The Middle Science: Traversing Scale In Complex Many-Body Systems

1. THE MANY-BODY CHALLENGE

Nearly 60 years ago, physicist Richard Feynman gave a now famous presentation, entitled “There’s Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics.” In considering the manipulation of matter on an atom-by-atom basis, Feynman simultaneously issued the challenge of the ultimate miniaturization of technology while removing from the picture a far more potent demon—the many-body problem. Within real physical systems, all particles interact, whether it is in the correlated motions of electrons or the self-assembly of amphiphilic molecules to form an emulsion. By significant ingenuity, physicists, chemists, and materials scientists have defanged the many-body demon and accelerated the progress of science through systematically improved approximations that encompass only a few sets of interactions at a time (for example, through the mean field approximations or perturbation theories). However, it is well acknowledged that such approaches can come at a great cost, including the inability to predict the phase diagram of strongly correlated metals such as plutonium, the grand challenge of accurately simulating magnetic states and quantum phase transitions within nano- and quantum materials, and developing an understanding of excited state correlation effects within large photosystems relevant to light harvesting. The challenge leveled by Feynman in 1959 has led to the development of molecular machinery capable of carrying out complex chemical reactions that underpin not only exascale computing but also the so-called quantum revolution. Ironically, this advanced computing architecture lays the foundation for moving away from the science at the “bottom” and into a “middle science” that has the potential to use computing power combined with the latest advancements of data science (DS) to create computationally tractable approaches that account for the full impact of many-body interactions across lengths and time scales. Imagine near linear-scaling electronic structure methods that include high-level many-body treatments via coarse-grained Hamiltonians derived from machine learning (ML) guided parametrization. Alternatively, consider classical force field simulations of chemically extreme solution phase systems, where attributed graphs support prediction of phase phenomena that occur on incredibly long time scales. This article provides a roadmap of challenges and potential methods that seek to empower many-body theories. Advances in discrete and computational mathematics and the integration of new mathematics methods with physics-aware learning strategies are emphasized. The goal is to extend the impact of many-body theories beyond traditional domains of electronic structure theory and into complex condensed matter systems, where propagating system states in time must be able to self-consistently account for many-body effects (described by different granularities of information). As we traverse the many-body roadmap—from small to big, across space, and through time—we will introduce tools from DS and mathematics that are currently being used and which show potential for future development. These include extensions to chemical graph theory, where network structures of molecular vertices connected by edges of intermolecular interactions can model emergent phenomenon in many-body systems. The discrete notion of spatial proximity measured by graphs becomes continuous and higher-dimensional in the mathematical fields of geometry and topology that measure...
2. THE MANY-BODY ROADMAP

2.1. When Small Is Big. 2.1.1. Curse of Dimensionality and the Art of Coarse-Graining. Quantum mechanics provides a complete description of chemistry, biology, and materials. By simply specifying the particles (nuclei and electrons), a many-body Hamiltonian can be written alongside the time-dependent Schrödinger (or Dirac) equation whose solutions define any imaginable physical property of the system. This is a truly amazing achievement of theoretical physics and a major milestone of human innovation. Unfortunately, in between the initial knowledge of the interacting particles and the macroscopic physical properties of the system lies an object of enormous complexity, the unknown many-body wave function determined by the Schrödinger equation. The exact wave function is exponentially complex because of quantum entanglement between particles; this translates into the huge size of the wave function, whether represented in real space or in the space of parameters. Both the size of the wave function and the cost of solving the equations scale factorially with the system size—giving rise to the curse of dimensionality. Quantum chemistry has been very successful in cutting in-roads through the intimidating landscape of the many-body wave function by developing mathematically rigorous hierarchies of approximations.11 Approximate treatments of electron correlation, in combination with algorithmic developments (e.g., fast reduced-complexity solvers, lossless and lossy compression, powerful tensor libraries) and ever-increasing computational power, enable applications of quantum theory to real-world problems, ranging from high-accuracy thermochemistry and simulated spectroscopy to materials modeling and drug design. Quantum chemistry can be defined as a science (or sometimes an art) of coarse-graining the complexity of the many-body problem. To illustrate this point, consider one mean-field theory (also known as the Hartree–Fock method)—it reduces the full many-body system to single particles interacting through an effective medium. Instead of a multidimensional many-body wave function, one can deal with individual particles that interact through an effective (averaged) field generated by all other particles. Despite such a drastic simplification, the mean-field treatment provides sufficiently accurate answers to many practical questions. Quantum chemistry uses it to build more sophisticated treatments, by using the states of (pseudo) noninteracting particles (Slater determinants composed of the Hartree–Fock orbitals) to build a many-body basis that allows one to gradually include more fine-grained interactions between the particles (electron correlation). Alternatively, mean-field theory can be used as a starting point for more approximated treatments, such as Hückel or Pariser–Parr–Pople models.

This idea is illustrated in Figure 1, where a detailed cat theory, including whiskers and all, is replaced by an effective cat theory that can provide sufficiently accurate description of the cat’s behavior. This effective description can be used as a stepping stone to build a more detailed description of the cat behaviors and inform us when to feed the cat, when to take it to the vet, or how to play with it. This mapping can be done either exactly or approximately. The coarse-grained, effective representation of a cat can be used as a starting point for including more fine-grained descriptions of the system (with more nuanced cat behaviors) or for building up the complexity, all the way up to a multiple cat theory and one for dynamically interacting cats.
or as a building block of a theory describing many cats. That is, an effective cat model can be used to construct a product space describing multiple cats and used to capture their interactions and dynamics (time dependence). This is an example of the bottom-up coarse-graining in which the model is rigorously derived from a more detailed description; however, one can also construct the model top-down by matching relevant experimental observables. For example, one can attempt to create an effective cat model phenomenologically by requiring that the model correctly predicts selected cat behaviors.

Quantum chemistry can be defined as a science (or sometimes an art) of coarse-graining the complexity of the many-body problem.

Coarse-graining of complexity (moving from left to right in Figure 1) can be carried out in a variety of ways: by hierarchical improvements in correlation treatment by including clusters of higher excitations (as in coupled-cluster theory) iteratively compressing/expanding the Hilbert space (as in density matrix renormalization group and other adaptive configuration interaction schemes), by using a Bloch formalism to construct effective Hamiltonians mapping the many-electron problem onto a small number of effective spins (as in treatments of magnetic phenomena), or by introducing more drastic shortcuts connecting the target (energy and properties) with the input (information about nuclei and electrons). A famous example from the latter category is density functional theory (DFT), which avoids the correlated many-body wave function altogether and instead operates with an energy functional, a mathematical device that can yield the energy directly from the electron density. The success of DFT is grounded in the fact that such exact mapping between the density and energy exists, as proven mathematically by the Hohenberg—Kohn theorems, or more vividly, by Wilson’s deduction. Since one can determine from the electron density the number of electrons (by integrating over the density) and the types and positions of the nuclei (from the cusps), and then one can write the Hamiltonian, solve the resulting Schrödinger equation, and obtain the exact energy. Although the existence of such a mapping does not provide a practical workable recipe, it justifies an empirical approach to the problem, that is, building the unknown functional by fitting its form and parameters to satisfy known exact properties of real or idealized systems (e.g., the uniform electron gas or the helium atom) or to directly reproduce experimental data. In the same spirit, semiempirical theories (tight-binding, SCC-DFTB, MNDO, INDO, CODA, PM3, AM1, Hubbard’s and Heisenberg Hamiltonians, classical force-fields) coarse-grain the complexity by representing the full interacting system of nuclei and electrons by much more compact effective Hamiltonians, or other coarse-grained descriptors like fictitious spins, “atoms” connected by springs and interacting by electrostatics and dispersion, or beads (groups of atoms folded into more coarse effective objects).

