Computational compound screening of biomolecules and soft materials by molecular simulations

Bereau, T.

DOI
10.1088/1361-651X/abd042

Publication date
2021

Document Version
Final published version

Published in
Modelling and Simulation in Materials Science and Engineering

IOP Publishing Modelling and Simulation in Materials Science and Engineering
Modelling Simul. Mater. Sci. Eng. **29** (2021) 023001 (38pp) https://doi.org/10.1088/1361-651X/abd042

Topical Review

Computational compound screening of biomolecules and soft materials by molecular simulations

Tristan Bereau*

Van 't Hoff Institute for Molecular Sciences and Informatics Institute, University of Amsterdam, Amsterdam 1098 XH, The Netherlands
Max Planck Institute for Polymer Research, 55128 Mainz, Germany

E-mail: t.bereau@uva.nl

Received 7 October 2020, revised 9 November 2020
Accepted for publication 3 December 2020
Published 2 February 2021

Abstract

Decades of hardware, methodological, and algorithmic development have propelled molecular dynamics (MD) simulations to the forefront of materials-modeling techniques, bridging the gap between electronic-structure theory and continuum methods. The physics-based approach makes MD appropriate to study emergent phenomena, but simultaneously incurs significant computational investment. This topical review explores the use of MD outside the scope of individual systems, but rather considering many compounds. Such an in silico screening approach makes MD amenable to establishing coveted structure-property relationships. We specifically focus on biomolecules and soft materials, characterized by the significant role of entropic contributions and heterogeneous systems and scales. An account of the state of the art for the implementation of an MD-based screening paradigm is described, including automated force-field parametrization, system preparation, and efficient sampling across both conformation and composition. Emphasis is placed on machine-learning methods to enable MD-based screening. The resulting framework enables the generation of compound-property databases and the use of advanced statistical modeling to gather insight. The review further summarizes a number of relevant applications.

Keywords: molecular simulations, computational screening, soft matter, biomolecules, high throughput, molecular dynamics

(Some figures may appear in colour only in the online journal)

*Author to whom any correspondence should be addressed.

1. Introduction

Ceder and Persson's *Scientific American* article *The Stuff of Dreams* refers to the 'golden age of materials design', a new era where computational methods—a mix of hardware and software implementation of physical laws and equations—assist scientists in designing new functional materials [1]. Designing better materials means selecting a chemical composition that yields superior materials properties. Traditional avenues have followed an Edisonian, trial-and-error approach, by experimentally screening as many compounds as possible—an approach that is typically both time-consuming and costly, due in no small part to synthesis, processing, and characterization. Computation offers a parallel route to search for compounds with desired characteristics, where the numerical solution of fundamental equations (e.g., the Schrödinger equation) can make predictions before going to the laboratory. The effort has gained momentum thanks to the development of computational hardware, software, and database tools, demonstrating exceptional potential to accelerate materials discovery in various fields [2–7].

There are good reasons to expand compound screening beyond the experimental realm. While high-throughput screening can probe impressive numbers of candidates, the requirements to synthesize, process, and/or characterize large libraries of compounds typically restricts the approach to particular systems and properties [8–13]. The computational route certainly also holds its share of system and property limitations, but are alleviated by the variety in resolutions, methods, and algorithms. Limitations may also arise from the set of compounds accessible: synthesized drugs form a minuscule subset of the chemical space of small organic molecules [14]. While not all compounds are expected to be necessary to satisfyingly interpolate the space, the level of subsampling unfortunately leads to a lack of uniformity: a database bias [15, 16]. Screening on the computer, on the other hand, needs no synthesis—though its virtual analog, model parametrization, often remains a challenge. More flexibility in choosing compounds enables avenues to exhaustively enumerate small subsets [17], find efficient ways to build up combinatorics [18], and select compounds using more sophisticated strategies, for instance active learning [19].

To remain robust across chemical space, computational methods must rely on fundamental, broadly applicable physical laws and equations. These physics-based methods—including the Schrödinger and Kohn–Sham equations at the electronic-structure level and Newton's classical equations of motion at the classical level—can make predictions that are grounded in the corresponding physics. Even classical simulations typically give rise to significant computational costs, which had until recently limited their penetration into the field of compound screening. Turning to density functional theory (DFT), the recent development yet rapid adoption of high-throughput schemes for various materials applications testifies to the escalating role of computation in materials screening and discovery [3, 20–22].

While some fields have already benefitted strongly from computational screening, others lag behind—such is the case for soft condensed matter. Marked by weak characteristic interaction energies on par with thermal energy, $k_{\mathrm{B}}T$, soft-matter systems embody a large class of materials, including not only polymers, liquid crystals, surfactants, colloids, but also biomolecules. When coupled to thermal fluctuations, soft matter display fascinating phenomena, such as spontaneous self assembly and mesoscopic architectures, simply navigating a rich free-energy landscape [23]. Fluctuations de facto require a careful consideration of entropic effects, and adequate computational methods to sample the accessible conformational space. Furthermore, soft-matter systems also typically display poor scale separation, challenging multiscale-modeling approaches [24].

The challenges of modeling biomolecules and soft matter have largely kept the field in a 'craftsmanship era'. Scientific studies focus on one or a handful of compounds, due to difficulties in parametrizing, preparing, sampling, and analyzing the system. These aspects all stand orthogonal to a screening strategy—automation reigns over the high-throughput paradigm. It is thus no surprise that machine learning and other data-driven techniques are rapidly penetrating the field of soft materials [25–27]. The rapid rise of high-throughput molecular simulations is the topic of this review.

1.1. Scope

Compound screening is a vast, quickly evolving area that connects to physics and chemistry, materials science, and even branches out to a plethora of applications, from organic photovoltaics to electrocatalysis to drug discovery to biomaterials [11, 28–31]. Despite its focus on biomolecular systems and soft matter, this compound-screening review will exclude studies originating from experimental data—arguably its largest subset. A large body of work has been devoted to the utilization of experimental compound databases, notably from quantitative structure–activity relationship (QSAR) methods in drug discovery [32–34]. Instead this review will focus not only on computational (in silico) screening, but those generated from *physics-based* methods. Physics-based methods consist of a hierarchy of multiscale-modeling methods, from quantum chemistry, to empirical force-field-based MD, to particle-based coarse-grained (CG) simulations, to continuum modeling [24, 35, 36]. They prevail in some key aspects essential to biomolecular materials and soft matter, specifically the modeling of emergent phenomena and entropy. Further, this hierarchy offers a conceptual bridge to the funnel-like nature of compound screening: quickly screen with fast methods and refine with more accurate models.

Current computational limitations strongly limit a purely quantum-chemical approach to a limited range of problems: primarily isolated molecules or relatively small and homogeneous environments [37]. Classical MD simulations prevail for biomolecules and soft matter, because of their ability to efficiently sample the vast conformational space. For a history and overview of MD simulations, we refer the reader to excellent books and reviews [38–42]. Though MD-based screening studies are dominated by an atomistic resolution, CG models take an increasingly large role, thanks to their more favorable computational load and ongoing improvements in linking the lower resolution to the underlying chemistry. This review will mostly revolve around *spatial* CG: particle-based models made of interaction sites (also called superparticles or beads), which correspond to groups of atoms. On the other hand, we will not touch upon methods that coarse-grain in *time*, due to (so far) limited impact on compound screening [43–47].

1.2. Inverse problems in soft matter

A material, entirely determined by its chemical composition—but also often its processing—will yield specific properties. Making measurements, either by experimental techniques or analytical/numerical calculations, boils down to establishing a mapping between the material composition and its properties. This is commonly denoted the *forward* problem, and is illustrated in figure 1(a) [48]. Materials design, on the other hand, aims at establishing the backward—or inverse—mapping: identifying the adequate structure given properties of interest. While the forward route is straightforward, there is no experiment or equations of motion to directly probe the backward problem. It instead typically requires solving an inverse problem: from a (small) number of forward measurements, infer the function that links chemistry to materials property. The notorious difficulty to solve inverse problems also applies in materials

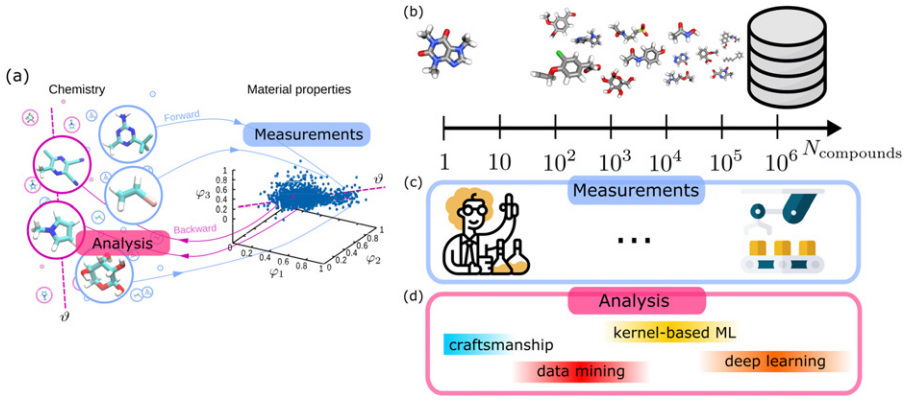

Figure 1. (a) Structure-property relationships are based on forward measurements and subsequent backward inference; (b) analogous to length- and time-scales in materials modeling, the number of compounds—the data-scale—is an essential variable of compound-screening problems; (c) measurements can only be performed manually for the lowest values of $N_{\text{compounds}}$, but otherwise require automation. (d) Different scales of $N_{\text{compounds}}$ are amenable to different types of statistical modeling. Part of the figure is adapted from [48], under a creative commons attribution (CC BY) license.

discovery, and leads to strenuous requirements on the number of measurements compared to the size of the interpolation space [49].

Though commonly referred to as *structure-property relationships*, this terminology hides that the structure itself is entirely determined by the material's chemical constituents. The review by Sherman *et al* clearly differentiates four different stages in the design of soft materials: (i) chemical synthesis or preparation leads to (ii) building blocks with effective, CG interactions, which drive their assembly into (iii) structures or morphologies, and imprint (iv) properties on the macroscopic scale [50]. This chemistry-building-block–structure-property framework does justice to the complexity, heterogeneity, and large scale separation that characterizes soft matter.

The chemistry to building-block step, (i $\rightarrow$ ii), is essential to reduce the overwhelming vastness of chemical space [14, 51] into a low-dimensional set of effective components with CG interactions. This requires a thorough understanding of the dominant driving forces: supramolecular interactions such as van der Waals, electrostatics, or hydrogen bonds [52]. Modeling has greatly taken advantage of building blocks by means of top-down coarse-graining, which parametrize simple models based on key phenomenological interactions, while staying close to the chemistry [53, 54]. The building-block to structure step, (ii $\rightarrow$ iii), has likely received the most attention. Relevant work largely consists of improving our understanding or finding practical routes at linking CG interactions to self assembly. Notable examples include the directed self assembly of diblock copolymer thin films using self-consistent field theory [55]; the 'materials design engine', using statistical mechanics as an automatic optimizer, with applications including the folding of a polymer and the directed self assembly of block copolymers [56]; design principles for colloidal self assembly with short-range interactions, establishing tight restrictions on the relative strength of the favorable and unfavorable interactions, as well as the number of components and energies [57]; a 'digital alchemy' framework to control self assembly by optimizing building blocks for a given target bulk structure [58]. The structure to property step, (iii $\rightarrow$ iv), has largely involved finite-element methods to

optimize material microstructures for specific design specifications, such as acoustic, elastic, and photovoltaic properties [59].

At equilibrium an additional consideration may prove useful in approaching inverse problems: the free-energy landscape. Central to any soft-matter system, the free-energy landscape shapes the self-assembly route, navigating down between conformational basins toward a (local) minimum. The free-energy landscape also conditions all observables, by its statistical weights over the conformational space. In the context of solving the inverse problem, the free-energy landscape thus stands as a powerful, physically meaningful intermediary between chemistry and building-block constituents on the one hand and structure/morphology and macroscopic properties on the other.

How does changing the chemistry affect the free-energy landscape? Various studies are tackling this question. Meng *et al* reported the free-energy landscape of clusters of attractive hard spheres, including a detailed characterization of the rotational entropy [60]. Scaling up, the field of protein folding has led to great insight into how the shape of the free-energy landscape impacts a protein's properties—the famous funnel-like shape is characteristic of many efficient folders [61–63]. These developments further enabled the design of new proteins, whose sequence and structure differ significantly from naturally occurring proteins [64]. Unfortunately not all free-energy landscapes display straightforward shapes; self assembly often results from a competition between conformational basins. Jankowski and Glotzer carefully studied the assembly pathway of patchy particles to grasp the diversity of possible final structures [65].

Coarse-graining likely has a strong role to play in the context of screening. As described below in section 3.8, a high-throughput study of drug–membrane thermodynamics linked CG features of small molecules with their potential of mean force of insertion in a lipid membrane [66]. The results suggest that exploring the diversity of top-down CG building blocks (step ii) fittingly *simplified* the structure–property relationship, making it easier to identify. CG models evidently coarsen the underlying free-energy landscape, and what could be criticized as a loss in accuracy or resolution can also be seen as a decisive advantage to tackle the inverse problem.

The system-size limitations associated with MD simulations naturally hinder the prospects of scaling up to genuine macroscopic properties. The systems remain instead micro- to mesoscopic and focus on basic structural, thermodynamic, and sometimes dynamical aspects. Their particle-based nature also naturally lend themselves to starting from the (i) chemistry or (ii) building-block steps.

1.3. Data-scales

One landmark property of most—if not all—materials is the large dynamic range of relevant length- and time-scales. Microscopic interactions lead to mesoscale architectures and morphologies, but also conformational transitions and aging behavior. It is not uncommon to observe phenomena spanning 10 or more orders of magnitude for either scale: from subnanometer to meter, and from femtosecond to seconds or more. Interestingly, these scales are relevant not only to understand the intrinsic properties of the system, but also to *probe* it: both experimental techniques and computational methods typically specialize in probing a (possibly small) subset of these scales [24, 35, 67]. For instance, quantum-chemical methods reign at small length- and time-scales, but fall short much beyond the nanometer- and picosecond-marks.

In this review we apply a similar conceptual framework to the *number of screened compounds*, $N_{\text{compounds}}$. This data-scale, unlike its other two counterparts, is not an intrinsic

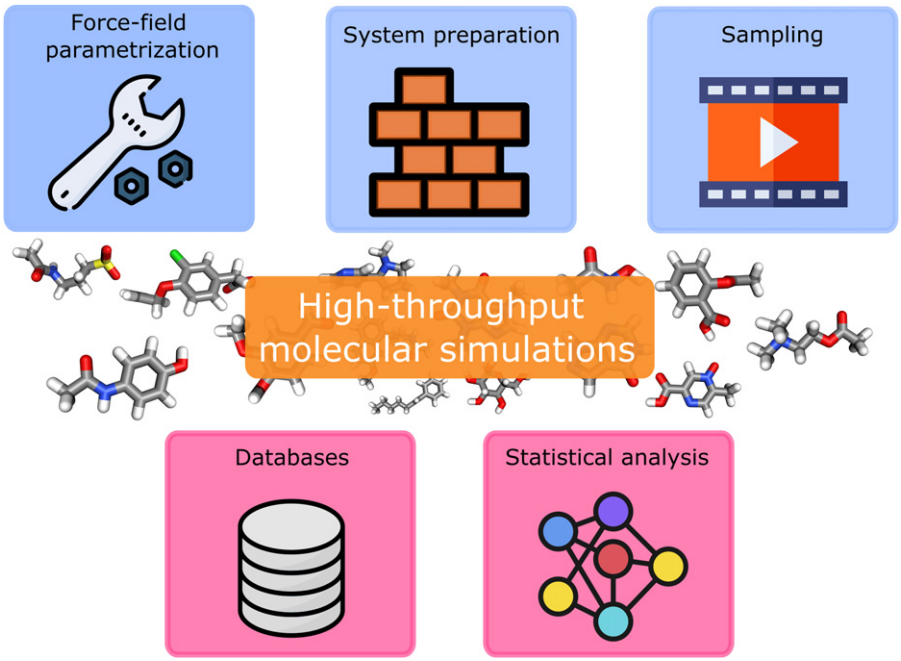

Figure 2. Protocol for high-throughput molecular simulations. Requirements include automated force-field parametrization schemes, system preparation, and efficient sampling (top; blue). It enables the generation of compound databases and statistical analysis to gather insight (bottom; pink).

variable—it is merely a practical consideration to help guide both the forward-measurement and backward-inference processes. We refer the reader to figure 1 for an illustration: establishing structure-property relationships (panel a) hinges upon the number of compounds screened (panel b). As will be described in section 3, MD studies typically work in the range $1 \leqslant N_{\text{compounds}} \lesssim 10^6$, though steady progress will likely rapidly push the upper bound. Working in higher regimes of the data scale will on the one hand strongly impact *requirements* on the forward-measurement protocol (figure 1(c)), but on the other hand *permit* more sophisticated statistical-analysis techniques (figure 1(d)). The data scale thereby forms an essential pillar to guide a compound-screening study, both to generate a database and garner insight from it.

