finally, data-driven approaches can be extremely beneficial when our understanding of physics and chemistry is lacking.

ACKNOWLEDGMENTS

We thank Aviel Chaimovich and Joseph F. Rudzinski for a critical reading of the manuscript. Funding from the Grant No. SFB-TRR146 of the German Research Foundation (DFG) is gratefully acknowledged.

[1] A. Jahan, M. Ismail, S. Sapuan, and F. Mustapha, "Material screening and choosing methods—A review," Mater. Des. **31**, 696–705 (2010).
[2] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, "Combinatorial and high-throughput screening of materials libraries: Review of state of the art," ACS Comb. Sci. **13**, 579–633 (2011).
[3] R. S. Bohacek, C. McMartin, and W. C. Guida, "The art and practice of structure-based drug design: A molecular modeling perspective," Med. Res. Rev. **16**, 3–50 (1996).
[4] C. M. Dobson, "Chemical space and biology," Nature **432**, 824–828 (2004).

[5] C. Lipinski and A. Hopkins, "Navigating chemical space for biology and medicine," Nature **432**, 855–861 (2004).
[6] M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, and H. Waldmann, "Charting biologically relevant chemical space: A structural classification of natural products (SCONP)," Proc. Natl. Acad. Sci. U. S. A. **102**, 17272–17277 (2005).
[7] P. M. Dean, *Molecular Similarity in Drug Design* (Springer Science and Business Media, 2012).
[8] T. I. Oprea and J. Gottfries, "Chemography: The art of navigating in chemical space," J. Comb. Chem. **3**, 157–166 (2001).
[9] R. E. Newnham, *Structure-Property Relations* (Springer Science and Business Media, 2012), Vol. 2.
[10] G. M. Maggiora, "On outliers and activity Cliffs why QSAR often disappoints," J. Chem. Inf. Model. **46**, 1535 (2006).
[11] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, 2005).
[12] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).
[13] H. Meuer, E. Strohmaier, J. Dongarra, and H. Simon, Top500 supercomputing sites, 2011.
[14] S. H. Fuller and L. I. Millett, "Computing performance: Game over or next level?," Computer **44**, 31–38 (2011).
[15] M. Vendruscolo and C. M. Dobson, "Protein dynamics: Moore's law in molecular biology," Curr. Biol. **21**, R68–R70 (2011).
[16] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, "Finding density functionals with machine learning," Phys. Rev. Lett. **108**, 253002 (2012).
[17] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).
[18] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, "Assessment and validation of machine learning methods for predicting molecular atomization energies," J. Chem. Theory Comput. **9**, 3404–3419 (2013).
[19] P.-G. De Gennes and J. Badoz, *Fragile Objects: Soft Matter, Hard Science, and the Thrill of Discovery* (Springer Science and Business Media, 2012).
[20] T. Schlick, R. D. Skeel, A. T. Brunger, L. V. Kalé, J. A. Board, J. Hermans, and K. Schulten, "Algorithmic challenges in computational molecular biophysics," J. Comput. Phys. **151**, 9–48 (1999).
[21] B. Palsson *et al.*, "The challenges of in silico biology," Nat. Biotechnol. **18**, 1147–1150 (2000).
[22] K. M. Merz and B. Roux, *Biological Membranes: A Molecular Perspective from Computation and Experiment* (Springer Science and Business Media, 2012).
[23] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw, "Biomolecular simulation: A computational microscope for molecular biology," Ann. Rev. Biophys. **41**, 429–452 (2012).
[24] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, "To milliseconds and beyond: Challenges in the simulation of protein folding," Curr. Opin. Struct. Biol. **23**, 58–65 (2013).
[25] J. R. Perilla, B. C. Goh, C. K. Cassidy, B. Liu, R. C. Bernardi, T. Rudack, H. Yu, Z. Wu, and K. Schulten, "Molecular dynamics simulations of large macromolecular complexes," Curr. Opin. Struct. Biol. **31**, 64–74 (2015).
[26] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli *et al.*, "Aflowlib.org: A distributed materials properties repository from high-throughput *ab initio* calculations," Comput. Mater. Sci. **58**, 227–235 (2012).
[27] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, "Predicting crystal structures with data mining of quantum calculations," Phys. Rev. Lett. **91**, 135503 (2003).
[28] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," Phys. Rev. B **89**, 094104 (2014).
[29] J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff, and J. K. Nørskov, "Computational high-throughput screening of electrocatalytic materials for hydrogen evolution," Nat. Mater. **5**, 909–913 (2006).
[30] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 1989).
[31] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, 2014).
[32] R. B. Best, N.-V. Buchete, and G. Hummer, "Are current molecular dynamics force fields too helical?," Biophys. J. **95**, L07–L09 (2008).
[33] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic validation of protein force fields against experimental data," PLoS One **7**, e32131 (2012).
[34] S. Piana, J. L. Klepeis, and D. E. Shaw, "Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations," Curr. Opin. Struct. Biol. **24**, 98–105 (2014).
[35] A. Botan, F. Favela-Rosales, P. F. Fuchs, M. Javanainen, M. Kanduc, W. Kulig, A. Lamberg, C. Loison, A. Lyubartsev, M. S. Miettinen *et al.*, "Toward atomistic resolution structure of phosphatidylcholine headgroup and glycerol backbone at different ambient conditions," J. Phys. Chem. B **119**, 15075–15088 (2015).
[36] C. Neale, W. D. Bennett, D. P. Tieleman, and R. Pomès, "Statistical convergence of equilibrium properties in simulations of molecular solutes embedded in lipid bilayers," J. Chem. Theory Comput. **7**, 4175–4188 (2011).
[37] C. Kutzner, R. Apostolov, B. Hess, and H. Grubmüller, "Scaling of the gromacs 4.6 molecular dynamics code on superMUC," in *Advances in Parallel Computing* (IOS Press, 2013), Vol. 25.
[38] C. Kutzner, S. Páll, M. Fechner, A. Esztermann, B. L. de Groot, and H. Grubmüller, "Best bang for your buck: GPU nodes for gromacs biomolecular simulations," J. Comput. Chem. **36**, 1990–2008 (2015).
[39] A. C. Pan, D. W. Borhani, R. O. Dror, and D. E. Shaw, "Molecular determinants of drug–receptor binding kinetics," Drug Discovery Today **18**, 667–673 (2013).
[40] R. O. Dror, H. F. Green, C. Valant, D. W. Borhani, J. R. Valcourt, A. C. Pan, D. H. Arlow, M. Canals, J. R. Lane, R. Rahmani *et al.*, "Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs," Nature **503**, 295–299 (2013).
[41] J. E. Stone, D. J. Hardy, I. S. Ufimtsev, and K. Schulten, "GPU-accelerated molecular modeling coming of age," J. Mol. Graphics Modell. **29**, 116–125 (2010).
[42] R. B. Best, "Atomistic molecular simulations of protein folding," Curr. Opin. Struct. Biol. **22**, 52–61 (2012).
[43] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9 (1-39)," J. Am. Chem. Soc. **132**, 1526–1528 (2010).