These approaches provide an enormous simplification and reduction of computational costs, enabling treatment of very large systems, such as those representing the chemistry of solutions and interfaces, materials, and biological systems. Of course, coarse-grained approaches are limited in the scope of what they can describe: mean-field theories assume the dominance of one-electron terms. Yet, by making physically motivated choices, one can construct multilayered models that combine high-level quantum-mechanical treatments of the core of the system with more coarse-grained treatments of the periphery such as in the QM/MM approach or other types of embedding.

### 2.1.2. Complexity Reduction and Opportunities for ML

Within any particular theory depicted in Figure 1 many challenges persist, as in the curse of scaling. Even the lowest-rung quantum theories scale polynomially. For example, in DFT or mean-field calculations, one needs to tackle nonlinear eigenvalue problems that scale as $O(N^2)$ with the system size $N$, even if the Fock matrix build is done with linear scaling. This limits both the length- and time-scale of the simulations. Significant efforts have been devoted to developing reduced-complexity approaches that computationally scale only linearly with the system size. These methods are typically based on so-called “divide-and-conquer” approaches, or numerically thresholded sparse matrix algebra techniques. The divide-and-conquer strategy can also be used to build up an extended system from smaller fragments. The components are treated quantum mechanically, and the results are then coarse-grained into effective Hamiltonians or fragment wave functions that are used to describe extended quantum systems—such coarse-graining can be done in real space, such as in fragment-based approaches, or in Fock space. Here again, physics dictates the most appropriate strategy: divide-and-conquer approaches rely on the locality of electron correlation, whereas coarse-graining in the Fock space is appropriate for strongly correlated systems.

Some of these promises have already been realized. Thanks to recent progress in methods and reduced-complexity algorithms, very large systems can now be treated by atomic or even fully quantum mechanical simulations. Multiscale approaches such as QM/MM and other types of embedding, as well as accelerated molecular dynamics and low-complexity coarse-grained charge equilibration models for molecular dynamics simulations, further extend the scope, both in terms of size of the systems and the time scales of the simulations (see section 2.3).

The authors anticipate that many more will be developed in the future; however, to realize the full potential of DS broadly (and machine learning (ML) specifically) in molecular sciences, more fundamental and interdisciplinary work is needed. Only by combining physically motivated coarse-grained theory with reduced-complexity algorithms and DS/ML will it be possible to significantly extend the roadmap and find transparent shortcuts to enable new, yet unexplored, areas of research.

Challenges abound toward a shared goal of both increasing simulation system time scale, length scale, and complexity while accounting for many-body effects at the appropriate level. Some phenomena inherently require high-level many-body treatments on relatively large scales. Notoriously difficult examples include excited-state processes (as in photobiology and photovoltaic materials), strong correlation (as in magnetic systems), long-range charge and excitation transfer, and long-range electrostatics. Can ML help us to address these challenges? Figure 2 outlines pathways that connect the system specification (the Hamiltonian) with the end result, physical properties and observables. Travel by these roads is enabled by various flavors of coarse-graining. DS/ML promise to provide new, powerful vehicles to make this travel faster (more computationally effective), better (more rigorous and transparent), and accessible to non-experts, as well as to allow us to build additional shortcuts.
ML techniques are already being deployed to carry out a variety of coarse-graining tasks. For example, ML has been successful in the automatic generation of force-fields and accelerating sampling in biosimulations, as well as the design of accurate short-range parametrizations for interatomic potentials that describe the formation energies and the dynamics of molecules and solids (Figure 3). This is achieved by constructing a complex, high-dimensional function described by a deep neural network (NN, see Supporting Information S.7). The inputs are rotationally invariant descriptors derived from the atom types and the molecular structure similar to the embedded atom method, and the output are observables, such as the energy and the interatomic forces. The optimization is performed by minimizing penalty functions that are determined by the difference between the predicted output observables and massive data sets of precalculated properties from high-level electronic structure theory. ML has been also used, with impressive results, to improve parametrization of density functionals, leading to qualitatively better functionals at each rung of DFT.

Artificial neural networks have also been effective in tackling ill-conditioned inverse problems, such as recovering single-particle spectral density in the frequency domain from the (noisy) Fermionic Green’s function in the imaginary time domain. The authors envision that the same ML approaches can also be used to make new or better shortcuts with other theoretical coarse-graining tools. For example, ML may be used to enhance the accuracy of fast low-level quantum chemistry methods such as tight-binding DFT or semiempirical quantum chemistry methods. ML can be also deployed for simplified or automated development of effective spin-Hamiltonians for extended systems. More generally, ML can connect high-level quantum chemistry with effective single-particle treatments or effective Hamiltonians, to deliver true predictive capability required for systems with strong electron correlation where regular DFT fails. Of course, to realize these ideas, many important questions should be addressed. What kind of optimization tools should be used and for what types of parametrization? Which data points from costly high-level wave function calculations will achieve the best parametrization? What is the best way to represent the data from electronic structure calculations? Or, alternatively, can we simply use the functional forms and the many-body physical picture to inform how the parametrization...
of the effective single-particle Hamiltonians should be performed, e.g., DFT+U. How can long-range charge interactions and relaxations be included without a large computational overhead? At the same time, how do we conserve constants of motions, symmetries, and other relevant properties? How is transparency kept through step-by-step coarse-graining in combination with ML algorithms that often use nonphysics deep neural network structures? All these questions present both new opportunities and challenges that will require intellectual investment and development.

ML-enhanced electronic-structure methods can be further combined with linear-scaling techniques and parallel algorithms, to enable studying very large systems with an unprecedented rigor. These new electronic structure tools, in combination with emerging exascale computing capabilities, can be expected to extend the effective reach of electronic structure calculations and AIMD (ab initio molecular dynamics) simulations by many orders of magnitude in time and length scale.

Data science techniques can be leveraged to analyze, learn, and predict the complex relationships that underpin chemical heterogeneity in the Middle Science.