2. Computational high-throughput paradigm

Before moving onto applications (section 3), we first describe the forward-measurement requirements and backward-analysis possibilities that a computational high-throughput paradigm both impose and enable, sketched in figure 2. The forward-measurement steps necessary to build the compound database—the blue boxes in figure 2—embody the computational analog of a laboratory's high-throughput screening experiment. The framework demands a strict and homogeneous protocol across compounds for two reasons: (i) it yields a *consistent* database amenable to extracting structure-property relationships; and (ii) it is practically convenient for automation purposes. The present section describes the various aspects of running MD simulations under these constraints.

When possible, the examples will be borrowed from the biomolecular and soft-matter fields. In other cases however, examples from other fields—in particular chemistry and hard condensed matter—may prove insightful of where developments may be headed.

2.1. Force-field parametrization

The scope and level of refinement of a number of biomolecular force fields attest to the remarkable developments in the molecular-simulation field: some of them are decades in the making, amounting to thousands of finely tuned parameters, and have endured relentless evaluations [68–73]. Unlike more empirical methods (e.g., statistical scoring in drug discovery), the physics-based nature of force fields grounds the model in the physics considered. It relies on specific potentials that encode relevant interactions [74–76]. Unfortunately force fields are difficult beasts to tame: their complexity can easily make any (re)parametrization for new compounds laborious, because they do not always offer systematic strategies.

Automated force-field parametrization is an old idea that is difficult to practically implement. Why is that? Quantum mechanics ought to provide us with a sure-fire way to derive classical potentials. Unfortunately the physics encoded in force fields is rather limited: for instance, most force fields are not explicitly polarizable. The limited physics of the model clouds the relationship to quantum mechanics and instead warrants a parametrization based on experimental properties. Major biomolecular force fields, such as CHARMM and OPLS, typically use a combination of reference information to parametrize across the chemical compound space (CCS; more on that in section 2.3.2) of drug-like small molecules: like others the CHARMM general force field (CGenFF) uses quantum mechanics to optimize charges and bonded interactions, while Lennard–Jones parameters rely on experimentally determined liquid density and heat of vaporization [77]. The need for experimental quantities can be problematic, and is alleviated by identifying chemical groups or fragments found in previously analyzed molecules. The gradual incorporation of model compounds allows CGenFF to broadly interpolate across a large subset of CCS, while retaining high fidelity of structural and thermodynamic properties. A similar strategy has been applied by OPLS [78, 79], GROMOS [80], and AMBER [81].

Arguably the incorporation of experimental data in a computational-screening pipeline is unfortunate: experimental data are limited to a minuscule subset of CCS, and it might well defeat the purpose of a virtual compound-discovery study. Despite their broad coverage of CCS, the above-mentioned biomolecular force fields largely avoid this issue by sharing and reusing information between molecules. The piece of information that is typically shared is the *atom type*. Beyond the chemical element itself, it represents the atom in a molecule given a local environment, for instance an sp^2 carbon in an alkene. The more chemically specific, the better—in other words the larger incorporation of neighboring atoms will more precisely characterize the local environment, and offer all the more resolution. The above-mentioned automated force-field strategies primarily aim at selecting the right atom types, and extract the corresponding parameters from a database. While these atom types have historically been handcrafted by chemical intuition, ongoing efforts aim at generalizing its concept using more robust annotators. For instance, the Open Force Field Initiative is applying so-called direct chemical perception by the use of SMIRKS patterns—linear notations encoding atoms and bonds [82].

The tendency to encode increasingly many atom types begs the question: is there a continuum limit? In effect this is precisely what is probed by machine learning (ML) models that span (subsets of) CCS. While we defer a broader discussion on the topic to section 2.5, we note that kernel-based methods, such as Gaussian process regression (GPR), assume and enforce

smoothness of the input space by the kernel function [83]. It leads to a continuum description of a so-called *atom-in-molecule* representation, a concept strongly utilized in hard condensed matter [75]. ML models learn a smooth interpolation between many-body atom-in-molecule representations and a target property of interest. ML has rapidly demonstrated impressive capabilities to interpolate increasingly large subsets of the CCS to complex electronic properties. Examples include atomization energies [84], dipole polarizability tensor [85], and multipole electrostatic coefficients [86].

How do we incorporate ML models into force fields? One straightforward approach is to work simultaneously with both: physics-based force fields encode the functional forms and asymptotes that we know, while ML models predict composition- and conformation-specific environments. This approach can lead to excellent accuracy and transferability, reproducing highly accurate coupled-cluster calculations across several molecular datasets, and without the need for any reparametrization [87]. Li *et al* have used ML models to predict quantum-mechanical properties, used as input for a polarizable force field, and match liquid-state observables [88]. In both cases the high-resolution of the physics-based models—they are both explicitly polarizable—enable a purely *ab initio* parametrization.

The more ML-centric alternative is to let go of functional forms entirely. Several applications show that this can lead to excellent many-body ML potentials for a variety of molecules and materials [89–91]. Moving beyond single systems and toward subsets of CCS is still a subject of ongoing research: most of these approaches have so far focused on a careful interpolation of the conformational space, and the compounded interpolation of composition requires significant adaptations (section 2.3.2). We point out the ML neural network potential ANI as a notable example in this direction [92]. We also note the challenge of accurately modeling long-range interactions, for instance by appropriate physically inspired kernels [93].

Going down in resolution, developing CG models takes the simulator down either one of two main tracks: top-down or bottom-up [53]. The top-down approach, which builds from phenomenological considerations, may turn out easier to automate in the case that there is a straightforward link between the reference information and the interaction potential. A variety of powerful models have been developed in the past, and we turn the interested reader to relevant reviews [53, 54]. Consider the popular CG Martini force field for biomolecular systems [94]. The automated CG Martini parametrization scheme can read in any small organic molecule, optimize a mapping using a set of heuristics, and predict a chemical fragment water/octanol partitioning coefficient from a neural network for each bead type [95]. Bead types of CG models can be further redefined to best accommodate for the diversity of compounds in the CCS [96]. On the other hand, the bottom-up route starts from microscopic information of a higher-resolution simulation. Systematic parametrization schemes exist, such as iterative Boltzmann inversion or force matching, accompanied by convenient software platforms [97, 98]. Aside from the CG potentials, bottom-up strategies can strongly benefit from a more systematic optimization of the mapping itself [99, 100]. Combinations of structure-based CG and ML have recently sparked interest and are quickly enabling new avenues, see below section 2.5.3.

2.2. System preparation

System preparation for an MD study has two main tenets: (i) the initial configurations and (ii) the procedure to run the simulation and compute observables (e.g., structural parameter or free energy). Controlling the latter is typically relatively easy, as it often boils down to applying the same simulation pipeline. Building initial configurations in an automated and consistent way, on the other hand, can require more sophisticated approaches: a screening study

that focuses on protein–ligand binding must first dock every single compound in the protein pocket. Beyond the proper geometric alignment of the ligand, the condensed phase of a liquid calls for packing of the molecules involved, and thus a delicate placement to avoid steric clashes. This has led to a variety of tools to initialize condensed-phase, soft-matter systems: Martínez *et al* designed PACKMOL to create simple liquids, mixtures, and more complex architectures, such as micelles and lipid bilayers [101]; Polymer Modeler is a polymer chain builder [102]; CHARMM-GUI is a sophisticated web server to facilitate the initial configuration of biomolecular systems, such as solvated proteins, and phospholipid membranes [103]; the INSANE script sets up complex phospholipid-membrane mixtures for the CG Martini force field [104]; MemProtMD elegantly prepares CG configurations of membrane proteins by *self-assembling* the phospholipid membrane around the experimentally resolved protein structure (section 3.6) [105]; both the Python-based MoSDeF and Hoobas frameworks offer extensible molecular-building capabilities (e.g., patchy DNA-grafted colloids in Hoobas), and the use of Python allows for deeper integration of system initialization and simulation/analysis [106, 107].

2.3. Sampling

Sampling lies at the heart of molecular simulations: both MD (with appropriate thermostat) and Monte Carlo simulations implement efficient importance-sampling algorithms to navigate a representative subset of the conformational space [39]. But sampling takes on a whole new dimension in the context of this review: not only does a simulation aim at sampling conformational space, compound screening is *also* a sampling problem—this one in compositional space. Here we limit our overview to recent methods that aim at sampling either space. The use of similar techniques to tackle both spaces is no coincidence, it highlights their resemblance and the associated sampling challenges.

2.3.1. Conformational sampling. The conformational space represents the structural distribution function of the system. A collection of N particles will give rise to a continuous $3N$-dimensional space of microstates. The statistical ensemble used to probe the system biases the weighting of the states, e.g., the Boltzmann distribution in the canonical ensemble. This bias means that not all microstates contribute equally, and instead an efficient conformational-sampling strategy should focus only on the more important ones.

More conformational sampling is almost always desired: simulating larger and more complex systems potentially opens up new insight unattainable before, but also helps testing for convergence issues [108, 109]. Limited computational resources limit how long the simulations can be, and instead offset many efforts in sampling more efficiently. Several excellent reviews cover the vast and rich area of enhanced-sampling techniques [110–113].

ML, and in particular deep learning, has opened up a number of new avenues in terms of facilitating conformational sampling [25]. For instance, autoencoders display an architecture that is prone to enhanced sampling: its symmetric bow-tie network, while simply aiming at reconstructing the input sample, forces an information bottleneck in the so-called *latent space*. Describing a system through this reduced dimensional latent space bridges naturally to the use of collective variables in enhanced sampling. A famous variant to autoencoders, the variational autoencoder, uses a variational approach to learn the latent representation, resulting in both a generative model and a smooth latent space that enables interpolation [114]. Various studies have leveraged the architecture of a (variational) autoencoder to learn a low-dimensional latent representation of the input conformational space [115, 116] or extract the long-time kinetics [117]. The added accuracy one can gain by using ML often comes at the cost of interpretability:

how do we express the latent-space dimensions—the collective variables—in terms of simple, physically meaningful coordinates? Ribeiro *et al* proposed to iteratively refine a set of proxy reaction coordinates that best emulates the latent-space distribution [118].

Other approaches do away with collective variables, and instead use unsupervised learning as a way to chart a low-dimensional free-energy surface. Chiavazzo *et al* have devised a method that iteratively proceeds between MD and nonlinear manifold learning techniques to expand the system away from regions already explored [119]. Expanding conformational space using dimensionality reduction was also proposed by Kukharenko *et al* [120]. They used the multidimensional-scaling scheme sketch-map [121] to project the points and initiate swarms of simulations from sparsely (but existing) sampled regions. The generation of molecular configurations that have not been previously sampled was subsequently proposed by means of a loss function that combined an autoencoder reconstruction loss and the sketch-map cost function [122]. The combination of the two approaches effectively appears to achieve features in line with the variational autoencoder: the data-driven learning of a smooth latent-space distribution, coupled to a generative model.

Beyond techniques aiming at enhancing the conformational space sampled, others have tried to blend in qualitative external knowledge—a prior of sorts—to drive the MD. Perez *et al* employed Bayesian inference to guide protein-folding MD from coarse physical knowledge, such as 'form a hydrophobic core' [123]. Folding times were reduced by several orders of magnitude, illustrating that the body of insight about protein folding can be leveraged to speed up protein simulations. This example illustrates well the dichotomy between what is systematic (e.g., algorithms) and what is not (e.g., our intuition), and the Bayesian scheme provides a formalism to bridge the two approaches. Strategies to blend numerical methods or algorithms with heuristic prior knowledge is bound to be useful in other areas.

2.3.2. Compositional sampling. The chemical compound space (CCS)—the space of all possible molecules or compounds—differs from the conformational space in at least two major ways: first, its discreteness. Conformational space permits continuous transformations between any pair of microstates. On the other hand, different molecules cannot be arbitrarily close, because of basic chemical rules (e.g., valency). In other words, very few spatial arrangements of atoms will lead to chemically stable compounds. Although there are computational treatments to continuously transform molecules (*vide infra*), the common setting is to dedicate different simulations for different molecules.

The second defining feature of CCS is its size: the dimensionality of the space is not a simple function of the number of particles. Natural proteins can be built by combinations of 20 amino acids, meaning that there are 20^n unique sequences of chain length n. For very short peptides of length $n = 10$—barely long enough to stabilize any secondary structure—this already leads us to a space of 10^{13} compounds. The increased variety of chemical groups in synthetic polymers will evidently yield a much larger CCS. Now consider small-drug like molecules that obey Lipinski's 'rule of five'—restricting the molecular weight, hydrophobicity, and number of hydrogen bonds—which capture the physicochemical properties of most orally active drugs [124], its space is estimated at 10^{60} chemically stable molecules [14]. There are not enough carbon atoms in the universe to synthesize all of them! What can we do, then? Just like microstates, not all molecules are made equal—most will yield uninteresting properties. Focusing on the ones with desired properties is precisely the answer to solving the inverse problem (section 1.2).

While overwhelmingly large, important steps in better grasping the size and scope of the CCS of drugs have been made. Reymond and co-workers have sidestepped the minuscule, inconsistent collection of synthesized drug-like molecules by instead constructing them

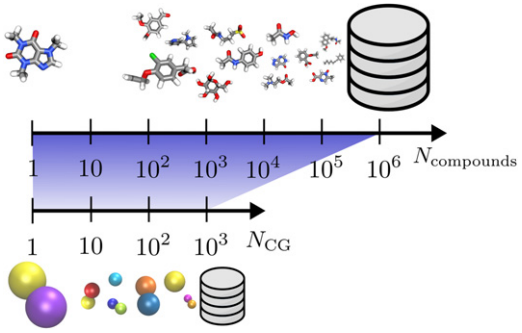

Figure 3. Transferable coarse-grained models can reduce the size of chemical compound space: fewer coarse-grained (CG) compounds are required to probe a subset of chemical space. They make use of a finite set of bead types to introduce a degeneracy in the CG representations of chemical compounds [95].

algorithmically [18, 51]. Graph-based methods combined with valency rules offer a systematic way to enumerate large subsets of CCS—most of which have never been synthesized. The so-called 'generated database' (GDB) enumerates a dense coverage of molecules containing a set of elements up to a threshold in number of heavy atoms: the GDB-17 contains 10^{11} molecules up to 17 heavy atoms of C, N, O, S, and halogens [125]. Beyond their identity, computing *properties* of these dense subsets has subsequently been subject to much activity, because they enable the training of ML models (section 2.5). The GDB has been used for the calculation of electronic properties, typically from density-functional theory (DFT), of increasingly many compounds: Rupp *et al* calculated the atomization energy of $7 \cdot 10^3$ molecules [84]; Ramakrishnan *et al* computed various electronic properties for $1 \cdot 10^5$ molecules [126]; and Hoja *et al* more recently reported a database of $4 \cdot 10^6$ molecules [127].

When tackling the exploration of CCS, coarse-graining can offer significant advantages. Top-down, phenomenological CG models focus the modeling on the essential ingredients or driving forces at play [53]. This minimalistic approach can lead to generic—if not universal—behavior that broadly applies to many systems. One famous example is the Kremer–Grest polymer model [128, 129]. Zhang *et al* demonstrated that a melt of this phenomenological model can broadly be backmapped to many different types of homopolymers [130]. Everaers *et al* recently matched the generic large-scale behavior of Kremer–Grest simulations to chemistry-specific experiments via the Kuhn length [131].