[44] J. D. Chodera and F. Noé, "Markov state models of biomolecular conformational dynamics," Curr. Opin. Struct. Biol. **25**, 135–144 (2014).
[45] N. Plattner and F. Noé, "Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models," Nat. Commun. **6**, 7653 (2015).
[46] K. H. Bleicher, H.-J. Böhm, K. Müller, and A. I. Alanine, "Hit and lead generation: Beyond high-throughput screening," Nat. Rev. Drug Discovery **2**, 369–378 (2003).
[47] G. Schneider and U. Fechner, "Computer-based de novo design of drug-like molecules," Nat. Rev. Drug Discovery **4**, 649–663 (2005).
[48] G. M. Keserű and G. M. Makara, "Hit discovery and hit-to-lead approaches," Drug Discovery Today **11**, 741–748 (2006).
[49] E. Kerns and L. Di, *Drug-Like Properties: Concepts, Structure Design and Methods: From ADME to Toxicity Optimization* (Academic Press, 2010).
[50] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande, "Alchemical free energy methods for drug discovery: Progress and challenges," Curr. Opin. Struct. Biol. **21**, 150–160 (2011).
[51] W. L. Jorgensen, "The many roles of computation in drug discovery," Science **303**, 1813–1818 (2004).
[52] J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design* (John Wiley & Sons, Inc., 1999).
[53] G. Schneider, "Virtual screening: An endless staircase?," Nat. Rev. Drug Discovery **9**, 273–276 (2010).
[54] C. A. Lipinski, "Drug-like properties and the causes of poor solubility and poor permeability," J. Pharmacol. Toxicol. Methods **44**, 235–249 (2000).
[55] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, "Massively multitask networks for drug discovery," e-print arXiv:1502.02072 (2015).
[56] I. Buch, T. Giorgino, and G. De Fabritiis, "Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations," Proc. Natl. Acad. Sci. U. S. A. **108**, 10184–10189 (2011).
[57] D. W. Borhani and D. E. Shaw, "The future of molecular dynamics simulations in drug discovery," J. Comput.-Aided Mol. Des. **26**, 15–26 (2012).
[58] J. Mortier, C. Rakers, M. Bermudez, M. S. Murgueitio, S. Riniker, and G. Wolber, "The impact of molecular dynamics on drug design: Applications for the characterization of ligand–macromolecule complexes," Drug Discovery Today **20**, 686 (2015).
[59] W. Sinko, S. Lindert, and J. A. McCammon, "Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design," Chem. Biol. Drug Des. **81**, 41–49 (2013).
[60] J. Wereszczynski and J. A. McCammon, "Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition," Q. Rev. Biophys. **45**, 1–25 (2012).
[61] Y. Okamoto, "Generalized-ensemble algorithms: Enhanced sampling techniques for Monte Carlo and molecular dynamics simulations," J. Mol. Graphics Modell. **22**, 425–439 (2004).
[62] D. M. Zuckerman, "Equilibrium sampling in biomolecular simulation," Ann. Rev. Biophys. **40**, 41 (2011).
[63] A. M. Ferrenberg and R. H. Swendsen, "Optimized Monte Carlo data analysis," Phys. Rev. Lett. **63**, 1195 (1989).
[64] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "Multidimensional free-energy calculations using the weighted histogram analysis method," J. Comput. Chem. **16**, 1339–1350 (1995).
[65] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys. **129**, 124105 (2008).
[66] B. Roux, "The calculation of the potential of mean force using computer simulations," Comput. Phys. Commun. **91**, 275–282 (1995).
[67] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, "Long-time protein folding dynamics from short-time molecular dynamics simulations," Multiscale Model. Simul. **5**, 1214–1226 (2006).
[68] Y. Deng and B. Roux, "Computations of standard binding free energies with molecular dynamics simulations," J. Phys. Chem. B **113**, 2234–2246 (2009).
[69] P. Nielaba, M. Mareschal, and G. Ciccotti, *Bridging the Time Scales: Molecular Simulations for the Next Decade* (Springer Science and Business Media, 2002), Vol. 605.
[70] G. R. Bowman, V. S. Pande, and F. Noé, *An Introduction to Markov State Models and their Application to Long Timescale Molecular Simulation* (Springer Science and Business Media, 2013), Vol. 797.
[71] J. L. MacCallum, A. Perez, and K. A. Dill, "Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference," Proc. Natl. Acad. Sci. U. S. A. **112**, 6985 (2015).
[72] T. Bereau and K. Kremer, "Automated parametrization of the coarse-grained Martini force field for small organic molecules," J. Chem. Theory Comput. **11**, 2783–2791 (2015).
[73] T. Bereau, D. Andrienko, and O. A. von Lilienfeld, "Transferable atomic multipole machine learning models for small organic molecules," J. Chem. Theory Comput. **11**, 3225–3233 (2015).
[74] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[75] J. Behler, "Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations," Phys. Chem. Chem. Phys. **13**, 17930–17955 (2011).
[76] Z. Li, J. R. Kermode, and A. De Vita, "Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces," Phys. Rev. Lett. **114**, 096405 (2015).
[77] P. Gasparotto and M. Ceriotti, "Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond," J. Chem. Phys. **141**, 174110 (2014).
[78] Z. D. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Müller, and G. Henkelman, "Optimizing transition states via kernel-based machine learning," J. Chem. Phys. **136**, 174101 (2012).
[79] R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt, and A. Aspuru-Guzik, "Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics," Energy Environ. Sci. **4**, 4849–4861 (2011).