2.1.3. The Challenge of Data Management. These new opportunities will further exacerbate challenges of handling and utilizing the data produced in large-scale simulations. Consider the question of how to best capture, analyze, and understand the vast amount of data generated from large-scale electronic structure calculations and quantum mechanically based molecular dynamics simulations. In the early days of AIMD simulations, when only small systems could be studied during fairly short simulation times, many properties could be analyzed in detail by direct inspection. Even in relatively small AIMD calculations, the wave function is treated as an intermediate object by direct inspection. Even in relatively small AIMD calculations, the wave function is treated as an intermediate object and is processed on-the-fly to extract a predetermined subset of electronic properties. Traditional analyses have been based on the atomic configuration and velocities only; however, more recent developments in graph theory, persistent homology, and other topological analyses are starting to reveal new correlating relationships in complex systems (section 2.2 and the Supporting Information). Such methods become even more important with increased availability of electronic-structure data. The non-local electronic entanglement, spin, and charge densities give a direct and detailed representation of bonding and response properties that are absent in a classical force-field simulation. A major challenge is thus how to take advantage of the additional electronic structure data in large-scale AIMD simulations. How can the evolving electron distribution be used in combination with the phase space of the atomic configurations to extract the best available information? For example, it may be possible to use nonlocal quantum-response properties to analyze and understand events and potentially hidden features in AIMD simulations. Along these lines, can one use the combination of the atomic configuration with the electronic structure to automatically extract chemically significant subgroups? If so, what is the best approach to store the combined information for follow-on studies? Consider the exciting opportunities that emerge if the results of AIMD on solar-cell materials (for example) could be used to develop coarse-grained models to model large-scale energy transport through a macroscopic device.

In this context, physically motivated and mathematically robust data compression techniques become critically important. Luckily, quantum mechanics offers some solutions. The many-body wave functions are expressively nonseparable and entangled, but electrons (or bosons) are indistinguishable. Consequently, to compute actual observables (energy, gradient, dipole moment, etc.), it is sufficient to use reduced quantities, such as density matrices. For example, to compute the expectation value of any one-electron operator, one needs only a one-particle density matrix, an object that depends on six coordinates rather than 3N. Further compression is possible by using diagonal representation (e.g., natural orbitals) and keeping only the values that are above a threshold. These techniques are well-known in quantum chemistry and are used in most practical codes. If these ideas are deployed in large-scale systems, they can be used as instruments to both manage big data and as inputs to ML. Similar concepts exist for excited states and have recently been generalized and extended into nonlinear response domains. The authors foresee that such reduced quantities, such as state and transition density matrices, natural orbitals, spin-correlators, and exciton descriptors, provide an opportunity to effectively store, analyze, and mine the massive information produced by large-scale AIMD simulations. A major challenge to this approach includes the generalization to other properties, standardization, and development of tools for automated analysis.

2.1.4. Anticipated Breakthroughs Enabled by ML. While major leaps in the efficiency of quantum chemistry have been made, these methods are still limited in their reach. Even if quantum-based simulations with hundreds of thousands of atoms over nanoseconds of simulation time could be achieved, this would still not be sufficient for many important applications. Further coarse-graining is needed, for example, where the single electron orbital resolution is replaced by equilibrated atomic net charges and multipole moments. These polarizable charge equilibration models can still describe many of the basic features of electronic structure theory, such as charge transfer, screening, and long-range electrostatic interactions. In combination with ML, these methods can provide a good balance between accuracy and computational efficiency for many problems.

The enhanced physical transparency of large-scale AIMD simulations in combination with analysis tools based on, for example, graph theory, ML, model reduction techniques, and quantum response properties, can potentially provide powerful platforms to understand and detect novel features of complex chemical systems (section 2.2). Known mechanisms that are governed by the electronic structure can be detected and traced, and anomaly detection algorithms could be used to discover yet unknown processes. Once these tools exist (and some already do), electronic structure can be used directly as the input to ML models for structure—property relationships—in the form of 2D or 3D images, graphs, or persistent homology barcodes (see Supporting Information S.3). The large number of snapshots of atomic configurations in combination with the electronic structure in large-scale AIMD simulations can also be used to learn relationships between atomic configurations and electronic structure to enable further enhancements to speeding up electronic structure calculations or even discovering new types of coarse-graining (section 2.3).
2.1.5. Practical Answers versus Fundamental Insights?  
While ML allows one to make shortcuts and go from atoms and electrons straight to practical answers (e.g., drug effectiveness), ML is not a replacement for science based on rigorous fundamental treatments, as in quantum chemistry and atomistic simulations. However, they can work together for mutual benefit—ML can help the community make shortcuts, and we can help develop ML tools by providing information representing the underlying physics.102 Consider that in principle a sufficiently deep neural network, trained on an infinite set of data, can give the exact answer to any specific question based on just system specification. Yet this is just as impractical as the brute-force solution of the exact Schrödinger equation. By using our knowledge of physics we can use semiempirical Hamiltonians to simulate the system, where the Hamiltonians are built using ML.87,88,102−104 This is more transparent, general, and manageable, and is also aligned with well-established traditions in physics,103 emphasizing the value of insight in addition to numerics. We conclude by emphasizing that there is a lot of work to be done from the electronic structure side in order to take full advantage of ML. The community needs better methods (e.g., robust reduced-complexity solvers), better models (e.g., for electron correlation), and better ways to compress information (e.g., using reduced density matrices, coarse-grained charges, or other reduced quantities such as effective Hamiltonians). There is also need for interdisciplinary approaches, where we can leverage graph theory, low complexity solvers, and mathematical topology to develop ways to analyze reduced electronic quantities and create better effective descriptors.

2.2. The Architecture of Space. As the system size and chemical complexity of a system is increased, the manifestations of many-body effects often change scale. Chemical complexity can grow via molecular degrees of freedom, chemical composition, or diversity of intermolecular interactions, all of which increase the breadth of thermally populated particle (electron, nuclear, molecular) configurations. As an example, in the solution process, the increased concentration of solutes can cause collective organization of molecules to occur across length-scale, and thus a broad ensemble of local environments can lead to heterogeneity moving to the micro- and even macroscopic scale.

There are many examples of condensed matter systems where many-body interactions at the scale of molecules, aggregates, or nanoparticles lead to broad ensembles of environments that are distributed in a nonperiodic manner and the system is characterized as being spatially “heterogeneous”. Further, this heterogeneity is essential to the behavior of the system—from the most basic science, to industrial applications and function. Consider the importance of spatial heterogeneity to the fundamental science of microemulsions, emulsions, and foams; nucleation, gelation, and crystallization; and the partitioning of solutes across interfaces within biology, separations, and purification. Importantly, the heterogeneity of spatial organization, caused by a delicate balance of intermolecular forces, reflects the presence of many minima on the configurational energy landscape. If the equilibrium ensemble is sufficiently sampled (section 2.3), then the breadth of the ensemble distribution is a direct reflection of the rugged nature of the energy landscape, and much can be learned about the relative energetics and interconversions between different local and global spatial arrangements, including phase phenomena.

To gain such insight, an ever-expanding domain of research is centered upon using state-of-the-art DS techniques to analyze, learn, and predict the complex relationships that underpin chemical heterogeneity driven by many-body interactions. As an illustration, we point to Figure 4, which shows the rugged and mountainous energy landscape of a complex chemical system. On the left side, a graph representation of ice is comprised of H2O nodes and hydrogen bond edges overlaid upon a chemical reaction coordinate for the transformation from one structural form to another,106 while on the right-hand side the topology of the energy landscape is encoded using sublevels of the energy,107 keeping track of the number of local minima and critical points (including higher-dimensional features). Within the many-body roadmap, we foresee new graph theoretical representations and topological analyses that may be employed in learning methods to empower chemists to not only elucidate chemical heterogeneity but leverage it to control chemical phenomena and develop new processes.