While the link between top-down CG models and the underlying CCS often remains qualitative, there can be approaches to establish it. Many of these top-down models are transferable, in that they define a limited set of interactions of *bead types* to encode the variety of chemical groups. In case of the popular Martini model the bead types roughly span the hydrophobicity scale [94]. This limited chemical resolution means that molecules alike will often map to the same CG mapping. This critically introduces a degeneracy in CG representation of small molecules, and effectively a *reduction* in the size of CCS. Figure 3 illustrates the use of Martini for small molecules: it can lead to a reduction in chemical space by roughly 3 orders of magnitude. The mapping from molecules to CG representations is straightforward to establish using automated parametrization schemes of GDB-type libraries [95, 132]. This reduction of the size of CCS can be applied to significantly boost the compound screening of thermodynamic properties—one example will be covered in the context of drug–membrane interactions, section 3.8.

Beyond mere enumeration or serendipitous picks, there are more efficient ways to explore CCS. Virshup *et al* devised an algorithm to stochastically grow an initial set of compounds to maximally diversify it, restricted to specific properties (e.g., drug-likeness) [133]. They reported a library of 10^4 compounds representative of the GDB-13, yielding a 10^4 reduction factor while retaining its diversity. Such an approach is likely to go hand in hand with the training of ML models, which require a good balance between chemical similarity and a representative coverage of the interpolation space. At the other end of the spectrum, Hoksza *et al* presented the MOLPHER framework, which provides a (discrete) path in chemical space between a pair of compounds [134]. It performs a series of simple structural molecular changes, such as atom addition or removal, from start to target molecule.

Other approaches at sampling CCS emphasize the (bio)chemistry or physics of navigating across molecules. Taking inspiration from nature has led to the adaptation of Darwinian-type directed evolution [135]. Computational directed evolution has so far mostly been applied to protein design, and more specifically to enzymes [136]. Leveraging the aptness of computational physics to perform importance sampling, a Markov chain Monte Carlo scheme can efficiently sample across CCS [137]. Closer to reproducing a laboratory experiment, Wang *et al* implemented an *ab initio* nanoreactor, leading to spontaneous chemical reactions and the formation of molecules through a variety of pathways [138]. Such a computational setting holds great promise in studying in more detail the origins of life [139].

While most of these approaches tackle CCS in its discrete form, continuous explorations may well prove extremely strategic. However, connecting compounds in a continuous manner requires some craft. One notable example is the alchemical transformation, a powerful tool in statistical mechanics to compute free-energy differences [140]. It relies on a crucial property: state functions do not depend on the path taken, and instead permit non-physical—alchemical—interpolations between two compounds (more on this in sections 3.2 and 3.3). A corresponding framework can be used to compute *ab initio* energy gradients and other changes in properties upon local changes in CCS [141, 142]. Aside from the relevant materials properties, the inclusion of *derivatives* may help in more efficiently mapping structure-property relationships [143].

Another strategy to circumvent the discreteness of CCS consists of imposing a continuous proxy. Such a proxy will enable continuous-optimization schemes, thereby facilitating molecular design. Wang *et al* employed a linear combination of atomic potentials to establish a continuous property landscape [144]. In a similar vein, von Lilienfeld *et al* relied on an energy functional based on the nuclear and electronic chemical potential [145]. With the advent of deep learning, new solutions have been proposed: Gómez–Bombarelli *et al* used a variational autoencoder (covered in section 2.3.1) to not only reduce the CCS, but more importantly to smoothen it [146]. Built in the variational autoencoder, the representation of the latent space allows a continuous exploration of the CCS. The architecture was connected to a surrogate model, whose objective was to predict a target property in the reduced latent space, enabling continuous optimization. This active-learning, Bayesian-optimization approach has lately been applied in the context of soft-matter systems by Shmilovich *et al*, as described in section 3.7 [147].

2.4. Data infrastructure

Assuming all technical requirements permit MD simulations at high throughput, the question arises: what to do with the data? Handling large collections of MD simulations can easily require extensive storage solutions. More importantly, it poses the problem of data *sharing*—not only between group members and collaborators, but across the community at large.

Recent cultural shifts in science are increasingly encouraging the dissemination of research data. A collaborative and open-source approach to scientific endeavors can strongly accelerate the pace of research [148]. Databases of experimentally determined materials properties, for instance for polymers, can prove invaluable to extract structure-property relationships and assist in designing better materials [149–151].

What to do, then, to publish large collections of MD simulations? An increasing number of online repositories dedicated to hosting scientific data have come about, Zenodo [152], figshare [153], or the open science framework [154], to name but a few. These databases are generic in that they are agnostic to the type of scientific data, unlike, say, the protein data bank (PDB), which specializes in biomacromolecular structures [155]. The next question is the data format. One straightforward solution is to simply compress all the input and output files of a set of MD trajectories and upload them as is—a strategy our group adopted to publish hundreds of umbrella-sampling MD trajectories [156]. This lets anyone freely access the data, but presents caveats. Notably, (i) it does not facilitate automated strategies to search and collect information about the data, and (ii) the input/output formats are tied to MD software used to generate the simulation trajectories. This is more formally denoted by a lack of data labeling—or *metadata*—and data normalization, respectively. The convenient access, retrieval, and categorization of heterogeneously generated data is key to assemble large databases, amenable to training ML models (more on that in section 2.5). Such a framework has been formalized by the FAIR principles: data that is findable, accessible, interoperable, and reusable [157]. The new era of computational materials design mentioned in the Introduction is in no small part made possible by a robust data infrastructure in materials science [158]. Publishing large FAIR datasets is becoming increasingly widespread, thanks to solutions like the Materials Data Facility [159]. The development of a number of data-infrastructure platforms, such as NOMAD and the Materials Project, strive to label electronic-structure calculations by detailed metadata, parse many codes and normalize the input and output information, and offer access via a webpage or a programmatic interface [160, 161]. Several consortia are working their way toward more robust data infrastructures for molecular simulations, including OpenKIM [162, 163], MOLSSI [164], and FAIR-DI [165]. Recent examples show that the interconnection of specialized databases can help automate the metadata annotation process, as will be described in section 3.6.

2.5. Data analysis

Once the difficult task of generating MD-based compound databases is over, a second one starts: the data analysis. Here we will rely on the concept of data-scale, already introduced in section 1.3. Figure 1 illustrates that the number of compounds largely determines the type of statistical modeling. This constraint stems from the expressivity of a statistical model, which depends largely on the number of parameters of the architecture and dimensionality of the representation, which themselves require larger training set sizes. We structure what follows in terms of the data-scale by means of the variable $N_{\text{compounds}}$, from the traditional setting of craftsmanship, to data mining in the low-data regime, to kernel-based ML methods, to deep learning.

2.5.1. Craftsmanship. Working in a regime $N_{\text{compounds}} \sim 1$ leaves little room for data-driven analysis methods. It instead embodies the traditional setting of gathering insight driven by physical theories, experiments, prior computer simulations, or simply intuition.

2.5.2. Data mining. Moving up to $N_{\text{compounds}} \gtrsim 10$ can offer enough information to systematically search for simple structure-property relationships. The low number of samples puts

a strong limit on the dimensionality of the sample information—the descriptors. Relating low-dimensional descriptors to materials property has enjoyed great attention for decades, embodied for instance by so-called quantitative structure–property relationships (QSPR) [34, 166]. QSPR is a well-established, powerful method to functionally relate chemical structure to property. Applications include largely drug discovery [167, 168], but we also note other soft-matter systems, such as the self assembly of conjugated oligopeptides (more on that in section 3.7) [169] and the tribology of functionalized, lubricating monolayer films [107]. QSPR relies on a set of descriptors, typically combined using a (multivariate) linear fit. More recent applications have turned to using the kernel trick to convert a non-linear problem into a linear one, support vector machines can then highlight the most important descriptors, and we further note the increasing use of artificial neural networks [167, 168]. Practically however, these data-mining models tend to be less limited by algorithmic developments than by the data itself: small values of $N_{\text{compounds}}$ can easily lead to a large dependence to the training set. This aspect calls for particular attention to *model generalization*: how similar do the predicted molecules need to be from the training set [167].

A more recent take on the functional discovery of structure-property relationships brings us to learning more complex equations. Compressed-sensing methods extend QSPR to expand the complexity of the functional relationships tested. They rely on a large combinatorial consideration of trial candidate equations, and a greedy l^1-norm optimization scheme to minimize the number of non-zero coefficients. Examples include the symbolic regression of nonlinear dynamical systems [170] and equations from the *Feynman lectures on physics* [171]. Ghiringhelli *et al* used least absolute shrinkage and selection operator (LASSO) to extract functional relationships between descriptors that can accurately classify between zinc blende and rocksalt semiconductors [172]. Ouyang *et al* refined the approach using the sure independence screening and sparsifying operator (SISSO), which hierarchically searches for combinations of descriptors [173]. Rather than building a single surrogate model aimed at explaining the entire dataset, another method called subgroup discovery focuses on coherent homogeneous subsets. Goldsmith *et al* revisited the zinc-blende/rocksalt semiconductor problem and identified separate regions with strict constraints [174]. These models are of particular interest at a time where ML models are increasingly criticized for their lack of interpretability: identifying the explicit role of the input variable in the structure–property mapping.

By and large, these approaches aim at capturing the *essential* variables or descriptors that dictate the target property. This dimensionality reduction aims at garnering insight into the problem at hand, ideally by visualizing how the minimal set of descriptors link to the property. The systematic construction of reduced dimensional representations is a vast field, one that naturally connects to unsupervised-learning techniques [175].

2.5.3. Kernel-based supervised learning. The regime $N_{\text{compounds}} \gtrsim 10^3$ is amenable to the optimization of much more expressive models. These are often called surrogate models: they aim at learning the (oftentimes complex) relationship between input and output parameters, so as to yield a computationally efficient prediction. These models strive for accuracy and generalization: to make a precise prediction over a large interpolation domain. At best, the accuracy of the estimation can be on par with the reference method [176]. We refer the reader to several excellent reviews on the use of (kernel-based) ML for molecular systems [4, 6, 25, 90, 177]. Compared to QSPR methods, ML methods are free of fixed functional forms, and instead offer flexible interpolation between training points in a high-dimensional feature space [83, 178]. ML models exploit similarity in several ways: they first impose a *metric*, allowing us to measure distances in CCS, a critical ingredient to both explore and sample from that space

(section 2.3). Similarity is explicitly assumed by enforcing smoothness between input space and target property—an aspect that helps interpolate between training points.

Naturally, ML is not free of pitfalls. The application of ML to materials modeling—and more specifically to molecular systems—requires domain knowledge. To be competitive, an ML model should outperform an ambitious baseline: our own understanding of physics and chemistry! An appealing strategy is to *construct* physics or chemistry inside ML models—an aspect we outline below.

The increased expressivity of ML relies on the use of higher-dimensional input information, *representations*, rather than mere descriptors. Representations offer a more detailed—many-body—description of the system, such as a molecule or an atom in its local environment [84, 179–181]. A higher-dimensional representation also means more difficulties in probing how broadly the ML model can be deployed: at which point does it start extrapolating? How will we know? While there are many facets to these questions, one crucial piece of information we can take advantage of is the *underlying physics*. Given that my system obeys a conservation law or symmetry, can we constrain an ML model to satisfy this constraint *a priori*? The need to account for physical symmetries was recognized early on [182]. The Noether theorem states that symmetries in a physical system lead to conservation laws and invariants. Empirically learning these invariants often requires significant amount of training data—encoding them in the representation or the ML architecture can lead to significant learning improvement [183]. As a result, translation, rotation, or (when applicable) permutation invariance often form the basic requirements for ML representations. Symmetries can be added to the kernel *itself*, notable examples include the learning of vectors by covariant kernels [184] or energy-conserving force fields via the Hessian [91, 185, 186]. Additional constraints can be added as well, for instance a decomposition ansatz when the target property lumps several terms, useful to decompose reference forces [89], atomic dipole moments [187], or free energies [188]. Kernels turn out to be extremely convenient to encode physical constraints because they work within the realm of linear algebra. Extending these properties to neural networks and deep learning is more challenging, though the improved expressivity has motivated active developments (*vide infra*).

The lessons learned to build ML models in chemistry and materials science largely transfer to soft matter and biomolecules, where similar constraints on the representation prevail [189]. Screening studies that make use of kernel-based ML have become prominent, for instance in protein–ligand binding, but many typically use experimental data [179]. Using MD, the relevant data-scale regimes typically require a CG approach. For instance in drug-membrane thermodynamics, CG simulations of $\sim 10^3$ systems led to predictions for $1.3 \cdot 10^6$ molecules, thanks to the CG model's reduction of CCS [137]. The predictions satisfied thermodynamic relations observed on smaller data sets, strongly suggesting robust generalization. While this study was based on a top-down CG model, systematic approaches like the variational force-matching method bode elegantly well with the loss function of an ML model. This has resulted in several studies, and in particular efforts at addressing the challenging question of mapping many atomistic configurations to a single CG geometry [186, 190–192].

Several challenges still lie ahead for a more robust description of condensed liquid-state systems. For instance, a (macro)molecule is never isolated, but embedded in its environment, such that a representation may benefit by incorporating the neighboring solvent's degrees of freedom [193]. The nature of the systems naturally calls for the development of ML-based force fields that incorporate long-range interactions [194], as well as more particle types. We also point out the critical role of the configurational aspect: a single geometry is not representative, but rather should incorporate information about the underlying Boltzmann distribution [188]. More than anything else, high-quality ML models require extensive training data. Soft

matter needs large, homogeneous databases analogous to what has been developed from DFT calculations for electronic properties, e.g., the QM9 database [126].

2.5.4. Deep learning. The extraordinary results achieved with deep learning in so many scientific and technological fields have to do with the added expressivity of these models. Using a neural-network architecture that connects several layers of nodes, input and output can be mapped to generalize surprisingly well [195]. Compared to the above-mentioned regimes, the added expressivity of deep learning comes at a price: they rely on an overwhelming number of parameters, and a non-convex problem to solve. Practically this entails many more training data points necessary to parametrize a model, typically in the range $N_{\text{compounds}} \gtrsim 10^6$.

The benefits of deep learning are far reaching: notably for drug discovery—though so far with data generated from experiments [196, 197], we also outlined some of the distinct conceptual advantages a deep-learning approach offers for sampling both across conformations and compositions (section 2.3.1). In terms of representing molecules, the inclusion of symmetries is also an essential aspect, requiring extensive methodological work [198, 199]. They open the door to so-called *physics-informed neural networks*, which aim at a synergistic combination of the two approaches to reduce the training data, effectively regularizing in small data-scale regimes [200]. Deep learning offers exciting opportunities: for instance graph convolutional neural networks (CNNs) offer a physically intuitive representation for molecules, where nodes and edges represent atoms and bonds. Graph CNNs offer appealing features: differentiable, more easily interpretable, and better performing than commonly used molecular fingerprints [201].

Harnessing the full potential of deep-learning models puts stringent requirement on the number of compounds, which severely restricts what can be achieved in terms of screening studies. Few MD studies have reached data-scale regimes amenable to deep learning, but impressive first steps show much promise, such as the prediction of transfer free energies in lipid membranes [202]. It offers a glance at the use of MD-based studies to train deep-learning models across the CCS of biomolecular and soft materials.

3. Screening applications

The following describes a number of MD-based screening applications for various soft-matter and biomolecular systems. We order the applications roughly in the number of compounds screened, from low to high, and grouped by topics when deemed fitting. Beyond the range of screening sizes, some of these applications result from intense and long-standing scientific activities. For those, the present review cannot do justice to the breadth of these research topics, but will hopefully stimulate the reader in diving into complementary readings.

3.1. Exploring conformational space with swarms of trajectories

Far from a screening at high throughput, this first application focuses on the study of *individual* (macro)molecules. While slightly deviating from the greater objective to screen across compounds, the conceptual approach and implementation undertaken here is relevant for our topic, as it provides innovative solutions to exploring conformational space.

The problem at heart involves the determination of kinetic properties for systems exhibiting relevant processes at long time scales—*long* compared to what would be considered reasonably achievable by a single trajectory on a supercomputer. Supercomputers tackle ambitious simulations by means of CPU or GPU parallelization. Unfortunately, not everything is easy to parallelize: While one can easily segment a simulation box to treat smaller cells concurrently,

MD numerically integrates the equations of motion in a serial fashion—it is difficult to parallelize time. Folding@Home tackled the problem by introducing two complementary aspects: a conceptual approach to circumvent the long-time-scale sampling problem, and a platform to implement it [203].