[80] V. Rühle, A. Lukyanov, F. May, M. Schrader, T. Vehoff, J. Kirkpatrick, B. Baumeier, and D. Andrienko, "Microscopic simulations of charge transport in disordered organic semiconductors," J. Chem. Theory Comput. **7**, 3335–3345 (2011).
[81] P. Kordt, J. J. M. van der Holst, M. Al Helwi, W. Kowalsky, F. May, A. Badinski, C. Lennartz, and D. Andrienko, "Modeling of organic light emitting diodes: From molecular to device properties," Adv. Funct. Mater. **25**, 1955–1971 (2015).
[82] C. Poelking, K. Daoulas, A. Troisi, and D. Andrienko, "Morphology and charge transport in P3HT: A theorist's perspective," in *P3HT Revisited—From Molecular Scale to Solar Cell Devices*, Advances in Polymer Science Vol. 265, edited by S. Ludwigs (Springer, Berlin, Heidelberg, 2014), pp. 139–180.
[83] C. Poelking and D. Andrienko, "Design rules for organic donor-acceptor heterojunctions: Pathway for charge splitting and detrapping," J. Am. Chem. Soc. **137**, 6320–6326 (2015).
[84] M. C. Scharber, D. Wuhlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, and C. L. Brabec, "Design rules for donors in bulk-heterojunction solar cells–Towards 10% energy-conversion efficiency," Adv. Mater. **18**, 789 (2006).
[85] C. Poelking, M. Tietze, C. Elschner, S. Olthof, D. Hertel, B. Baumeier, F. Würthner, K. Meerholz, K. Leo, and D. Andrienko, "Impact of mesoscale order on open-circuit voltage in organic solar cells," Nat. Mater. **14**, 434–439 (2014).
[86] P. Deglmann, A. Schaefer, and C. Lennartz, "Application of quantum calculations in the chemical industry—An overview," Int. J. Quantum Chem. **115**, 107–136 (2015).
[87] Y. Unger, T. Strassner, and C. Lennartz, "Prediction of the emission wavelengths of metal-organic triplet emitters by quantum chemical calculations," J. Organomet. Chem. **748**, 63–67 (2013).
[88] F. May, M. Al-Helwi, B. Baumeier, W. Kowalsky, E. Fuchs, C. Lennartz, and D. Andrienko, "Design rules for charge-transport efficient host materials for phosphorescent organic light-emitting diodes," J. Am. Chem. Soc. **134**, 13818–13822 (2012).