2.2.1. Characterizing Spatial Heterogeneity. Historically, descriptors of local geometry have been labeled “order parameters” and may include geometric criteria (e.g., tetrahedrality), analysis of contacts to a particle (e.g., coordination number), or geometric partitioning methods (e.g., Voronoi analysis).108−119 Within the past decade, there has been significant development in...
atom-centered descriptors that attempt to account for the chemical environment about an atom and or similarity/dissimilarity therein. This has been particularly relevant to atomistic ML models that attempt to find structure–property relationships within the materials science community. These include the Smooth Overlap of Atomic Positions (SOAP),\textsuperscript{120,121} as well as Atom-Centered Symmetry Functions (ACSF),\textsuperscript{122} and the Faber–Chistensen–Huang–Lilienfeld (FCHL) method.\textsuperscript{123} However, the balance of forces that cause spatial and chemical heterogeneity challenges the development of well-defined and transferrable criteria that describe local and nonlocal variations of organization. This is particularly true for noncovalently interacting species, as in molecular assemblies; yet, a breadth of new techniques are changing this research landscape.

Graph theoretical depictions of chemical systems have a long history, particularly in the construction, analysis, and descriptors of molecular graphs, those based upon the covalent bond skeletons.\textsuperscript{124,125} Significant applications include pharmacology and drug development\textsuperscript{126,127} as well as the exploration of chemical synthesis routes.\textsuperscript{128,129} At the same time, network representations of materials (specifically crystals) have flourished with the rapid growth of the crystal design strategies based upon predefined geometric frameworks.\textsuperscript{130−132} Yet, these methods are expanding their impact as a more diverse suite of chemistry data is being represented in a network formalism and as algorithmic improvements allow them to be used in massively parallelized architectures.\textsuperscript{133−135} Broadly, Cartesian coordinate point-cloud data (from experiment or simulation) may be represented as a graph, where individual nodes of the graph can be an atom, molecule, or molecular assembly, and edges constitute a measure of the interactions between them. The connectivity (adjacency) matrix representing the graph (or related matrices) are square and often symmetric, contributing to a breadth of linear algebra tools used to identify graph characteristics (i.e., spectral graph theory).

Although historical emphasis has been placed upon molecular graphs (which map intramolecular bond connectivity), more recent efforts have become increasingly creative; for example, a complete network of all intermolecular interactions may be partitioned into subgraphs that constitute just those nodes that represent a specific type (based on a node attribute like atom type) or edge type (based on the nature of the interparticle interaction as in hydrogen bonding, ion-dipole interactions, etc.). In such intermolecular graphs, significant domain expertise must be employed both in construction (because definitions of intermolecular interactions are nuanced) and analysis, where biases may be introduced through sampling or the choice of criterion employed in their construction.\textsuperscript{136} In general, there remain significant opportunities to enhance the graph representation to include more information that, in turn, will enhance the chemical insight obtained through their analysis. This includes feature vectors $\mu(v)$ attributed to each node $v \in V$, or edge feature vectors $w(e)$ for each edge $e \in E$. Node features within the total graph of the system could include atom/molecule type, or descriptors of electronic structure, while the edge feature could simply be an indicator of whether the edge $e = \{u, v\}$ is connecting similar or dissimilar chemical species $u$ and $v$, or characteristics of the interaction (e.g., hydrogen bonding, ion-dipole, etc.). Within the ensuing discussion, this type of graph is referred to as a heterogeneous graph representation.

Pattern, or subgraph, identification is one means to identify chemically relevant species within a large ensemble present in a graph representation. Locally, algorithms that include edge distribution, spectral graph theory, and clustering have been employed to study chemical heterogeneity, for example, in multicomponent solutions and crystalline systems.\textsuperscript{137,138} In some cases, neural networks have been implemented to learn about local environments of nodes (specifically atoms in crystals) and predict potentially new organizational structures within materials design (Figure 5).\textsuperscript{138} However, the limited information content in traditional chemical graphs highlights an arbitrary nature of many of the choices associated with such algorithms. It is desirable to develop clustering and other algorithms to handle heterogeneous graph inputs. One potential approach is to develop appropriate (distance) metrics for the node feature spaces and convert the input into a discrete metric space $(V, d_v)$. One can then perform standard clustering algorithms over $(V, d_v)$. Defining the metric for these node features $\mu(V)$ is crucial, and prior domain knowledge should be incorporated if possible. We anticipate the possibility of invoking ML for this task, where a graph neural network (GNN, see Supporting Information S.8) learns an embedding of the graph nodes in a better latent space (thus better metric) suitable to preserve the chemical information, in an unsupervised manner.

Appropriate training data, input that is an attributed graph representation of the chemical system, and the ability to compute better feature representations of graph nodes (called node embedding in the literature) all combine to enable an improved feature representation within the graph that represents the chemical system. In other words, the input to a graph neural network (GNN) could be a graph $G = (V, E)$ with node features $\mu: V \rightarrow \mathbb{R}^d$. Using training data, we can train a GNN to output a new embedding $\rho: V \rightarrow \mathbb{R}^d$ for the nodes, so that algorithms such as clustering could take advantage of the new embedding and create more meaningful identification of chemical species. Once trained, this GNN can then be applied to other chemical systems of the same type, potentially at a larger scale. As one moves to larger scales, where molecular-scale clusters are coarse-grained into nodes in a new graph representation, it is possible to still retain molecular-scale information as features of
the larger-scale graph. For example, a node may be associated with the persistence diagram summary of the shape of the cluster corresponding to a node (for an intermolecular graph), with such capabilities being recently reported for graphs in ref 139. The challenge is to develop appropriate (enhanced) GNN architectures that can handle the heterogeneous input graphs more effectively.

An additional option is to incorporate the chemists’ notion that a chemical species is defined by a certain threshold of energetic stability, meaning its associated intermolecular graph persists for a specific time scale within a chemical system. It is valuable to incorporate these concepts into the identification of relevant graph patterns. For example, to compute clusters that are stable w.r.t. dynamic evolution of the system, a simple approach would be to treat the graphs of systems at a different time $t_k$ as multiple observed instances of the same input and compute a “stable clustering” from these instances.\textsuperscript{140,141}

### 2.2.2. Characterizing the Geometry and Properties of Chemical Assemblies

The shape and geometry of the chemical species provide important insights into the structural correlations and energetics of the systems. An illustrative example of this comes from nucleation. Nucleation refers to the birth of a new (stable) phase from a metastable phase. This is often characterized by the formation of clusters that emerge and dissolve rapidly due to the imbalance of the favorable energy of the cluster and unfavorable energy of the interface between the metastable and stable states. This interface is strongly influenced by the shape of the clusters formed. Being able to robustly identify the shapes and their fluctuations is therefore important to understand the nucleation process and the associated energetics. To characterize the “shape” of components in a chemical system, as well as the interaction among them, we again can view the input data as different forms: weighted point clouds (e.g., representing atom positions, or the centers of mass of molecules, etc.), surface/volume models, or graphs (representing chemical systems at different scales).