The dynamics of complex systems is typically dominated by free-energy barriers: thermal fluctuations will lead a system to dwell in a conformational basin (i.e., a local minimum), before being spontaneously pushed over a barrier. Assuming single-exponential kinetics with (unknown) rate k, the probability for the system to cross the barrier at time t is given by $P_1(t) = k\exp(-kt)$. Rather than wait for a single trajectory to cross over once, let many copies attempt it over a short time. In the case of M simulations, the probability for the first simulation to cross at the same time t is now $P_M(t) = Mk\exp(-Mkt)$, exhibiting an effective rate that is M times faster. The pioneering work of Pande and co-workers demonstrated the value of the approach: running multiple instances of a short simulation boosts the chances of seeing early crossing events, and sufficiently many occurrences allow them to estimate the rate k, as illustrated on the folding of small peptides and polymers [204].

The second breakthrough of the Folding@Home consortium was to establish a distributed-computing platform, powered by idle CPU power contributed by anonymous users over the internet [203]. Running many short, uncoupled simulations meant that they did not need to run on the same supercomputer. All simulation instances need no communication, since they independently sample the same conformational space. Practically this was simply realized by M copies of the same initial configuration (typically with different seeds and velocities), since the stochasticity of the dynamical process will quickly lead to diverging trajectories.

One of the early examples of the Folding@Home project aimed at the folding kinetics of two mutants of the designed, 23-residue-long mini-protein BBA5 [205]. With a mean folding time on the order of 10 μs, it is considered a fast-folding protein, yet very much a challenging time-scale for an all-atom MD simulation—especially at the time the research was conducted. Following the above-mentioned reasoning for single-exponential kinetics, they estimated that for such a folding timescale, roughly 10 out of 10 000 individual trajectories should fold after 10 ns. Using an implicit-solvent united-atom model, they showed that an impressively large number of short simulations yielded excellent agreement with laser temperature-jump experiments.

Folding@Home has made significant contributions in elucidating the protein-folding problem in silico [63, 206]. Early applications were then superseded with Markov state models, a more robust memoryless master-equation treatment of the kinetics, pioneered by Noé, Pande, Chodera, Bowman, and others [47, 207–210].

Moving away from protein folding, a more recent application of distributed-computing platforms focused on protein–ligand binding. Using their distributed-computing platform GPUGRID, De Fabritiis and co-workers demonstrated the value of the approach for PMF calculations for standard binding free energies [211]. Buch *et al* reported an impressive study of the enzyme-inhibitor complex trypsin–benzamidine: they performed 495 unbiased MD simulations of the unbound ligand for 100 ns each [212, 213]. They sampled a variety of binding events, but also several *pathways*, allowing them to robustly estimate both the binding free energy, as well as the on and off binding rates. Extensions to the modeling of protein–protein association kinetics form to date one of the most impressive developments in this area [214].

Distributed-computing platforms have had a conceptual impact as to how the community increasingly approaches MD simulations: from handcrafted, individual instances to swarms of trajectories. The associated need for automation paves the way for different kinds of

high-throughput MD simulations. Spawning MD trajectories has since been extended to exploring uncharted regions of the free-energy landscape using machine learning [119].

3.2. Protein–ligand binding

The ever-growing penetration of computational chemistry in drug discovery has experienced its shares of challenges [215]. Like any complex engineering problem, the design of a drug entails many considerations and complementary problems to solve. From membrane penetration, to toxicity, to pharmacokinetic and pharmacodynamic considerations, we focus here solely on the determination of protein–ligand binding.

Basic structure-based drug-design methods typically assume rigid drug-target structures: starting from a crystal structure or homology modeling, a ligand is docked near the receptor's active site; the molecular configuration is then used to estimate binding, often using empirical scoring functions as a proxy. While this type of virtual screening accommodates a large number of compounds, it models the complex as mostly rigid. The lack of flexibility is an issue, given the recognized role of the conformational ensemble in biomolecular activity [216]. The field moved from a static lock-and-key binding paradigm to more dynamic pictures, such as induced fit or conformational selection. This emphasizes the need for physics-based methods that model not only structural flexibility, but more broadly the relevant emergent phenomena following binding [217].

Beyond flexibility, an accurate account of the binding free energy is desired. Free energies are *ensemble* properties, making the scoring of any individual configuration a conceptually peculiar exercise. Several methods have been developed and tested over the years—the drug-design field having explored many methodologies to strike the right balance between accuracy and throughput: from end-point methods to rigorous calculations derived from statistical mechanics.

One prominent example of an end-point method combines MD simulations on the bound and unbound configurations, using an implicit solvent and a Poisson–Boltzmann surface area solvation term (MM-PBSA). Brown and Muchmore applied MM-PBSA to a set of 308 ligands bound to one of three protein receptors [218]. The breadth and scope of the study is laudable: moving toward a high-throughput MD scheme to extract free energies of binding. The moderate correlation coefficients (Pearson coefficient $R^2 = 0.5$–0.7) are unfortunately a testament to the difficulties end-point methods display in reliably directing drug discovery [219, 220].

Alchemical transformations provide a rigorous framework to compute binding free energies [140]. Though many methodologies exist [221, 222], we mention one equilibrium techniques that aims at calculating the free energy upon transforming from state **A** to **B**: free-energy perturbation, introduced by Zwanzig [223], relies on exponential averaging

$$\Delta G_{\mathbf{A}\to\mathbf{B}} = G_{\mathbf{B}} - G_{\mathbf{A}} = -k_{\mathrm{B}}T \ \ln\left\langle \exp\left(-\frac{\mathcal{H}_{\mathrm{B}}(\boldsymbol{r}) - \mathcal{H}_{\mathrm{A}}(\boldsymbol{r})}{k_{\mathrm{B}}T}\right)\right\rangle_{\mathbf{A}}, \tag{1}$$

where $\boldsymbol{r}$ denotes the system's particle coordinates, $\mathcal{H}_{\mathrm{A}}$ is the Hamiltonian of state **A**, and $\langle\cdot\rangle_{\mathbf{A}}$ is an ensemble average at state point **A**.

Three decades ago, the pioneering study of Wong and McCammon presented an alchemical transformation between benzamidine bound to the enzyme trypsin [224]. A fascinating review by Jorgensen describes some of the successes of MD coupled with alchemical transformations to advance the drug-discovery pipeline [219]. While the generation of new scaffolds (i.e., entirely different structures) is naturally sought, so-called hit-to-lead optimization—refinement of the binding of a promising starting compound—is where alchemical transformations really shine. There are two reasons for this: (i) the computational expense of each

alchemical transformation limits the screening to relatively few compounds, thereby limiting the chances of finding new scaffolds; and (ii) the interpolative nature of an alchemical transformation (i.e., overlap in the conformational spaces, see 1) leads to better convergence for similar molecules.

Alchemical transformations took a more systematic turn with the study of Wang *et al* [225]. They reported relative free-energy calculations at an all-atom level with explicit solvent for an impressive 200 ligands. This feat was aided by the deployment of MD simulations on graphics processing units (GPU), as well as a streamlined procedure to prepare and run alchemical transformations. Critically, they optimized a 'perturbation graph', which measures the maximum common substructure between any pair of compounds [226]. The algorithm minimizes the number of alchemical transformations, while accommodating for both multiple pathways to estimate statistical error and the presence of closed cycles (which ought to yield no free-energy difference). With a total of 330 perturbations, they reported a root-mean-squared error against experiments of only 1.1 kcal mol^{-1}. More recent work has reported alchemical transformations for up to several thousands of ligands [227]. Force-field improvements, from OPLS2.1 to OPLS3 and OPLS3e have yielded systematic improvements in binding free energies [78, 79].

Three decades of MD-based computational drug design have shown impressive developments: not only in the sheer number of compounds (from 1 to thousands reported in a single study), but more importantly in the convergence of the calculations via significantly longer simulation trajectories, and an overall improvement of the force fields. The significant contributions of industrial actors is a testament to both the pressing needs of the pharmaceutical industry and the opportunities offered by physics-based MD methods.

3.3. Solvation of small molecules

The free energy of solvation of small molecules is in many ways an antechamber to protein-ligand binding: it consists of the free-energy difference of transferring a small molecule from the gas into a condensed-phase environment. Rather than a protein pocket, solvation is performed in a bulk liquid. The homogeneity of the medium makes the calculations easier to converge, typically allowing for broader studies that may accommodate significantly more compounds.

The study of Jorgensen and Ravimohan pioneered alchemical transformations by converting methanol into ethane [228]. They applied free-energy perturbation (covered in section 3.2) to compute the relative free-energy difference in hydration—solvation in water—of the two compounds. An alchemical transformation between these two similar molecules helps the calculation: it only requires decoupling the hydroxyl group and coupling a methyl in its stead.

Modeling solvation has had significant impact as a proxy for more complex phenomena—a prime example being protein folding (some of which was covered in section 3.1). The protein-folding problem was always strongly pushed by computer simulations [63]. Huang *et al* reported an insightful study on hydrophobic solvation, they calculated the free energy of solvation for hard-sphere solutes of various sizes [229]. These solutes, though not directly linked to any particular chemistry, aimed at a better phenomenological understanding of possibly large hydrophobic regions exposed to water, such as in protein folding. Of particular interest was the systematic change in the solute size and comparison of the asymptotics against theory. In the same vein, the early 2000s witnessed intense activities in accurate calculations of hydration and transfer free energies of (neutral) amino-acid side-chain analogs [230–233].

Mobley *et al* reported hydration free energies for a set of 44 small, neutral molecules [234]. A larger set of 239 small neutral organic molecules was later tested against various force-field parameters and charge models [235–237]. In parallel, Mobley *et al* released the FreeSolv database, a set of 504 neutral small organic molecules, with comparison against experiments [238]. Such studies have led to the more routine incorporation of hydration free energies in validating force fields [78, 79]. Scaling up, Bennett *et al* recently reported an impressive $15 \cdot 10^3$ water–cyclohexane transfer free-energy calculations from all-atom MD [202].

Experimental free-energy datasets such as FreeSolv are useful because they cover much of the diversity of small drug-like molecules, although the small number of compounds necessarily limits how representative they are. ML models of in silico hydration free energies trained on different datasets—both experimental and combinatorially generated—did not appropriately generalize across each other, highlighting biases in the chemical space covered [188]. Still, the increased size and breadth of the spanned chemical space allow researchers to identify systematic problems with force-field parameters for classes of compounds. The same holds true at the CG level: the automated Martini parametrization scheme for small molecules facilitates the calculation of partitioning free energies for several hundred molecules [95]. It helped identify systematic issues with certain chemical groups, such as rings or halogens, which new versions of the force field aim at correcting [239].

With a growing number of computational techniques to compute free energies, how can one compare their predictive accuracy in a fair way? Nicholls *et al* set up an informal blind-test study, comparing different methodologies for 17 small molecules [240]. This was later formalized through the SAMPL challenge [241, 242]. The blind tests consisted of teams applying their method to compounds for which solvation free energies are known but unpublished or relatively inaccessible. It avoids the risks of tuning model parameters that would skew results to seem artificially more favorable. SAMPL2 introduced an explanatory section to gain insight in (disclosed) unexpected experimental results [243]. Later challenges have since occurred and keep helping benchmark and refine computational methods [244].

3.4. Ionic liquids

Ionic liquids (ILs) are salts. They exhibit a melting point or glass-transition temperature below 100°, while so-called 'room-temperature' ILs remain liquid below 0°. ILs typically exhibit good thermal stability, low vapor pressures, and are able to dissolve many compounds. This makes ILs interesting solvents in sustainable chemistry, with technological applications such as solvent for biomolecules or catalysis [245]. Critically, ILs are also conductive, which makes them candidates for use in electrochemical applications. In parallel, the combinatorics of association of cation–anion pairs leads to an extraordinary number of possible ILs. The combination of the breadth of chemical structures available and the variety of properties of interest has motivated a number of quantitative structure-property relationship modeling, albeit so far mostly exclusively from experimental data [34].

Computer simulations have played a significant role in better understanding ILs. Maginn pointed out that interests in ILs rose coincidentally with the advent of computer simulations, which have proven increasingly capable of shedding light on complex fluids [246]. The complex structural, thermodynamic, and dynamical aspects, including behavior at interfaces, viscosity, and dynamical heterogeneity motivated computational studies at various scales, from quantum-mechanical calculations to classical atomistic to CG modeling [246–249].

Turning to computational screening, Osti *et al* reported an insightful study aimed at probing ion interactions and transport in solvated ILs [250]. They fixed the IL cation–anion

pair (1-butyl-3-methyl-imidazolium bis(trifluoromethylsulfonyl)), but screened across four organic solvents: acetonitrile (CH_3CN), methanol (CH_3OH), tetrahydrofuran (C_4H_8O), and dichloromethane (CH_2Cl_2). The potential of mean force of separating a cation–anion pair suggested clear correlations between the energetics of the interaction and solvent polarity: a larger dipole moment is better able to screen ion–ion interactions, thereby decreasing the free energy of solvation. This clear trend was mirrored in the dynamics: ion diffusivity showed a linear increase against the solvent dipole moment. The results were corroborated by quasi-elastic neutron scattering experiments, overall offering clear structure-property relationships.

A larger, follow-up screening yielded surprising results [251]. Thompson *et al* extended the set of systems they studied, both in terms of IL–solvent mixtures (18 increments in the range 0.1–0.95 mass fraction) and solvent chemistry (22 solvents including nitriles, alcohols, halocarbons, carbonyls, and glymes) for a total of 396 state points. This study both further confirmed a previously observed trend—IL mass fraction against IL diffusivity—and uncovered a new one—solvent diffusivity against IL diffusivity. Critically, they revisited the previously observed trend by Osti *et al* between IL diffusivity and solvent dipole moment [250]: the incorporation of more compounds indicated no strong correlation across the entire data set. The authors hinted at the role of complementary solvent order parameters to recover clear trends. Combined, the two studies by Osti *et al* and later Thompson *et al* illustrate a decisive aspect: the inference of structure-property relationships hinges on a representative set of chemical compounds.

3.5. Silicate glasses

Glasses—materials that have been cooled significantly but without crystallizing—are known as structurally similar to but dynamically very different from liquids [252]. Glassy materials play a key role in many technological areas, motivating the optimization of their mechanical properties, from hardness to fracture strength to elastic properties [253]. Glasses embody an overwhelming class of materials, when considering not only the compositional aspects—potentially including a large number of elements of the periodic table—but also its strong out-of-equilibrium nature, meaning that the processing of the material can easily lead to kinetic traps.

Yang *et al* recently presented a high-throughput MD study of silicate glasses, in an effort to predict their Young's modulus [254]. They covered the ternary diagram of calcium aluminosilicate (CAS), $CaO–Al_2O_3–SiO_2$, by use of 231 compositions over the domain in 5% mol regular increments. The authors ran MD simulations with tailored force fields [255] using a melt-quench procedure to prepare the configurations. All efforts were made at providing a consistent system-preparation and simulation protocol throughout the compositional space studied, but some limiting regimes required specific treatments: (i) Higher initial melting temperature for samples with high SiO_2 concentrations, due to their higher glass-transition temperatures; and (ii) faster cooling rate for samples with high CaO concentrations, as they otherwise tend to crystallize. These aspects illustrate the challenges faced by the need for consistent protocols across large regions of chemical/compositional space.

From the simulation data, they predicted the Young's modulus across the compositional space using different statistical models. All their approaches—from polynomial regression to various flavors of machine learning—led to excellent results, indicative of both a dense sampling of the compositional domain and a smooth mapping to the target property. Interestingly, they showed that fitting models to available experimental data ($\sim$ 100 points) led to severe biases: (i) clustering of the available data leaves large domains without any training points; and (ii) significant uncertainty and systematic errors between experiments can lead to large

variations. While the latter aspect can be alleviated by means of adequate regularization, the former recalls the ever-present dangers of extrapolation.

3.6. Membrane proteins

Building up on the modeling of soluble proteins (see section 3.1), membrane proteins form an important subset due to their biochemical impact: they form roughly 25% of all human proteins [256] and half of current drug targets [257]. Membrane proteins typically exert significantly more complexity than their soluble counterparts. Transmembrane proteins in particular—those that span the membrane bilayer—evolve in a highly complex environment at the interface between the membrane and the aqueous environment. This complex environment is compounded by the large sizes that membrane proteins typically exhibit, often made of numerous α helices or a prominent β barrel. As a result, the size and heterogeneity of membrane proteins have made them challenging, not only for structure determination [258, 259], but also for computer simulations [260–263].

The computational modeling of membrane proteins has benefitted heavily from particle-based CG models. An all-atom treatment of a protein and its surrounding lipid membrane remains to date a heroic effort: protein folding happens over much longer time scales in the membrane, due to the much larger correlation times exerted in the bilayer. Peptide folding and insertion in a lipid membrane has been reported at an atomistic level, although using an implicit-membrane description, thereby speeding up the peptide dynamics in the membrane environment [264]. Alternatively, CG models offer an appealing way to study peptide folding and insertion in explicit membranes, thereby offering the means to monitor how the peptide perturbs membrane structure [265, 266].

A CG description of membrane proteins does not only allow to study folding and insertion for one of them, it can also be used to study a larger number of systems. Sansom *et al* presented more than a decade ago an impressive protocol to automate the preparation of transmembrane proteins [267]. Starting from experimentally determined protein structures—typically deposited in the protein data bank (PDB) [155, 268]—these macromolecules typically lack structural information about the aqueous and membrane environments. Running MD simulations of a membrane protein requires first to solvate it in both a lipid membrane and an aqueous environment. Atomistic protocols typically start from *equilibrated* lipid bilayers and place a hole to incorporate the macromolecule [269]. Instead, the CG protocol of Sansom *et al* did not order the lipids in any way, but rather incorporated them as an unstructured 'soup'. The soup spontaneously rearranged into a bilayer, thanks to self assembly and the speedy molecular diffusion at the CG level. Other CG based schemes have been developed to ease and automate the generation of complex lipid bilayers [270] and the assembly of membrane-protein multimers [271]. We note that the Martini-like CG model does not allow for secondary or tertiary structure reorganization, and is instead restrained around the crystal structure [272].

The pioneering database of Sansom *et al* contained 91 membrane proteins and was made available together with a web server to easily visualize structural information [267]. Though no longer available today, the Sansom group later released an expanded database of membrane proteins: MemProtMD [273]. Based on a more sophisticated pipeline, the CG-based preparation protocol was amended by a backmapping to atomistic resolution [274]. They also more systematically imported structures from the PDB. The shear size and incomplete data annotation of the PDB led them to design structural descriptors to detect α-helical and β-barrel membrane proteins. An ensemble analysis across structures allowed them to gain insight in the probabilities of occurrence of amino acid side chains with respect to the depth in the bilayer.

The MemProtMD database and associated web server contains more than 3500 PDB entries [275]. A systematic connection with other databases brings in additional metadata to group structures according to their constituent proteins and family. The network of protein databases helps automatically annotate these structures with valuable information.

Beyond the screening of membrane proteins themselves, cell membranes embed these biomolecules in complex plasma membranes, made of a wide diversity of compounds. Corradi *et al* studied the protein–lipid interactions for 10 membrane proteins embedded in a model plasma membrane made of 60 lipid species [276]. The authors identified clear 'lipid fingerprints': preferential association of certain lipid species to parts of the protein. This study highlights the combinatorial challenge involved, not only through the shear sampling of each system, but the extreme compositional diversity at hand.

3.7. Oligopeptide self assembly

The use of oligopeptides, consisting of a small number of residues, to self assemble nanostructures offers the promise of tunable supramolecular functionalities, yet with ease of preparation, biocompatibility, and degradability [277, 278]. They are proving viable contenders for applications in biomedicine and nanotechnology [279, 280]. Various types of nanostructures can be achieved, including fibers, tubes, and sheets [281, 282]. This diversity stems from the vast combination of 20 natural amino acids into sequences.

In a series of studies, Frederix *et al* have set up a systematic MD-based virtual screening protocol to establish clear structure-property relationships between the amino-acid sequence of short peptides and self assembly under aqueous conditions. Using the CG Martini force field, they first probed the ability to reproduce structural features of the well-characterized diphenylalanine (FF) peptide [17]. The aggregation of 1600 dipeptides for 1.5 μs of simulation time (approximately accounting for the acceleration due to coarse-graining) generated a tubular nanostructure whose dimensions are in agreement with x-ray diffraction analysis of crystallized FF nanotubes [283]. This indicated that despite structural limitations of the Martini force field to model protein secondary structure, it could yield reasonable self-assembling features. Beyond the final structure, the simulations also helped understand the *mechanism* of formation: from an initial random placement to quick ordering into sheet-like aggregates, to vesicle formation, and finally long hollow tubes.

Scaling up, Frederix *et al* screened *exhaustively* the space of all possible $20^2 = 400$ dipeptide combinations [17]. Although coarse-graining significantly speeds up the simulations, the scope of the study led the researchers to rapidly probe early determinants of aggregation. They followed the self assembly of 300 dipeptides for 400 ns. They scored the peptides' aggregation propensity by means of the solvent-accessible surface area, relative to the initial well-mixed configuration. The score was in good qualitative agreement with experimentally resolved structures, for the few sequences available. Though in need of atomistic refinement, the study highlights how CG simulations can sketch the mapping between sequence and self-assembled nanostructure.

A follow-up study aimed at the broader exploration of all tripeptides: $20^3 = 8000$ in total [284]. They sought compounds that simultaneously favored aggregation propensity *and* hydrophilicity. While *a priori* contradicting requirements, their results testify to the broad diversity of possible systems, including subtle intermediates capable of displaying surprising properties. Extending the dipeptide study, their aggregation-propensity score was combined with the water–octanol partitioning coefficient to measure hydrophobicity. They identified a significant number of peptides that were not strongly hydrophobic, yet exhibit aggregation. The screening confirmed and extended design rules for the placement of specific amino acids in a

particular position [285, 286]. This includes steric effects in the placement of aromatic residues close to the N-terminus, but also charged amino acids on positions 1 and 3 as an architecture for intermolecular salt-bridge formation. Critically, their virtual screening procedure led for the first time to the subsequent synthesis and experimental characterization of tripeptides able to form hydrogels at neutral pH.

More complex oligopeptides were considered more recently by Thurston and Ferguson: a synthetic peptide–Π–peptide symmetric triblock architecture of the form NXXX–Π–XXXN, where X are amino acids and Π is a conjugated aromatic core [169]. To limit the space of candidates, they initially restricted their study to one of two aromatic cores, naphthalenediimide and perylenediimide, and the five amino acids A, F, G, I, and V were motivated by prior work. Aiming at optoelectronic functionality, their design objective targeted the stabilization of π–π stacking between neighboring oligopeptides, measuring the distance between aromatic cores as a proxy for electronic delocalization. They relied on an atomistic resolution with an implicit-solvent model to more efficiently sample the conformational space. Both free energies of dimerization and trimerization were calculated using enhanced-sampling MD on 26 peptides. Intermediate values of the dimerization and trimerization free energies led to the most favorable properties, as a tradeoff between sufficient interaction strength to drive assembly, yet little enough to avoid kinetic traps. A quantitative structure-property relationship (QSPR) model was then trained on these select peptides and a large set of 247 molecular descriptors, based on the PaDEL software package [287]. The authors motivated their choice over more sophisticated machine learning approaches both for its interpretability, as well as the dataset's *high-dimensional*, *low-sample size* regime. Further MD validation confirmed the predictability of the QSPR for largely apolar sequences—similar to the 26 training peptides—and proposed a new sequence unstudied by experiment. While the QSPR lacked transferability to strongly polar residues, the results indicate that adding a limited set of MD simulations should be straightforward and effective.

A wider study, also aiming at optimizing optoelectronic properties, was recently reported by Shmilovich *et al* [147]. The synthetic architecture DXXX–OPV3–XXXD used a three-repeat oligophenylenevinylene π core, for its ability to assemble into optically and electronically active nanoaggregates [288]. Compared to the study of Thurston and Ferguson, the wider space of $20^3 = 8000$ peptides was tackled by two complementary strategies: (i) CG simulations using the Martini force field, and (ii) a deep representational active learning approach. Following the pioneering work of Gómez–Bombarelli *et al* [146], they projected the discrete sequence space into a low-dimensional *continuous* representation. A variational autoencoder was used to train a latent-space embedding [114], based on basic topological features of the CG beads of the Martini model. They trained a Gaussian process regression (GPR) on the latent-space embedding to predict the propensity of self assembly, and used a Bayesian optimization to select the 'next best' candidates to be simulated. Iterating over several generations of this loop, they were able to converge the GPR model by only simulating 2.3% of the space of sequences. This computational design platform, which aptly combines molecular simulations for compound measurement and data-driven methods to efficiently sample the sequence space, holds many promises for the virtual screening of biomolecular and soft materials.

3.8. Drug-membrane permeabilities

One beloved application of biomolecular simulations is the cell membrane. Though composed of a large variety of molecules, many are phospholipids. These amphiphiles can spontaneously self assemble to form large mesoscale structures, such as vesicles. This compartmentalization of the cell can still allow for exchange of (macro)molecules—either via active transport

(biology), or passively by simple diffusion (thermodynamics). This latter aspect can be considered by the concentration gradient of a solute molecule, such as a drug, across a soft interface between two aqueous environments. Expressing this as a one-dimensional Smoluchowski equation along the normal to the membrane, z, leads to the inhomogeneous solubility-diffusion model [289, 290]. The resulting quantity is the permeability coefficient, P, a flux that accounts for the heterogeneity of the environment by integration over z the energetics of crossing together with the local diffusivity

$$P^{-1} = \int \mathrm{d}z \frac{\exp\left[\beta G(z)\right]}{D(z)}. \tag{2}$$

In this equation, $\beta = 1/k_{\mathrm{B}}T$ is the inverse temperature, $D(z)$ is the local diffusivity, and $G(z)$ is the potential of mean force (PMF)—it is the free energy required to cross the interface as a function of the order parameter z. Interestingly this quantity is not readily accessible from current experimental techniques, leaving computer simulations as the gold standard.

The use of enhanced-sampling techniques, such as umbrella sampling, offer the means to compute the PMF at an atomistic resolution and gather unprecedented insight [167, 291]. Unsurprisingly the calculation of $G(z)$ is tremendously difficult to converge: approximately 10^5 CPU-hours is required for a small rigid molecule crossing a single-component lipid membrane using explicit-solvent atomistic models. This unfortunately limits an atomistic throughput to $\sim$ 10 molecules per study [292–295].

Here again, CG models allow for a significant step up in the number of compounds that can be screened. Beyond the reduced representation speeding up convergence of each simulation, the mapping to a Martini representation easily leads to large numbers of compounds (section 2.3). Menichetti *et al* reported the PMFs of $4.6 \cdot 10^5$ small molecules in a one-component 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC) membrane [296]. This collection of compounds resulted from the exhaustive screening of all CG small molecules made of one and two neutral Martini beads, 14 and 105, respectively. The resulting set of PMFs showed a strict variety, which could be accurately correlated to the water/octanol partitioning of the solute—a bulk quantity to relate to structural features at the membrane interface. The mapping between chemistry and CG representations was established by coarse-graining subsets of the GDB [18], keeping compounds that mapped to one- and two-bead representations.

A follow-up study extended the screening from PMFs to the permeability coefficient (2) [66]. The CG simulations did not inform the diffusivity term (problematic due to inconsistent accelerations of the CG dynamics [297]), but were instead taken from atomistic simulations, indicating weak dependence on the solute's chemistry [292]. The results showed excellent agreement with atomistic simulations and correlation with experiments, despite the minimalistic modeling approach. Permeability coefficients were predicted for $5.1 \cdot 10^5$ small organic molecules. Projecting the permeability surface onto two physically motivated descriptors (hydrophobicity and acidity, i.e., pK_{a}) highlighted the localization of key chemical groups, and their influence on the target property. It also challenged earlier phenomenological models of solute permeation [298].

A further scale up in the number of compounds 'simply' comes down to a broader screening toward larger CG representations: from one- and two-bead constructs to more. The combinatorics of the Martini bead types, while more favorable than atomically-detailed chemistry, still grow exponentially: 14, 105, 1470, and 19 306 for one- to four-bead constructs—only considering linear chains. Instead of an exhaustive account, Hoffmann *et al* presented an importance-sampling scheme to navigate the space of compounds [137]. A metropolis-chain Monte Carlo

scheme was devised by daisy-chaining compounds via alchemical transformations, and using the relative free energy in the metropolis criterion. This led to a large network of compounds sampled, and the use of closed thermodynamic cycles allowed for small corrections to the free energies. The space of compounds that was *not* sampled was subsequently predicted using a simple kernel-based ML model. Some of the predictions were explicitly validated, but all followed simple linear relationships between transfer free energies that had been identified for the smaller compounds [296]—the thermodynamics of the system acted as an ML physical constraint *global to the compound dataset*. Overall it boosted the prediction of transfer free energies to $1.3 \cdot 10^6$ small organic molecules.

Extending the high-throughput CG framework, compound screening can be used to better understand differential stabilization between lipid domains, as a proxy for small molecules modulating complex multi-component lipid membranes [299]. The difference in PMF minima between the relevant environments stands as a computationally appealing proxy for large-scale simulations of membrane reorganization. The results could identify families of compounds that could induce membrane mixing or demixing. Compound screening and their effect on membrane thermodynamics may help us better understand the mechanism of action of certain anesthetics [300].

4. Outlook

The path toward in silico compound screening of biomaterials and soft materials seems clear, but still contains a number of important hurdles before reaching large data-scale regimes. Automating the preparation, parametrization, and analysis of MD simulations is necessary to reach a high throughput, and has largely embodied the scope of this review. The other critical aspect is our capacity to run enough MD simulations, clearly the main bottleneck. In this sense, CG modeling has an important role to play: its ability to emulate a complex systems with fewer degrees of freedom offers a significant scale-up in the context of screening. The added capability to reduce the size of chemical space seems to be a promising way to ease the analysis and extraction of structure-property relationships.

Beyond statics, in silico compound screening will likely hold essential to target *dynamical* properties, such as mean-first passage times, folding and nucleation rates, or even aging dynamics. To achieve this, force-field methods need to improve the modeling of dynamics—a statement that holds at all scales, though in particular at the CG level. The perspective to move toward non-equilibrium systems will require the means to incorporate *processing* effects in materials, leading to structure-process-property relationships. Getting there will be challenging: non-equilibrium systems have no well-defined free-energy surface, and they critically depend on how the system is prepared [23].

Last, compound screening needs tighter integration with experiments. This is not only in light of verifying the in silico predictions, but a collaborative procedure between simulations and experiment that is poised to further accelerate soft-materials discovery.

Acknowledgments

I thank Andrew L Ferguson and Joseph F Rudzinski for critical reading of the manuscript. I am grateful to several colleagues and collaborators for insightful discussions on topics pertaining to this review, including Denis Andrienko, Andrew L Ferguson, Kiran H Kanekal, Kurt Kremer, Anatole von Lilienfeld, Roberto Menichetti, and Joseph F Rudzinski. Icons on figures 1–3 made by Becris, Eucalyp, Freepik, and Monkik from www.flaticon.com.