For graphs, classical summaries include simple graph based statistics (e.g., diameter of the graph, average degrees) or spectral properties. Consider that eigenvalues or eigenvectors of the adjacency matrix (or its related matrices) can be related to specific edge patterns within a graph, as in the use of PageRank as a tool to identify polyhedral arrangement of particles and as a collective variable for chemical transformations.\textsuperscript{93,142,143}

Leveraging recent developments in computational topology, one can use other meaningful features to characterize the graph topology together with metric information, such as the length sequence of the minimal cycle base of a weighted graph.\textsuperscript{144} A particular general framework is via persistent homology, which can map an input (metric or weighted) graph into a persistence diagram feature representation, via the use of the intrinsic Cech filtration over the input graph (equipped with the shortest path metric),\textsuperscript{145} or induced by the clique complexes (for both directed or undirected graphs),\textsuperscript{146–149} or induced by the so-called path homology when the input graph is directed.\textsuperscript{150–152} Once mapped into persistence diagram summaries, one can endow an appropriate distance metric on the (feature) space of persistence diagrams (see Supporting Information S.3) or further map this feature space to a finite/infinite dimensional vector space, so as to perform downstream analysis such as clustering, classification, or property prediction \textit{(vide infra)}. In recent work, such an approach was taken to understand the quantum delocalization of nuclei and the structures of ion aggregates in electrolytes.\textsuperscript{92,206} Within dynamically evolving systems, studying the fluctuations of the persistent homology or distance metrics (i.e., the Fourier transform) can also yield important information about the characteristic time scales of chemical phenomena.\textsuperscript{152}

More rich features can be potentially extracted by combining geometric and topological methods. The elevation function\textsuperscript{153} has been proposed as a way to identify protrusions and cavities in a molecular surface for, e.g., the protein docking problem.\textsuperscript{154} To obtain the protrusions and cavities, one can use Morse theory to segment the input surface into “mountains” and “valleys” based on the elevation function $f: M \rightarrow \mathbb{R}$. Alternatively, one could employ concepts from geometric measure theory, algebraic topology, and optimization. The \textit{flat norm} has been used to define a distance between generalized hypersurfaces within a Euclidean space. In applications to a soft matter surface, protrusions and hierarchically organized structures have been identified by minimization of a generalized area of the soft matter interface.\textsuperscript{155} The surface was flattened toward a horizontal plane with the same $x$, $y$ dimensions, by minimization of the flat norm function, and the volumes were systematically identified as protrusion features. In general, by choosing an appropriate descriptor function defined on the surface, or defined in the space (volume) around the structures of interest, one can again use persistent homology to map it to a persistence diagram feature representation induced by this function.

In the case where the input is a point cloud data, say representing a system of $n$ particles, one can build a weighted simplicial complex representation using the interaction among multiple particles (intuitively, the interaction of $p$ particles gives rise to a $(p-1)$-simplex) and perform persistent homology to filtrations induced by such weighted complex. If the goal is to characterize the interaction between two or more subsystems, then one can potentially only consider those simplicies spanned by particles from multiple systems. Finally, in addition to persistence diagram summaries, it is also interesting to identify geometric representations of features such as significant “holes” and “bubbles” based on either the graph representation or the point cloud data of the Cartesian coordinates of the chemical system, as has been done with protein structures.\textsuperscript{156} Such features again can be captured with the help of the language of topology, where intuitively, the $p$-dimensional homology group of a space roughly describes $p$-dimensional holes. Finding good geometric representations of such features then amounts to computing generating cycles with good properties (e.g., tightest loops hugging certain one-dimensional holes) for such homological features.\textsuperscript{157–160} These tools provide potentially new ways to feature vectorize a chemical system or its individual components. It will also be interesting to combine these features with neural network architectures to further improve the sensitivity and representation power of neural networks, e.g., ref 139.

#### 2.2.3. Heterogeneity Across Scale

Graph patterns of complex chemical systems may also form groups (clusters) at different length-scales, and \textit{a priori}, it is not always clear which scale one should inspect data in an agnostic way. Instead of the \textit{flat clustering} as discussed above, it is also natural to consider the \textit{hierarchical clustering} of input data and inspect the possible grouping at all scales. This was recently demonstrated for heterogeneous chemical solutions using modularity optimization.\textsuperscript{161} It has also been argued by the authors of ref 162 that hierarchical clustering has better properties than flat clustering, which has an intrinsic \textit{impossibility theorem} as shown by Kleinberg.\textsuperscript{163}
The aforementioned persistence homology can potentially be used to characterize an input system across multiple scales simultaneously. As described in the Supporting Information S.3, if persistent homology summarizes the creation and death of topological features along a sequence of filtration (e.g., distance) values as the scale increases, then its resulting persistence diagram summary encodes the lifetime of features across all scales simultaneously. If we could create a filtration that can encode the evolution of the chemical system as the scale increases, then persistent homology can be used to capture and summarize features of hierarchical organization and spatial heterogeneity. Within a chemical system, simplicial complexes could first be used to represent the organization of atoms into molecules. As we increase scale to encompass simplicial complexes of intermolecular organizational structure, larger substructures gradually emerge. If we aim to capture only the intermolecular or longer-range interactions and organization, we may collapse a group of points at the atomistic scale into a supernode at a larger scale, which further gives rise to a simplicial map among their corresponding simplicial complex representations. This could be performed iteratively to capture macroscopic organizational features. Interestingly, this idea shares some similarity with the sparsification of the so-called Rips complex (and their variants) across scales.164–166

2.2.4. Notions of Distances. Chemical insight can be derived from systematically studying the differences in chemical networks or persistent homology, as well as the associated descriptors that identify heterogeneity and broad ensembles of chemical environments. There exists significant literature regarding binary similarity or distance measures that are used for classification purposes.167 Some of which have been employed for molecular graphs.168 Further, the graph representations of materials systems have enabled the development of similarity metrics for querying large structural databases.169 Yet, there remain significant opportunities for developing distance metrics (see Supporting Information S.1) that quantitatively measure differences in the heterogeneous graphs described above. Such distances would ideally be invariant of rigid transformation or internally normalized so that they can be compared across different chemical systems and system sizes.

Distances to compare the precise graph combinatorial structure, e.g., the graph edit distance, are typically computationally hard (even for trees170). One way to sidestep this is by comparing instead a “meaningful” summary of graphs. Simple summaries include diameter of graph and degree sequence/distribution. A popular family of summaries is via spectral structures, such as the spectra of the graph Laplace operator or the diffusion operator associated with input graphs. Recent advancements in topological data analysis provide new ways to compare graphs via persistent homology. The high-level idea is that given a graph \( G \), one can construct an appropriate filtration \( F_t \) on \( G \) and use its resulting persistence diagram as a topological summary for \( G \). Classical distance metrics for persistence diagrams, e.g., the bottleneck distance171 and the \( p \)-th Wasserstein distance,172 enjoy stability properties but lack nice (e.g., inner-product) structure. Recently, a series of methods have been developed to map persistence diagrams to a vector space or Hilbert space.173–175 Regarding graphs of chemical networks, it could be interesting to explore what are meaningful filtrations on them (by incorporating domain knowledge so as to produce useful persistence-based summaries) and how to incorporate multiple such summaries in defining the final distance metric.