ORCID iDs

Tristan Bereau https://orcid.org/0000-0001-9945-1271

References

[1] Ceder G and Persson K 2013 The stuff of dreams *Sci. Am.* **309** 36–40
[2] Jain A *et al* 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002
[3] Curtarolo S, Hart G L W, Nardelli M B, Mingo N, Sanvito S and Levy O 2013 The high-throughput highway to computational materials design *Nat. Mater.* **12** 191–201
[4] Pyzer-Knapp E O, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J and Aspuru-Guzik A 2015 What is high-throughput virtual screening? a perspective from organic materials discovery *Annu. Rev. Mater. Res.* **45** 195–216
[5] Jain A, Shin Y and Persson K A 2016 Computational predictions of energy materials using density functional theory *Nat. Rev. Mater.* **1** 15004
[6] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 Machine learning in materials informatics: recent applications and prospects *npj Comput. Mater.* **3** 54
[7] Tkatchenko A 2020 Machine learning for chemical discovery *Nat. Commun.* **11** 4125
[8] Mishra K P, Ganju L, Sairam M, Banerjee P K and Sawhney R C 2008 A review of high throughput technology for the screening of natural products *Biomed. Pharmacother.* **62** 94–8
[9] Mayr L M and Bojanic D 2009 Novel trends in high-throughput screening *Curr. Opin. Pharmacol.* **9** 580–8
[10] Macarron R *et al* 2011 Impact of high-throughput screening in biomedical research *Nat. Rev. Drug Discov.* **10** 188–95
[11] Potyrailo R, Rajan K, Stoewe K, Takeuchi I, Chisholm B and Lam H 2011 Combinatorial and high-throughput screening of materials libraries: review of state of the art *ACS Comb. Sci.* **13** 579–633
[12] Muster T H, Trinchi A, Markley T A, Lau D, Martin P, Bradbury A, Bendavid A and Dligatch S 2011 A review of high throughput and combinatorial electrochemistry *Electrochim. Acta* **56** 9679–99
[13] Du G, Fang Q and den Toonder J M J 2016 Microfluidics for cell-based high throughput screening platforms-a review *Anal. Chim. Acta* **903** 36–50
[14] Dobson C M 2004 Chemical space and biology *Nature* **432** 824–8
[15] Hert J, Irwin J J, Laggner C, Keiser M J and Shoichet B K 2009 Quantifying biogenic bias in screening libraries *Nat. Chem. Biol.* **5** 479–83
[16] Lin A, Horvath D, Afonina V, Marcou G, Reymond J-L and Varnek A 2018 Mapping of the available chemical space versus the chemical universe of lead-like compounds *ChemMedChem* **13** 540–54
[17] Frederix P W J M, Ulijn R V, Hunt N T and Tuttle T 2011 Virtual screening for dipeptide aggregation: toward predictive tools for peptide self-assembly *J. Phys. Chem. Lett.* **2** 2380–4
[18] Fink T and Reymond J-L 2007 Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery *J. Chem. Inf. Model.* **47** 342–53
[19] Warmuth M K, Liao J, Rätsch G, Mathieson M, Putta S and Lemmen C 2003 Active learning with support vector machines in the drug discovery process *J. Chem. Inf. Comput. Sci.* **43** 667–73
[20] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD) *JOM* **65** 1501–9
[21] Jain A, Shin Y and Persson K A 2016 Computational predictions of energy materials using density functional theory *Nat. Rev. Mater.* **1** 15004
[22] Himanen L, Geurts A, Foster A S and Rinke P 2019 Data-driven materials science: status, challenges, and perspectives *Adv. Sci.* **6** 1900808
[23] Doi M 2013 *Soft Matter Physics* (Oxford: Oxford University Press)

[24] Peter C and Kremer K 2010 Multiscale simulation of soft matter systems *Faraday Discuss.* **144** 9–24
[25] Ferguson A L 2017 Machine learning and data science in soft materials engineering *J. Phys.: Condens. Matter.* **30** 043002
[26] Bereau T 2018 Data-driven methods in multiscale modeling of soft matter *Handbook of Materials Modeling: Methods: Theory and Modeling* ed W Andreoni and S Yip (Berlin: Springer) pp 1–12
[27] Jackson N E, Webb M A and de Pablo J J 2019 Recent advances in machine learning towards multiscale soft materials design *Curr. Opin. Chem. Eng.* **23** 106–14
[28] Greeley J, Jaramillo T F, Bonde J, Chorkendorff I and Nørskov J K 2006 Computational high-throughput screening of electrocatalytic materials for hydrogen evolution *Nat. Mater.* **5** 909–13
[29] Simon C G and Lin-Gibson S 2010 Combinatorial and high-throughput screening of biomaterials *Adv. Mater.* **23** 369–87
[30] Kitchen D B, Decornez H, Furr J R and Bajorath J 2004 Docking and scoring in virtual screening for drug discovery: methods and applications *Nat. Rev. Drug Discov.* **3** 935–49
[31] Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera R S, Gold-Parker A, Vogt L, Brockway A M and Aspuru-Guzik A 2011 The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid *J. Phys. Chem. Lett.* **2** 2241–51
[32] Hellberg S, Sjoestroem M, Skagerberg B and Wold S 1987 Peptide quantitative structure-activity relationships, a multivariate approach *J. Med. Chem.* **30** 1126–35
[33] Topliss J 2012 *Quantitative Structure-Activity Relationships of Drugs* vol 19 (Amsterdam: Elsevier)
[34] Le T, Epa V C, Burden F R and Winkler D A 2012 Quantitative structure-property relationship modeling of diverse materials properties *Chem. Rev.* **112** 2889–919
[35] Tadmor E B and Miller R E 2011 *Modeling Materials: Continuum, Atomistic and Multiscale Techniques* (Cambridge: Cambridge University Press)
[36] van der Giessen E *et al* 2020 Roadmap on multiscale materials modeling *Modelling Simul. Mater. Sci. Eng.* **28** 043001
[37] Szabo A and Ostlund N S 2012 *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (Mineola, NY: Dover)
[38] Binder K 1995 *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* (Oxford: Oxford University Press)
[39] Frenkel D and Smit B 2001 *Understanding Molecular Simulation: From Algorithms to Applications* vol 1 (Amsterdam: Elsevier)
[40] Karplus M and McCammon J A 2002 Molecular dynamics simulations of biomolecules *Nat. Struct. Biol.* **9** 646–52
[41] Hansson T, Oostenbrink C and van Gunsteren W 2002 Molecular dynamics simulations *Curr. Opin. Struct. Biol.* **12** 190–6
[42] Rapaport D C 2004 *The Art of Molecular Dynamics Simulation* (Cambridge: Cambridge University Press)
[43] Mori H 1965 Transport, collective motion, and brownian motion *Prog. Theor. Phys.* **33** 423–55
[44] Zwanzig R 1973 Nonlinear generalized Langevin equations *J. Stat. Phys.* **9** 215–20
[45] Tuckerman M, Berne B J and Martyna G J 1992 Reversible multiple time scale molecular dynamics *J. Chem. Phys.* **97** 1990–2001
[46] Gear C W, Hyman J M, Kevrekidid P G, Kevrekidis I G, Runborg O and Theodoropoulos C 2003 Equation-free, coarse-grained multiscale computation: enabling mocroscopic simulators to perform system-level analysis *Commun. Math. Sci.* **1** 715–62
[47] Chodera J D and Noé F 2014 Markov state models of biomolecular conformational dynamics *Curr. Opin. Struct. Biol.* **25** 135–44
[48] Bereau T, Andrienko D and Kremer K 2016 Research update: computational materials discovery in soft matter *APL Mater.* **4** 053101
[49] Kaipio J and Somersalo E 2006 *Statistical and Computational Inverse Problems* vol 160 (Berlin: Springer Science and Business Media)
[50] Sherman Z M, Howard M P, Lindquist B A, Jadrich R B and Truskett T M 2020 Inverse methods for design of soft materials *J. Chem. Phys.* **152** 140902
[51] Reymond J-L 2015 The chemical space project *Acc. Chem. Res.* **48** 722–30
[52] Stone A 2013 *The Theory of Intermolecular Forces* (Oxford: Oxford University Press)

[53] Noid W G 2013 Perspective: coarse-grained models for biomolecular systems *J. Chem. Phys.* **139** 09B201
[54] Ingólfsson H I, Lopez C A, Uusitalo J J, de Jong D H, Gopal S M, Periole X and Marrink S J 2013 The power of coarse graining in biomolecular simulations *WIREs Comput. Mol. Sci.* **4** 225–48
[55] Hannon A F, Gotrik K W, Ross C A and Alexander-Katz A 2013 Inverse design of topographical templates for directed self-assembly of block copolymers *ACS Macro Lett.* **2** 251–5
[56] Miskin M Z, Khaira G, de Pablo J J and Jaeger H M 2015 Turning statistical physics models into materials design engines *Proc. Natl. Acad. Sci.* **113** 34–9
[57] Hormoz S and Brenner M P 2011 Design principles for self-assembly with short-range interactions *Proc. Natl. Acad. Sci.* **108** 5193–8
[58] van Anders G, Klotsa D, Karas A S, Dodd P M and Glotzer S C 2015 Digital alchemy for materials design: colloids and beyond *ACS Nano* **9** 9542–53
[59] Jain A, Bollinger J A and Truskett T M 2014 Inverse methods for material design *AIChE J.* **60** 2732–40
[60] Meng G, Arkus N, Brenner M P and Manoharan V N 2010 The free-energy landscape of clusters of attractive hard spheres *Science* **327** 560–3
[61] Bryngelson J D, Onuchic J N, Socci N D and Wolynes P G 1995 Funnels, pathways, and the energy landscape of protein folding: a synthesis *Proteins* **21** 167–95
[62] Shakhnovich E 2006 Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet *Chem. Rev.* **106** 1559–88
[63] Dill K A and MacCallum J L 2012 The protein-folding problem, 50 years on *Science* **338** 1042–6
[64] Kuhlman B and Baker D 2004 Exploring folding free energy landscapes using computational protein design *Curr. Opin. Struct. Biol.* **14** 89–95
[65] Jankowski E and Glotzer S C 2012 Screening and designing patchy particles for optimized self-assembly propensity through assembly pathway engineering *Soft Matter* **8** 2852
[66] Menichetti R, Kanekal K H and Bereau T 2019 Drug-membrane permeability across chemical space *ACS Cent. Sci.* **5** 290–8
[67] Serdyuk I N, Zaccai N R, Zaccai J and Zaccai G 2017 *Methods in Molecular Biophysics* (Cambridge: Cambridge University Press)
[68] Maple J R, Dinur U and Hagler A T 1988 Derivation of force fields for molecular mechanics and dynamics from *ab initio* energy surfaces *Proc. Natl. Acad. Sci.* **85** 5350–4
[69] Halgren T A and Damm W 2001 Polarizable force fields *Curr. Opin. Struct. Biol.* **11** 236–42
[70] Wang W, Donini O, Reyes C M and Kollman P A 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** 211–43
[71] Ponder J W and Case D A 2003 Force fields for protein simulations *Adv. Protein Chem.* **66** 27–85
[72] Mackerell A D 2004 Empirical force fields for biological macromolecules: overview and issues *J. Comput. Chem.* **25** 1584–604
[73] Wang L-P, McKiernan K A, Gomes J, Beauchamp K A, Head-Gordon T, Rice J E, Swope W C, Martínez T J and Pande V S 2017 Building a more predictive protein force field: a systematic and reproducible route to amber-fb15 *J. Phys. Chem.* B **121** 4023–39
[74] Halgren T A 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** 7827–43
[75] Tkatchenko A, DiStasio R A Jr, Car R and Scheffler M 2012 Accurate and efficient method for many-body van der waals interactions *Phys. Rev. Lett.* **108** 236402
[76] Van Vleet M J, Misquitta A J, Stone A J and Schmidt J R 2016 Beyond born-mayer: improved models for short-range repulsion in *ab initio* force fields *J. Chem. Theory Comput.* **12** 3851–70
[77] Vanommeslaeghe K and MacKerell A D Jr 2015 CHARMM additive and polarizable force fields for biophysics and computer-aided drug design *Biochim. Biophys. Acta* **1850** 861–71
[78] Harder E *et al* 2015 OPLS3: a force field providing broad coverage of drug-like small molecules and proteins *J. Chem. Theory Comput.* **12** 281–96
[79] Roos K *et al* 2019 OPLS3e: extending force field coverage for drug-like small molecules *J. Chem. Theory Comput.* **15** 1863–74
[80] Malde A K, Zuo L, Breeze M, Stroet M, Poger D, Nair P C, Oostenbrink C and Mark A E 2011 An automated force field topology builder (ATB) and repository: version 1.0 *J. Chem. Theory Comput.* **7** 4026–37

[81] Wang J, Wang W, Kollman P A and Case D A 2006 Automatic atom type and bond type perception in molecular mechanical calculations *J. Mol. Graph. Model.* **25** 247–60
[82] Mobley D L *et al* 2018 Escaping atom types in force fields using direct chemical perception *J. Chem. Theory Comput.* **14** 6076–92
[83] Rasmussen C E 2004 Gaussian processes in machine learning *Advanced Lectures on Machine Learning* (Berlin: Springer) pp 63–71
[84] Rupp M, Tkatchenko A, Müller K-R and Von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[85] Wilkins D M, Grisafi A, Yang Y, Lao K U, DiStasio R A and Ceriotti M 2019 Accurate molecular polarizabilities with coupled cluster theory and machine learning *Proc. Natl. Acad. Sci.* **116** 3401–6
[86] Bereau T, Andrienko D and von Lilienfeld O A 2015 Transferable atomic multipole machine learning models for small organic molecules *J. Chem. Theory Comput.* **11** 3225–33
[87] Bereau T, DiStasio R A Jr, Tkatchenko A and von Lilienfeld O A 2018 Non-covalent interactions across organic and biological subsets of chemical space: physics-based potentials parametrized from machine learning *J. Chem. Phys.* **148** 241706
[88] Li Y, Li H, Pickard F C IV, Narayanan B, Sen F G, Chan M K Y, Sankaranarayanan S K R S, Brooks B R and Roux B 2017 Machine learning force field parameters from *ab initio* data *J. Chem. Theory Comput.* **13** 4492–503
[89] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104** 136403
[90] Behler J 2016 Perspective: machine learning potentials for atomistic simulations *J. Chem. Phys.* **145** 170901
[91] Chmiela S, Tkatchenko A, Sauceda H E, Poltavsky I, Schütt K T and Müller K-R 2017 Machine learning of accurate energy-conserving molecular force fields *Sci. Adv.* **3** e1603015
[92] Smith J S, Isayev O and Roitberg A E 2017 ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost *Chem. Sci.* **8** 3192–203
[93] Grisafi A and Ceriotti M 2019 Incorporating long-range physics in atomic-scale machine learning *J. Chem. Phys.* **151** 204105
[94] Marrink S J and Tieleman D P 2013 Perspective on the martini model *Chem. Soc. Rev.* **42** 6801–22
[95] Bereau T and Kremer K 2015 Automated parametrization of the coarse-grained martini force field for small organic molecules *J. Chem. Theory Comput.* **11** 2783–91
[96] Kanekal K H and Bereau T 2019 Resolution limit of data-driven coarse-grained models spanning chemical space *J. Chem. Phys.* **151** 164106
[97] Rühle V, Junghans C, Lukyanov A, Kremer K and Andrienko D 2009 Versatile object-oriented toolkit for coarse-graining applications *J. Chem. Theory Comput.* **5** 3211–23
[98] Dunn N J H, Lebold K M, DeLyser M R, Rudzinski J F and Noid W G 2017 BOCS: bottom-up open-source coarse-graining software *J. Phys. Chem* B **122** 3363–77
[99] Chakraborty M, Xu J and White A D 2020 Is preservation of symmetry necessary for coarse-graining? *Phys. Chem. Chem. Phys.* **22** 14998–5005
[100] Foley T T, Kidder K M, Shell M S and Noid W G 2020 Exploring the landscape of model representations *Proc. Natl. Acad. Sci.* **117** 24061–8
[101] Martínez L, Andrade R, Birgin E G and Martínez J M 2009 PACKMOL: a package for building initial configurations for molecular dynamics simulations *J. Comput. Chem.* **30** 2157–64
[102] Haley B, Wilson N, Li C, Arguelles A, Jaramillo E and Strachan A 2018 *Polymer Modeler*
[103] Jo S, Kim T, Iyer V G and Im W 2008 CHARMM-GUI: a web-based graphical user interface for CHARMM *J. Comput. Chem.* **29** 1859–65
[104] Wassenaar T A, Ingólfsson H I, Böckmann R A, Tieleman D P and Marrink S J 2015 Computational lipidomics with insane: a versatile tool for generating custom membranes for molecular simulations *J. Chem. Theory Comput.* **11** 2144–55
[105] Newport T D, Sansom M S P and Stansfeld P J 2019 The memprotmd database: a resource for membrane-embedded protein structures and their lipid interactions *Nucleic Acids Res.* **47** D390–7
[106] Girard M, Ehlen A, Shakya A, Bereau T and de la Cruz M O 2019 Hoobas: a highly object-oriented builder for molecular dynamics *Comput. Mater. Sci.* **167** 25–33
[107] Summers A Z, Gilmer J B, Iacovella C R, Cummings P T and MCabe C 2020 MoSDeF, a python framework enabling large-scale computational screening of soft matter: application to