2.3. Coarse-Graining Time. The spatial heterogeneity caused by many-body interactions can have a profound impact upon the propagation of its states in time and/or the ability to access or sample those states and their transformations. It is tempting to use the descriptions based upon graphs or persistent homology as coarse-grained variables for evolving the system in time; however, this is nontrivial because the dynamics of any fine-grained variable that has been coarse-grained out of the representation is necessarily inaccessible. A minimum time step is defined accordingly because it must be long enough that the specifics of the fine-grained variables are inconsequential. On the other hand, time scales that are too many orders of magnitude larger than the minimum time step are also inaccessible because of limitations in computing time. The need for matching both space and time as one coarse grains up in scales is illustrated in Figure 6, where the challenge

![Figure 6. Illustration of the space and time coarse-graining that must be captured with scale-consistent equations of motion. At the shortest scale 1, the resolution in time and space is that of atoms (e.g., Å and femtoseconds). At the next scale 2, the resolution is that of multiatom particles (e.g., 10s of Å and 10s of femtoseconds). At the mesoscale 3, the resolution is in even greater (e.g., 100s of Å and 100s of femtoseconds).](https://doi.org/10.1021/acscentsci.1c00685)
2.3.1. Coarse-Grained Equations of Motion. In chemistry, the solution of the scaled equations of motion is particularly important because one needs to control atomic and molecular composition, and the various relevant relaxations/motions of chemical systems that range from biomolecular events to macroscopic solutions. This spans time scales from femtoseconds to minutes or hours. Hence, the challenge of scales in time—here spanning more than 15 orders of magnitude—cannot be solved by brute computing alone in the foreseeable future. Meanwhile, the heterogeneity in time-dependent behavior and response of the system may require us to access time scales smaller than the minimum time step of the coarse-grained system. A second challenge is a so-called inverse problem in which we must determine the statistics—if not the specifics—of the dynamics of the fine-grained variables knowing only the dynamics of the coarse-grained variables. Thus, the first challenge is the construction of rigorous frameworks (equations of motion) and their solutions (specific trajectories or generalized integrators) that evolve a meso- to macroscale system in time. Such a framework and solution needs to have dynamical consistency,179,184 that is, fidelity to the coarse-grained dynamics of underlying system evolved with lower-scale equations of motion.

One approach to developing coarse-grained dynamics is to train artificial intelligence models directly from fine-grained data.185–187 For example, the ability to efficiently generate large-scale molecular dynamics simulations as described in section 2.1 can enable the training of mesoscale and continuum models that are fit to data from these simulations, leading to models that are able to predict phenomena at experimental scales while leveraging the accuracy of information calculated from ab initio methods. Traditionally, this has required significant expertise to hand-craft featurization, dimensionality reduction, discretization, and kinetic parametrization; machine learning methods developed in recent years have enabled increasing automation of this process.

A complementary approach to redeveloping or encoding the equations of motion entirely in an AI machine relies on mapping the potential at a given scale into an AI machine and using it within an integrator. The use of machine learning to accelerate high-accuracy electronic structure determination was discussed in section 2.1. In addition, significant progress has already been achieved in the deep encoding of ab initio potentials into force fields,70,78,91,186,188 and this approach continues to hold promise.

2.3.2. Large Length-Scale Dynamics. Data-driven model reduction of dynamical systems, especially with new techniques leveraging the power of deep learning in combination with more traditional techniques such as projection, have become increasingly of interest. There are a few main directions in this area: (1) model reduction of an existing known high-dimensional system in order to enable increased simulation speed, which may or may not lead to interpretable coarse-grained dynamics, and (2) model reduction of existing systems in order to analyze and interpret the dynamics, which may or may not lead to increased computational speed.

Recently, methods to combine Koopman analysis with deep learning to develop models for coarse-grained dynamics have been developed. VAMPnets enable an end-to-end framework for training a Markov state model from dynamical simulation data. The learned models automatically transform data from coordinate space to Markov states, rather than depending on handcrafted transformations based on physical expertise.189,190 A similar concept for crystal structures was developed combining Koopman analysis and graph convolutional neural networks.191 Variational autoencoders and several other ML-based approaches have also been proposed to automatically coarse-grain molecular dynamics.192–194 An important aspect of these methods is interpretability of the learned dynamics and their ability to satisfy known physical constraints. For example, Markov states learned from deep learning models can be difficult to interpret, unlike reaction networks learned from molecular dynamics simulations.

There are several potential opportunities for further exploration. First, it would be interesting to understand whether it is possible to use machine learning to learn adaptive time steps when solving the equations of motion. Second, more theoretical work is needed on extrapolation in time and uncertainty quantification. When reduced models and coarse-grained models are learned from data, they are trained from data from a certain time range and initial conditions. It is important to try to understand how reliable these reduced models are when we continue simulating in time. For example, learned reduced models must be careful to not be overfit to the time scale on which they are trained on.196

There has recently also been a growing body of work on data-driven identification of nonlinear dynamical systems. Given measurements from a system over time, the idea of data-driven discovery of dynamics is to estimate the unknown governing equations for this system197 at time scales that are large enough compared to the resolution of the data. Sparsity-based approaches in this machine learning task play the role of Occam’s razor: regularizers promote simplicity in the governing equations by promoting sparsity. However, these techniques need to be carefully implemented when there are important components at very different scales, as is typically the case in for example chemical reaction networks.198 Conversely, when the governing equations are known, data-driven methods can also be used to more efficiently solve systems of differential equations by direct prediction of the solution.199 While still nascent, eventually these techniques may have the potential to increase computational speed sufficiently to obviate the need for reduced-order models and coarse-graining.199

2.3.3. Coarse-Graining to Macroscales. In recent years, automatic learning of large chemical reaction networks from atomistic simulation data has been explored195,200,201 enabling these reaction networks to then be simulated using kinetic models, either through stochastic simulation or rate equations. This enables orders of magnitude faster propagation in both space and time.

For example, kinetic Monte Carlo models of chemical reaction networks have been trained from molecular dynamics simulations. The idea is to learn a set of “elementary” chemical reactions occurring in the system and their corresponding rate constants from molecular dynamics data so that the chemistry of the system can be modeled using the chemical master equation.202 This enables the simulations to propagate in time on the time scale of chemical reactions rather than that of
that can enable atomistic simulations to capture more rare completeness, accelerated sampling techniques have been developed as well as corresponding reaction rate parameters. To ensure reactions relevant to the system and time scales being modeled, simulation data is the ability to enumerate a complete set of much as an order of magnitude in time. However, there are without necessarily obeying physically realistic trajectories, dynamic conditions and heterogeneous systems where diffusion effects play an important role.