chemistry-property relationships in lubricating monolayer films *J. Chem. Theory Comput.* **16** 1779–93
[108] Neale C, Bennett W F D, Tieleman D P and Pomès R 2011 Statistical convergence of equilibrium properties in simulations of molecular solutes embedded in lipid bilayers *J. Chem. Theory Comput.* **7** 4175–88
[109] Shaw D E *et al* 2010 Atomic-level characterization of the structural dynamics of proteins *Science* **330** 341–6
[110] Abrams C and Bussi G 2013 Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration *Entropy* **16** 163–99
[111] Bernardi R C, Melo M C R and Schulten. K 2015 Enhanced sampling techniques in molecular dynamics simulations of biological systems *Biochim. Biophys. Acta* **1850** 872–7
[112] Valsson O, Tiwary P and Parrinello M 2016 Enhancing important fluctuations: rare events and metadynamics from a conceptual viewpoint *Annu. Rev. Phys. Chem.* **67** 159–84
[113] Camilloni C and Pietrucci F 2018 Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems *Adv. Phys.* X **3** 1477531
[114] Kingma D P and Welling M 2013 Auto-encoding variational bayes (arXiv:1312.6114)
[115] Chen W and Ferguson A L 2018 Molecular enhanced sampling with autoencoders: on-the-fly collective variable discovery and accelerated free energy landscape exploration *J. Comput. Chem.* **39** 2079–102
[116] Sultan M M, Wayment-Steele H K and Pande V S 2018 Transferable neural networks for enhanced sampling of protein dynamics *J. Chem. Theory Comput.* **14** 1887–94
[117] Wehmeyer C and Noé F 2018 Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics *J. Chem. Phys.* **148** 241703
[118] Ribeiro J M L, Bravo P, Wang Y and Tiwary P 2018 Reweighted autoencoded variational bayes for enhanced sampling (RAVE) *J. Chem. Phys.* **149** 072301
[119] Chiavazzo E, Covino R, Coifman R R, Gear C W, Georgiou A S, Hummer G and Kevrekidis I G 2017 Intrinsic map dynamics exploration for uncharted effective free-energy landscapes *Proc. Natl. Acad. Sci.* **114** E5494–503
[120] Kukharenko O, Sawade K, Steuer J and Peter C 2016 Using dimensionality reduction to systematically expand conformational sampling of intrinsically disordered peptides *J. Chem. Theory Comput.* **12** 4726–34
[121] Ceriotti M, Tribello G A and Parrinello M 2011 Simplifying the representation of complex free-energy landscapes using sketch-map *Proc. Natl. Acad. Sci.* **108** 13023–8
[122] Lemke T and Peter C 2019 EncoderMap: dimensionality reduction and generation of molecule conformations *J. Chem. Theory Comput.* **15** 1209–15
[123] Perez A, MacCallum J L and Dill K A 2015 Accelerating molecular simulations of proteins using bayesian inference on weak information *Proc. Natl. Acad. Sci.* **112** 11846–51
[124] Lipinski C A 2004 Lead- and drug-like compounds: the rule-of-five revolution *Drug Discovery Today: Technol.* **1** 337–41
[125] Ruddigkeit L, van Deursen R, Blum L C and Reymond J-L 2012 Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17 *J. Chem. Inf. Model.* **52** 2864–75
[126] Ramakrishnan R, Dral P O, Rupp M and Von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules *Sci. Data* **1** 140022
[127] Hoja J, Sandonas L M, Ernst B G, Vazquez-Mayagoitia A, DiStasio R A Jr and Tkatchenko A 2020 QM7-x: a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules (arXiv:2006.15139)
[128] Grest G S and Kremer K 1986 Molecular dynamics simulation for polymers in the presence of a heat bath *Phys. Rev.* A **33** 3628–31
[129] Kremer K and Grest G S 1990 Dynamics of entangled linear polymer melts: a molecular-dynamics simulation *J. Chem. Phys.* **92** 5057–86
[130] Zhang G, Stuehn T, Daoulas K C and Kremer K 2015 Communication: one size fits all: equilibrating chemically different polymer liquids through universal long-wavelength description *J. Chem. Phys.* **142** 221102
[131] Everaers R, Karimi-Varzaneh H A, Fleck F, Hojdis N and Svaneborg C 2020 Kremer–Grest models for commodity polymer melts: linking theory, experiment, and simulation at the Kuhn scale *Macromolecules* **53** 1901–16
[132] Bereau T 2014 Auto_martini repository https://github.com/tbereau/auto_martini

[133] Virshup A M, Contreras-García J, Wipf P, Yang W and Beratan D N 2013 Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds *J. Am. Chem. Soc.* **135** 7296–303
[134] Hoksza D, Škoda P, Voršilák M and Svozil D 2014 Molpher: a software framework for systematic chemical space exploration *J. Cheminformatics* **6** 7
[135] Joyce G F 1992 Directed molecular evolution *Sci. Am.* **267** 90–7
[136] Chowdhury R and Maranas C D 2019 From directed evolution to computational enzyme engineering—a review *AIChE J.* **66** e16847
[137] Hoffmann C, Menichetti R, Kanekal K H and Bereau T 2019 Controlled exploration of chemical space by machine learning of coarse-grained representations *Phys. Rev.* E **100** 033302
[138] Wang L-P, Titov A, McGibbon R, Liu F, Pande V S and Martínez T J 2014 Discovering chemistry with an *ab initio* nanoreactor *Nat. Chem.* **6** 1044–8
[139] Meisner J, Zhu X and Martínez T J 2019 Computational discovery of the origins of life *ACS Cent. Sci.* **5** 1493–5
[140] Mobley D L and Klimovich P V 2012 Perspective: alchemical free energy calculations for drug discovery *J. Chem. Phys.* **137** 230901
[141] von Lilienfeld O A 2009 Accurate *ab initio* energy gradients in chemical compound space *J. Chem. Phys.* **131** 164102
[142] Balawender R, Welearegay M A, Lesiuk M ł, De Proft F and Geerlings P 2013 Exploring chemical space with the alchemical derivatives *J. Chem. Theory Comput.* **9** 5327–40
[143] Baben M t, Achenbach J O and von Lilienfeld O A 2016 Guiding *ab initio* calculations by alchemical derivatives *J. Chem. Phys.* **144** 104103
[144] Wang M, Hu X, Beratan D N and Yang W 2006 Designing molecules by optimizing potentials *J. Am. Chem. Soc.* **128** 3228–32
[145] von Lilienfeld O A, Lins R D and Rothlisberger U 2005 Variational particle number approach for rational compound design *Phys. Rev. Lett.* **95** 153002
[146] Gómez-Bombarelli R *et al* 2018 Automatic chemical design using a data-driven continuous representation of molecules *ACS Cent. Sci.* **4** 268–76
[147] Shmilovich K, Mansbach R A, Sidky H, Dunne O E, Panda S S, Tovar J D and Ferguson A L 2020 Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation *J. Phys. Chem.* B **124** 3873–91
[148] Woelfle M, Olliaro P and Todd M H 2011 Open science is a research accelerator *Nat. Chem.* **3** 745–8
[149] Huan T D, Mannodi-Kanakkithodi A, Kim C, Sharma V, Pilania G and Ramprasad R 2016 A polymer dataset for accelerated property prediction and design *Sci. Data* **3** 160012
[150] Audus D J and de Pablo J J 2017 Polymer informatics: opportunities and challenges *ACS Macro Lett.* **6** 1078–82
[151] Barnett J W, Bilchak C R, Wang Y, Benicewicz B C, Murdock L A, Bereau T and Kumar S K 2020 Designing exceptional gas-separation polymer membranes using machine learning *Sci. Adv.* **6** eaaz4301
[152] European Organization For Nuclear Research and OpenAIRE 2013 Zenodo https://zenodo.org
[153] figshare LLP 2011 Figshare https://figshare.com
[154] Center for Open Science 2011 Open science framework https://osf.io
[155] Research Collaboratory for Structural Bioinformatics PDB 2000 Protein data bank https://rcsb.org
[156] Hoffmann C, Centi A, Menichetti R and Bereau T 2020 Molecular dynamics trajectories for 630 coarse-grained drug-membrane permeations *Sci. Data* **7** 1–7
[157] Wilkinson M D *et al* 2016 The FAIR guiding principles for sci. data management and stewardship *Sci. Data* **3** 160018
[158] Draxl C and Scheffler M 2020 Big data-driven materials science and its fair data infrastructure *Handb. Mater. Model.* 49–73
[159] Blaiszik B, Chard K, Pruyne J, Ananthakrishnan R, Tuecke S and Foster I 2016 The materials data facility: data services to advance materials science research *JOM* **68** 2045–52
[160] Jain A *et al* 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002
[161] Draxl C and Scheffler M 2018 Nomad: the fair concept for big data-driven materials science *MRS Bull.* **43** 676–82
[162] Tadmor E B, Elliott R S, Sethna J P, Miller R E and Becker C A 2011 The potential of atomistic simulations and the knowledgebase of interatomic models *JOM* **63** 17

[163] Tadmor E B, Elliott R S, Phillpot S R and Sinnott S B 2013 NSF cyberinfrastructures: a new paradigm for advancing materials simulation *Curr. Opin. Solid State Mater. Sci.* **17** 298–304
[164] MolSSI 2016 The molecular sciences software institute https://molssi.org
[165] FAIR-DI e.V. 2018 FAIR-DI https://fair-di.eu
[166] Lo Y-C, Rensi S E, Torng W and Altman R B 2018 Machine learning in chemoinformatics and drug discovery *Drug Discov. Today* **23** 1538–46
[167] Swift R V and Amaro R E 2013 Back to the future: can physical models of passive membrane permeability help reduce drug candidate attrition and move us beyond QSPR? *Chem. Biol. Drug Des.* **81** 61–71
[168] Zhang L, Tan J, Han D and Zhu H 2017 From machine learning to deep learning: progress in machine intelligence for rational drug discovery *Drug Discov. Today* **22** 1680–5
[169] Thurston B A and Ferguson A L 2018 Machine learning and molecular design of self-assembling -conjugated oligopeptides *Mol. Simul.* **44** 930–45
[170] Brunton S L, Proctor J L and Kutz J N 2016 Discovering governing equations from data by sparse identification of nonlinear dynamical systems *Proc. Natl. Acad. Sci.* **113** 3932–7
[171] Udrescu S-M and Tegmark M 2020 AI Feynman: a physics-inspired method for symbolic regression *Sci. Adv.* **6** eaay2631
[172] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 Big data of materials science: critical role of the descriptor *Phys. Rev. Lett.* **114** 105503
[173] Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M and Ghiringhelli L M 2018 Sisso: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates *Phys. Rev. Mat.* **2** 083802
[174] Goldsmith B R, Boley M, Vreeken J, Scheffler M and Ghiringhelli L M 2017 Uncovering structure-property relationships of materials by subgroup discovery *New J. Phys.* **19** 013031
[175] Ceriotti M 2019 Unsupervised machine learning in atomistic simulations, between predictions and understanding *J. Chem. Phys.* **150** 150901
[176] Faber F A *et al* 2017 Prediction errors of molecular machine learning models lower than hybrid dft error *J. Chem. Theory Comput.* **13** 5255–64
[177] Ramakrishnan R and von Lilienfeld O A 2017 *Machine Learning, Quantum Chemistry, and Chemical Space* (New York: Wiley) pp 225–56
[178] Rasmussen C E and Williams C K I 2006 *Gaussian Processes for Machine Learning* vol 1 (Cambridge: MIT press Cambridge)
[179] Bartók A P, De S, Poelking C, Bernstein N, Kermode J R, Csányi G and Ceriotti M 2017 Machine learning unifies the modeling of materials and molecules *Sci. Adv.* **3** e1701816
[180] Faber F A, Christensen A S, Huang B and von Lilienfeld O A 2018 Alchemical and structural distribution based representation for universal quantum machine learning *J. Chem. Phys.* **148** 241717
[181] von Lilienfeld O A 2018 Quantum machine learning in chemical compound space *Angew. Chem., Int. Ed. Engl.* **57** 4164–9
[182] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces *Phys. Rev. Lett.* **98** 146401
[183] Huang B and von Lilienfeld O A 2016 Communication: understanding molecular representations in machine learning: the role of uniqueness and target similarity *J. Chem. Phys.* **145** 161102
[184] Glielmo A, Sollich P and De Vita A 2017 Accurate interatomic force fields via machine learning with covariant kernels *Phys. Rev.* B **95** 214302
[185] Bartók A P and Csányi G 2015 Gaussian approximation potentials: a brief tutorial introduction *Int. J. Quantum Chem.* **115** 1051–7
[186] Scherer C, Scheid R, Andrienko D and Bereau T 2020 Kernel-based machine learning for efficient simulations of molecular liquids *J. Chem. Theory Comput.* **16** 3194–204
[187] Veit M, Wilkins D M, Yang Y, DiStasio R A and Ceriotti M 2020 Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles *J. Chem. Phys.* **153** 024113
[188] Rauer C and Bereau T 2020 Hydration free energies from kernel-based machine learning: compound-database bias *J. Chem. Phys.* **153** 014101
[189] Gkeka P *et al* 2020 Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems (arXiv:2004.06950)
[190] John S T and Csányi G 2017 Many-body coarse-grained interactions using Gaussian approximation potentials *J. Phys. Chem* B **121** 10934–49

[191] Wang J, Olsson S, Wehmeyer C, Pérez A, Charron N E, de Fabritiis G, Noé F and Clementi C 2019 Machine learning of coarse-grained molecular dynamics force fields *ACS Cent. Sci.* **5** 755–67
[192] Wang J, Chmiela S, Müller K-R, Noé F and Clementi C 2020 Ensemble learning of coarse-grained molecular dynamics force fields with a kernel approach *J. Chem. Phys.* **152** 194106
[193] Pipolo S, Salanne M, Ferlat G, Klotz S, Saitta A M and Pietrucci F 2017 Navigating at will on the water phase diagram *Phys. Rev. Lett.* **119** 245701
[194] Grisafi A and Ceriotti M 2019 Incorporating long-range physics in atomic-scale machine learning *J. Chem. Phys.* **151** 204105
[195] Sejnowski T J 2020 The unreasonable effectiveness of deep learning in artificial intelligence *Proc. Natl. Acad. Sci.* **117** 30033–8
[196] Goh G B, Hodas N O and Vishnu A 2017 Deep learning for computational chemistry *J. Comput. Chem.* **38** 1291–307
[197] Chen H, Engkvist O, Wang Y, Olivecrona M and Blaschke T 2018 The rise of deep learning in drug discovery *Drug Discov. Today* **23** 1241–50
[198] Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K and Riley P 2018 Tensor field networks: rotation-and translation-equivariant neural networks for 3d point clouds (arXiv:1802.08219)
[199] Kondor R and Trivedi S 2018 On the generalization of equivariance and convolution in neural networks to the action of compact groups (arXiv:1802.03690)
[200] Raissi M, Perdikaris P and Karniadakis G E 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations *J. Comput. Phys.* **378** 686–707
[201] Duvenaud D K, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A and Adams R P 2015 Convolutional networks on graphs for learning molecular fingerprints *Advances in Neural Information Processing Systems* pp 2224–32
[202] Bennett W F D, He S, Bilodeau C L, Jones D, Sun D, Kim H, Allen J, Lightstone F C and Ingólfsson H I 2020 Predicting small molecule transfer free energies by combining molecular dynamics simulations and deep learning *J. Chem. Inf. Model.* **60** 5375–81
[203] Shirts M and Pande V S 2000 COMPUTING: screen savers of the world unite! *Science* **290** 1903–4
[204] Pande V S *et al* 2002 Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing *Biopolymers* **68** 91–109
[205] Snow C D, Nguyen H, Pande V S and Gruebele M 2002 Absolute comparison of simulated and experimental protein-folding dynamics *Nature* **420** 102–6
[206] Lane T J, Shukla D, Beauchamp K A and Pande V S 2013 To milliseconds and beyond: challenges in the simulation of protein folding *Curr. Opin. Struct. Biol.* **23** 58–65
[207] Noé F 2008 Probability distributions of molecular observables computed from markov models *J. Chem. Phys.* **128** 244103
[208] Pande V S, Beauchamp K and Bowman G R 2010 Everything you wanted to know about markov state models but were afraid to ask *Methods* **52** 99–105
[209] Bowman G R, Pande V S and Noé F (ed) 2014 *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Netherlands: Springer)
[210] Husic B E and Pande V S 2018 Markov state models: from an art to a science *J. Am. Chem. Soc.* **140** 2386–96
[211] Buch I, Sadiq S K and De Fabritiis G 2011 Optimized potential of mean force calculations for standard binding free energies *J. Chem. Theory Comput.* **7** 1765–72
[212] Buch I, Harvey M J, Giorgino T, Anderson D P and De Fabritiis G 2010 High-throughput all-atom molecular dynamics simulations using distributed computing *J. Chem. Inf. Model.* **50** 397–403
[213] Buch I, Giorgino T and De Fabritiis G 2011 Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations *Proc. Natl. Acad. Sci.* **108** 10184–9
[214] Plattner N, Doerr S, De Fabritiis G and Noé F 2017 Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling *Nat. Chem.* **9** 1005–11
[215] Jorgensen W L 2004 The many roles of computation in drug discovery *Science* **303** 1813–8
[216] Boehr D D, Nussinov R and Wright P E 2009 The role of dynamic conformational ensembles in biomolecular recognition *Nat. Chem. Biol.* **5** 789–96
[217] De Vivo M, Masetti M, Bottegoni G and Cavalli A 2016 Role of molecular dynamics and related methods in drug discovery *J. Med. Chem.* **59** 4035–61