An important challenge in learning reaction networks from simulation data is the ability to enumerate a complete set of reactions relevant to the system and time scales being modeled, as well as corresponding reaction rate parameters. To ensure completeness, accelerated sampling techniques have been developed that can enable atomistic simulations to capture more rare events. Some techniques are intended to rapidly find reactions without necessarily obeying physically realistic trajectories, such as the nanoreactor. Other techniques use accelerated sampling methods such as DFT-based adaptive kinetic Monte Carlo (aKMC) to build more complete reaction networks on the fly.

Meanwhile, most of these techniques involve a step where rate parameters are calculated using ab initio and transition state theory. However, the development of satisfactory reaction networks that match the results of full atomistic simulations remains a difficult problem due to the complex nature of the reaction chemistry. Recent results that first enumerate a reaction network and then calculate rate parameters have been able to capture important features but not completely match atomistic simulations. Aside from the assumptions inherent to transition state theory, it is also not certain that reaction barriers calculated by looking at isolated reactions still hold when looking at the potential energy surface of the whole system. For high-temperature, high-pressure systems, it has also been shown that, in addition to reactions involving stable states, transient species may play a critical role in reaction dynamics. This cannot be captured by methods that require rates to be calculated from reaction barriers. Second, while theory has been developed around accelerated sampling techniques to provide information about local completeness, for example, in aKMC, it is not yet well-understood how to characterize completeness of the whole network.

There remain many opportunities for data science methods to attempt to address these issues. Transition state approaches can be used to obtain nonrecrossing dividing surfaces, and data science approaches have been used to capture the multidimensional normally hyperbolic invariant manifolds associated with the reaction geometry. Alternatively, one may be able to directly estimate reaction rates from atomistic simulations. Several excellent path sampling techniques such as transition path sampling, transition interface sampling, and forward flux sampling exist to enable this. However, they are computationally expensive and challenging to apply to ab initio simulations. The advent of neural network based potentials is helping overcome this barrier. To address completeness of the reaction network, it may be possible to use data-driven techniques to predict which reactions are likely to exist under certain thermodynamic conditions. For example, one observation from atomistic simulations is that the size of reaction networks does not seem to actually grow combinatorially with the number of observed species. Thus, one can imagine a machine learning algorithm that could identify likely species and reactions for a given system. This would help provide crucial information about the completeness of learned reaction networks.

3. CONCLUSION

The science and theory of the many-body problem creates an empyreal domain for chemists, data scientists, and mathematicians to work together to rapidly accelerate the development of methodologies that rationally traverse space and time. The roadmap presented in this article is a platform of ideas that we hope will inspire new collaborative and interdisciplinary efforts. Mathematicians and data scientists should appreciate how the many-body problem permeates the physical sciences and may serve as a basis for mathematical development (in both pure and applied science). At the same time, domain experts in chemistry, physics, and materials science should recognize the powerful capabilities that data science and mathematics can impart to reveal complex and multidimensional correlations within their data—advancing their ability to formulate new theories and models of physical systems. It is only through collaboration and mutual respect that the essential challenge of describing many-body phenomena will be tackled to reveal the scientific opportunities that lie within the Middle Science.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscentsci.1c00685.

A coda that provides further reading and references for math and data science tools (PDF)

AUTHOR INFORMATION

Corresponding Author

Aurora E. Clark — Department of Chemistry, Washington State University, Pullman, Washington 99163, United States; orcid.org/0000-0001-9381-721X; Email: auclark@wsu.edu

Authors

Henry Adams — Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523, United States; orcid.org/0000-0003-0914-6316

Rigoberto Hernandez — Departments of Chemistry, Chemical and Biomolecular Engineering, and Materials Science and Engineering, Colorado State University, Fort Collins, Colorado 80523, United States; orcid.org/0000-0001-9381-721X

https://doi.org/10.1021/acscentsci.1c00685

ACS Cent. Sci. 2021, 7, 1271–1287
Engineering, Johns Hopkins University, Baltimore, Maryland 21218, United States; orcid.org/0000-0001-8526-7414
Anna I. Krylov — Department of Chemistry, University of Southern California, Los Angeles, California 90089, United States; orcid.org/0000-0001-6788-5016
Anders M. N. Niklasson — Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States; orcid.org/0000-1836-4982
Sapna Sarupria — Department of Chemical and Biomolecular Engineering, Center for Optical Materials Science and Engineering Technologies (COMSET), Clemson University, Clemson, South Carolina 29670, United States; Department of Chemistry, University of Minnesota, Minneapolis, Minnesota 55455, United States; orcid.org/0000-0001-7692-8313
Yusu Wang — Halicioglu Data Science Institute, University of California, San Diego, La Jolla, California 92093, United States; orcid.org/0000-0001-7950-4348
Stefan M. Wild — Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, Illinois 60439, United States; orcid.org/0000-0002-6099-2772
Qian Yang — Computer Science and Engineering Department, University of Connecticut, Storrs, Connecticut 06269-4155, United States; orcid.org/0000-0001-5519-1092

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.1c00685

Funding
A.E.C acknowledges support from the Department of Energy, Basic Energy Sciences (DOE BES) Grant numbers DE-SC00001815 and FWP 30388 and the National Science Foundation (NSF) Grant No. OAC-1934725. H.A. acknowledges support from the National Science Foundation (NSF) Grant No. OAC-1934725. R.H. acknowledges support from the National Science Foundation under Grant No. CHE-2001611, the NSF Center for Sustainable Nanotechnology, A.I.K. acknowledges support from the U.S. Department of Energy, Office of Science (Grant No. DE-SC0019432, “Q4Q: Quantum Computation for Quantum Prediction of Materials and Molecular Properties”) and from the National Science Foundation (Grant No. ACI-1547580 for MolSSI institute). A.M.N.N acknowledges support from the U.S. Department of Energy, Office of Science of Basic Energy Sciences (LANLE8AN) and by the U.S. Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy Contract No. 89233218NCNA000001. S.S. acknowledges support from the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Award Number DE-SC0015448. Y.W. acknowledges support from the National Science Foundation (Grant No. OAC-2039794). S.M.W. acknowledges support from the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division, under Contract Number DE-AC02-06CH11357. Q.Y. acknowledges support from the National Science Foundation (NSF) Grant No. DMR-2102406.

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This article evolved from presentations and discussions at the workshop “At the Tipping Point: A Future of Fused Chemical and Data Science” held virtually September 21–23, 2020. The workshop was sponsored by the Council on Chemical Sciences, Geosciences and Biosciences of the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences. We thank the members of the council for their encouragement and assistance in developing this workshop.