[218] Brown S P and Muchmore S W 2009 Large-scale Application of high-throughput molecular mechanics with Poisson−Boltzmann surface area for routine physics-based scoring of Protein−Ligand complexes *J. Med. Chem.* **52** 3159–65
[219] Jorgensen W L 2009 Efficient drug lead discovery and optimization *Acc. Chem. Res.* **42** 724–33
[220] Borhani D W and Shaw D E 2011 The future of molecular dynamics simulations in drug discovery *J. Comput. Aided Mol. Des.* **26** 15–26
[221] Chipot C and Pearlman D A 2002 Free energy calculations. the long and winding gilded road *Mol. Simul.* **28** 1–12
[222] Michel J and Essex J W 2010 Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations *J. Comput. Aided Mol. Des.* **24** 639–58
[223] Zwanzig R W 1954 High-temperature equation of state by a perturbation method: I. Nonpolar gases *J. Chem. Phys.* **22** 1420–6
[224] Wong C F and McCammon J A 1986 Dynamics and design of enzymes and inhibitors *J. Am. Chem. Soc.* **108** 3830–2
[225] Wang L *et al* 2015 Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field *J. Am. Chem. Soc.* **137** 2695–703
[226] Liu S, Wu Y, Lin T, Abel R, Redmann J P, Summa C M, Jaber V R, Lim N M and Mobley D L 2013 Lead optimization mapper: automating free energy calculations for lead optimization *J. Comput. Aided Mol. Des.* **27** 755–70
[227] Abel R, Wang L, Harder E D, Berne B J and Friesner R A 2017 Advancing drug discovery through enhanced free energy calculations *Acc. Chem. Res.* **50** 1625–32
[228] Jorgensen W L and Ravimohan C 1985 Monte Carlo simulation of differences in free energies of hydration *J. Chem. Phys.* **83** 3050–4
[229] Huang D M, Geissler P L and Chandler D 2001 Scaling of hydrophobic solvation free energies† *J. Phys. Chem.* B **105** 6704–9
[230] Villa A and Mark A E 2002 Calculation of the free energy of solvation for neutral analogs of amino acid side chains *J. Comput. Chem.* **23** 548–53
[231] MacCallum J L and Tieleman D P 2003 Calculation of the water-cyclohexane transfer free energies of neutral amino acid side-chain analogs using the OPLS all-atom force field *J. Comput. Chem.* **24** 1930–5
[232] Shirts M R, Pitera J W, Swope W C and Pande V S 2003 Extremely precise free energy calculations of amino acid side chain analogs: comparison of common molecular mechanics force fields for proteins *J. Chem. Phys.* **119** 5740–61
[233] Shirts M R and Pande V S 2005 Solvation free energies of amino acid side chain analogs for common molecular mechanics water models *J. Chem. Phys.* **122** 134508
[234] Mobley D L, Dumont É, Chodera J D and Dill K A 2007 Comparison of charge models for fixed-charge force fields: small-molecule hydration free energies in explicit solvent *J. Phys. Chem.* B **111** 2242–54
[235] Shivakumar D, Deng Y and Roux B 2009 Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model *J. Chem. Theory Comput.* **5** 919–30
[236] Shivakumar D, Williams J, Wu Y, Damm W, Shelley J and Sherman W 2010 Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field *J. Chem. Theory Comput.* **6** 1509–19
[237] Shivakumar D, Harder E, Damm W, Friesner R A and Sherman W 2012 Improving the prediction of absolute solvation free energies using the next generation OPLS force field *J. Chem. Theory Comput.* **8** 2553–8
[238] Mobley D L, Bayly C I, Cooper M D, Shirts M R and Dill K A 2009 Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations *J. Chem. Theory Comput.* **5** 350–8
[239] Souza P C T, Thallmair S, Conflitti P, Ramírez-Palacios C, Alessandri R, Raniolo S, Limongelli V and Marrink S J 2020 Protein–ligand binding with the coarse-grained martini model *Nat. Commun.* **11** 3714
[240] Nicholls A, Mobley D L, Guthrie J P, Chodera J D, Bayly C I, Cooper M D and Pande V S 2008 Predicting small-molecule solvation free energies: an informal blind test for computational chemistry *J. Med. Chem.* **51** 769–79
[241] Guthrie J P 2009 A blind challenge for computational solvation free energies: introduction and overview *J. Phys. Chem.* B **113** 4501–7

[242] SAMPL Challenges. samplchallenges.github.io.
[243] Geballe M T, Skillman A G, Nicholls A, Guthrie J P and Taylor P J 2010 The SAMPL2 blind prediction challenge: introduction and overview *J. Comput. Aided Mol. Des.* **24** 259–79
[244] Mobley D L, Wymer K L, Lim N M and Guthrie J P 2014 Blind prediction of solvation free energies from the SAMPL4 challenge *J. Comput. Aided Mol. Des.* **28** 135–50
[245] Armand M, Endres F, MacFarlane D R, Ohno H and Scrosati B 2009 Ionic-liquid materials for the electrochemical challenges of the future *Nat. Mater.* **8** 621–9
[246] Maginn E J 2009 Molecular simulation of ionic liquids: current status and future opportunities *J. Phys.: Condens. Matter* **21** 373101
[247] Wang Y, Jiang W, Yan T and Voth G A 2007 Understanding ionic liquids through atomistic and coarse-grained molecular dynamics simulations *Acc. Chem. Res.* **40** 1193–9
[248] Lynden-Bell R M, Del Pópolo M G, Youngs T G A, Kohanoff J, Hanke C G, Harper J B and Pinilla C C 2007 Simulations of ionic liquids, solutions, and surfaces *Acc. Chem. Res.* **40** 1138–45
[249] Bhargava B L, Balasubramanian S and Klein M L 2008 Modelling room temperature ionic liquids *Chem. Commun.* **29** 3339–51
[250] Osti N C, Van Aken K L, Thompson M W, Tiet F, Jiang D-e, Cummings P T, Gogotsi Y and Mamontov E 2016 Solvent polarity governs ion interactions and transport in a solvated room-temperature ionic liquid *J. Phys. Chem. Lett.* **8** 167–71
[251] Thompson M W, Matsumoto R, Sacci R L, Sanders N C and Cummings P T 2019 Scalable screening of soft matter: a case study of mixtures of ionic liquids and organic solvents *J. Phys. Chem.* B **123** 1340–7
[252] Paul A 1989 *Chemistry of Glasses* (Berlin: Springer Science & Business Media)
[253] Wondraczek L, Mauro J C, Eckert J, Kühn U, Horbach J, Deubener J and Rouxel T 2011 Towards ultrastrong glasses *Adv. Mater.* **23** 4578–86
[254] Yang K, Xu X, Yang B, Cook B, Ramos H, Krishnan N M A, Smedskjaer M M, Hoover C and Bauchy M 2019 Predicting the Young's modulus of silicate glasses using high-throughput molecular dynamics simulations and machine learning *Sci. Rep.* **9** 8739
[255] Bouhadja M, Jakse N and Pasturel A 2013 Structural and dynamic properties of calcium aluminosilicate melts: a molecular dynamics study *J. Chem. Phys.* **138** 224510
[256] Fagerberg L, Jonasson K, von Heijne G, Uhlén M and Berglund L 2010 Prediction of the human membrane proteome *Proteomics* **10** 1141–9
[257] Bakheet T M and Doig A J 2009 Properties and identification of human protein drug targets *Bioinformatics* **25** 451–7
[258] White S H 2004 The progress of membrane protein structure determination *Protein Sci.* **13** 1948–9
[259] Moraes I, Evans G, Sanchez-Weatherby J, Newstead S and Stewart P D S 2014 Membrane protein structure determination-the next generation *Biochim. Biophys. Acta Biomembr.* **1838** 78–87
[260] Kandt C, Ash W L and Peter Tieleman D P 2007 Setting up and running molecular dynamics simulations of membrane proteins *Methods* **41** 475–88
[261] Ayton G S and Voth G A 2009 Systematic multiscale simulation of membrane protein systems *Curr. Opin. Struct. Biol.* **19** 138–44
[262] Chavent M, Duncan A L and Sansom M S 2016 Molecular dynamics simulations of membrane proteins and their interactions: from nanoscale to mesoscale *Curr. Opin. Struct. Biol.* **40** 8–16
[263] Jefferies D and Khalid S 2020 Atomistic and coarse-grained simulations of membrane proteins: a practical guide *Methods*
[264] Im W and Brooks C L 2005 Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations *Proc. Natl. Acad. Sci.* **102** 6771–6
[265] Bereau T and Deserno M 2014 Enhanced sampling of coarse-grained transmembrane-peptide structure formation from hydrogen-bond replica exchange *J. Membr. Biol.* **248** 395–405
[266] Bereau T, Bennett W F D, Pfaendtner J, Deserno M and Karttunen M 2015 Folding and insertion thermodynamics of the transmembrane WALP peptide *J. Chem. Phys.* **143** 243127
[267] Sansom M S, Scott K A and Bond P J 2008 Coarse-grained simulation: a high-throughput computational approach to membrane proteins *Biochem. Soc. Trans.* **36** 27–32
[268] Berman H M 2000 The protein data bank *Nucleic Acids Res.* **28** 235–42
[269] Jo S, Lim J B, Klauda J B and Im W 2009 CHARMM-GUI membrane builder for mixed bilayers and its application to yeast membranes *Biophys. J.* **97** 50–8
[270] Wassenaar T A, Ingólfsson H I, Böckmann R A, Tieleman D P and Marrink S J 2015 Computational lipidomics with insane: a versatile tool for generating custom membranes for molecular simulations *J. Chem. Theory Comput.* **11** 2144–55

[271] Wassenaar T A, Pluhackova K, Moussatova A, Sengupta D, Marrink S J, Tieleman D P and Böckmann R A 2015 High-throughput simulations of dimer and trimer assembly of membrane proteins. The DAFT approach *J. Chem. Theory Comput.* **11** 2278–91
[272] Monticelli L, Kandasamy S K, Periole X, Larson R G, Tieleman D P and Marrink S-J 2008 The MARTINI coarse-grained force field: extension to proteins *J. Chem. Theory Comput.* **4** 819–34
[273] Newport T D, Sansom M S P and Stansfeld P J 2018 MemProtMD Database http://memprotmd.bioch.ox.ac.uk/home/
[274] Stansfeld P J, Goose J E, Caffrey M, Carpenter E P, Parker J L, Newstead S and Sansom M S P 2015 MemProtMD: automated insertion of membrane protein structures into explicit lipid membranes *Structure* **23** 1350–61
[275] Newport T D, Sansom M S P and Stansfeld P J 2018 The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions *Nucleic Acids Res.* **47** D390–7
[276] Corradi V *et al* 2018 Lipid-protein interactions are unique fingerprints for membrane proteins *ACS Cent. Sci.* **4** 709–17
[277] Zelzer M and Ulijn R V 2010 Next-generation peptide nanomaterials: molecular networks, interfaces and supramolecular functionality *Chem. Soc. Rev.* **39** 3351
[278] Uhlig T, Kyprianou T, Martinelli F G, Oppici C A, Heiligers D, Hills D, Calvo X R and Verhaert P 2014 The emergence of peptides in the pharmaceutical business: from exploration to exploitation *EuPA Open Proteomics* **4** 58–69
[279] Burroughes J H, Bradley D D C, Brown A R, Marks R N, Mackay K, Friend R H, Burns P L and Holmes A B 1990 Light-emitting diodes based on conjugated polymers *Nature* **347** 539–41
[280] Koss K and Unsworth L 2018 Towards developing bioresponsive, self-assembled peptide materials: dynamic morphology and fractal nature of nanostructured matrices *Materials* **11** 1539
[281] Hartgerink J D 2001 Self-assembly and mineralization of peptide-amphiphile nanofibers *Science* **294** 1684–8
[282] Smith A M, Williams R J, Tang C, Coppo P, Collins R F, Turner M L, Saiani A and Ulijn R V 2008 Fmoc-diphenylalanine self assembles to a hydrogel via a novel architecture based on $\pi-\pi$ interlocked β-sheets *Adv. Mater.* **20** 37–41
[283] Görbitz C H 2006 The structure of nanotubes formed by diphenylalanine, the core recognition motif of Alzheimer's β-amyloid polypeptide *Chem. Commun.* **22** 2332–4
[284] Frederix P W J M, Scott G G, Abul-Haija Y M, Kalafatovic D, Pappas C G, Javid N, Hunt N T, Ulijn R V and Tuttle T 2015 Exploring the sequence space for (tri-)peptide self-assembly to design and discover new hydrogels *Nat. Chem.* **7** 30
[285] Chiti F, Stefani M, Taddei N, Ramponi G and Dobson C M 2003 Rationalization of the effects of mutations on peptide andprotein aggregation rates *Nature* **424** 805–8
[286] Marchesan S, Easton C D, Kushkaki F, Waddington L and Hartley P G 2012 Tripeptide self-assembled hydrogels: unexpected twists of chirality *Chem. Commun.* **48** 2195–7
[287] Yap C W 2010 PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints *J. Comput. Chem.* **32** 1466–74
[288] Wall B D and Tovar J D 2012 Synthesis and characterization of π-conjugated peptide-based supramolecular materials *Pure Appl. Chem.* **84** 1039–45
[289] Diamond J M and Katz Y 1974 Interpretation of nonelectrolyte partition coefficients between dimyristoyl lecithin and water *J. Membrain Biol.* **17** 121–54
[290] Marrink S-J and Berendsen H J C 1994 Simulation of water transport through a lipid membrane *J. Phys. Chem.* **98** 4155–68
[291] Orsi M and Essex J W 2010 Chapter 4 passive permeation across lipid bilayers: a literature review *Molecular Simulations and Biomembranes: From Biophysics to Function* (Cambridge: The Royal Society of Chemistry) pp 76–90
[292] Carpenter T S, Kirshner D A, Lau E Y, Wong S E, Nilmeier J P and Lightstone F C 2014 A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations *Biophys. J.* **107** 630–41
[293] Lee C T *et al* 2016 Simulation-based approaches for determining membrane permeability of small compounds *J. Chem. Inf. Model.* **56** 721–33
[294] Bennion B J *et al* 2017 Predicting a drug's membrane permeability: a computational model validated with *in Vitro* permeability assay data *J. Phys. Chem.* B **121** 5228–37
[295] Tse C H, Comer J, Wang Y and Chipot C 2018 Link between membrane composition and permeability to drugs *J. Chem. Theory Comput.* **14** 2895–909

[296] Menichetti R, Kanekal K H, Kremer K and Bereau T 2017 In silico screening of drug-membrane thermodynamics reveals linear relations between bulk partitioning and the potential of mean force *J. Chem. Phys.* **147** 125101
[297] Rudzinski J F 2019 Recent progress towards chemically-specific coarse-grained simulation models with consistent dynamical properties *Computation* **7** 42
[298] Menichetti R and Bereau T 2019 Revisiting the Meyer-Overton rule for drug-membrane permeabilities *Mol. Phys.* **117** 2900–9
[299] Centi A, Dutta A, Parekh S H and Bereau T 2020 Inserting small molecules across membrane mixtures: insight from the potential of mean force *Biophys. J.* **118** 1321–32
[300] Cornell C E, McCarthy N L C, Levental K R, Levental I, Brooks N J and Keller S L 2017 n -alcohol length governs shift in L o - L d mixing temperatures in synthetic and cell-derived membranes *Biophys. J.* **113** 1200–11