■ REFERENCES

(1) Feynman, R. P. There’s plenty of room at the bottom. Eng. Sci. 1960, 23, 22.
(2) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. Challenges for density functional theory. Chem. Rev. 2012, 112, 289–320.
(3) Kutzelnigg, W. How many-body perturbation theory (MBPT) has changed quantum chemistry. Int. J. Quantum Chem. 2009, 109, 3858–3884.
(4) Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. Dispersion-corrected mean-field electronic structure methods. Chem. Rev. 2016, 116, 5165–5154.
(5) Brito, W. H.; Kotliar, G. Correlation strength and orbital differentiation across the phase diagram of plutonium metal. Phys. Rev. B: Condens. Matter Mater. Phys. 2020, 102, 245111.
(6) Brus, L. Size dimensionality, and strong electron correlation in nanoscience. Acc. Chem. Res. 2014, 47, 2951–2959.
(7) Bennett, E.; Mydosh, J. A. Strongly correlated electron physics: From Kondo and spin glasses to heavy fermions, hidden order and quantum phase transitions. AIP Conf. Proc. 2012, 1485, 3–77.
(8) Basov, D. N.; Averitt, R. D.; Hsieh, D. Towards properties on demand in quantum materials. Nat. Mater. 2017, 16, 1077–1088.
(9) Mazziotti, D. A. Effect of strong electron correlation on the efficiency of photosynthetic light harvesting. J. Chem. Phys. 2012, 137, 074117.
(10) Mazziotti, D. A.; Skochdopole, N. Quantum Information and Computation for Chemistry; John Wiley & Sons, Ltd, 2014; p 355.
(11) Royal Swedish Academy of Sciences, Press Release, 1998: The Royal Swedish Academy of Sciences has awarded the 1998 Nobel Prize in Chemistry to Prof. Walter Kohn and Prof. John A. Pople. The Laureates have each made pioneering contributions in developing methods that can be used for theoretical studies of the properties of molecules and the chemical processes in which they are involved. Citation: “to Walter Kohn for his development of the density-functional theory and to John Pople for his development of computational methods in quantum chemistry.”
(12) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. J. Chem. Phys. 2013, 139, 090901.
(13) Bartlett, R. J. In Theory and Applications of Computational Chemistry; Dykstra, C., Frenking, G., Scuseria, G., Eds.; Elsevier, 2005.
(14) Malrieu, J. P.; Caballol, R.; Calzado, C. J.; de Graaf, C.; Guihery, N. Magnetic interactions in molecules and highly correlated materials: Physical content, analytical derivation, and rigorous extraction of magnetic Hamiltonians. Chem. Rev. 2014, 114, 429–492.
(15) Harrison, W. A. Electronic Structure and the Properties of Solids: The Physics of the Chemical Bond; Dover: New York, 1980.
(16) Foulkes, W. M. C.; Haydock, R. Tight-binding models and density-functional theory. Phys. Rev. B: Condens. Matter Mater. Phys. 1989, 39, 12520.
(17) Porezag, D.; Frauenheim, Th.; Köhler, Th.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. Phys. Rev. B: Condens. Matter Mater. Phys. 1995, 51, 12947–12957.
(18) Eistner, M.; Porezag, D.; Jungnickel, G.; Eilsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent charge density-functional tight-binding method for simulations of complex materials properties. Phys. Rev. B: Condens. Matter Mater. Phys. 1998, 58, 7260.
(19) Finnis, M. W.; Paxton, A. T.; Methfessel, M.; van Schilfgaarde, M. Crystal structures of zirconia from first principles and self-consistent tight binding. Phys. Rev. Lett. 1998, 81, 5149.

(20) Pople, J. A.; Santry, D. P.; Segal, G. A. Approximate self-consistent molecular orbital theory I. Invariant procedures. J. Chem. Phys. 1965, 43, 1219.

(21) Pople, J. A.; Beveridge, D.; Dobosh, P. Approximate self-consistent molecular orbital theory V. Intermediate neglect of differential overlap. J. Chem. Phys. 1967, 47, 2026.

(22) Dewar, M. J. S.; Thiel, W. A semiempirical model for the two-center repulsion integrals in the NDDO approximation. Theoret. Chim. Acta 1977, 46, 89–104.

(23) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. J. Am. Chem. Soc. 1985, 107, 3902–3909.

(24) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. J. Mol. Model. 2007, 13, 1173–1213.

(25) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. J. Mol. Model. 2013, 19, 1–32.

(26) Dral, P. O.; Wu, X.; Thiel, W. Semiempirical quantum-chemical methods with orthogonalization and dispersion corrections. J. Chem. Theory Comput. 2019, 15, 1743–1760.

(27) Bannon, C.; Caldeweyher, E.; Ehert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. Wiley Interdiscip. Rev.: Comput. Mol. Sci. 2021, 11, 1–49.

(28) Malone, W.; Nebgen, B.; White, A.; Zhang, Y.; Song, H.; Bjorgaard, J. A.; Sifain, A. E.; Rodriguez-Hernandez, B.; Freixas, V. M.; Fernandez-Alberti, S.; Roitberg, A. E.; Nelson, T. R.; Tretiak, S. NEXMD software package for nonadiabatic excited state molecular structure calculations. Theory Comput. 2019, 15, 5771–5783.

(29) Royal Swedish Academy of Sciences, Press Release, 1998: The Nobel Prize in Chemistry for 2013 to M. Karplus, Universite de Strasbourg, France and Harvard University, Cambridge, MA, USA, M. Levitt, Stanford University School of Medicine, Stanford, CA, USA, and A. Warshel, University of Southern California, Los Angeles, CA, USA “for the development of multiscale models for complex chemical systems”.

(30) Wesolowski, T. A.; Warshel, A. Frozen density functional approach for ab initio calculations of solvated molecules. J. Phys. Chem. 1993, 97, 8050–8053.

(31) Abrikosov, I. A.; Niklasson, A. M. N.; Simak, S. I.; Johansson, B.; Ruban, A. V.; Skriver, H. L. Order-N Green function technique for local environment effects in alloys. Phys. Rev. Lett. 1996, 76, 4203.

(32) Govind, N.; Wang, Y. A.; da Silva, A. J. R.; Carter, E. A. Accurate ab initio energetics of extended systems via explicit correlation embedded in a density functional environment. Chem. Phys. Lett. 1998, 295, 129–134.

(33) Huang, C.; Pavone, M.; Carter, E. A. Quantum mechanical embedding theory based on a unique embedding potential. J. Chem. Phys. 2011, 134, 154110.

(34) Welborn, M.; Manby, F. R.; Miller, T. F. Even-handed subsystem selection in projection-based embedding. J. Chem. Phys. 2018, 149, 144101.

(35) Golub, G.; van Loan, C. F. Matrix Computations, 3rd ed.; Johns Hopkins University Press: Baltimore, 1996.

(36) Goedecker, S. Linear scaling electronic structure methods. Rev. Mod. Phys. 1999, 71, 1085–1123.

(37) Bowler, D. R.; Miyazaki, T. O(N) methods in electronic structure calculations. Rep. Prog. Phys. 2012, 75, 036503–036546.

(38) Yang, W. Direct calculation of electron-density in density-functional theory. Phys. Rev. Lett. 1991, 66, 1438.

(39) Walker, P. D.; Mezei, P. G. Molecular electron-density LHS approach to molecular building. J. Am. Chem. Soc. 1993, 115, 12423.

(40) Yang, W. T.; Lee, T. S. A density-matrix divide-and-conquer approach for electronic-structure calculations of large molecules. J. Chem. Phys. 1995, 103, 5674–5678.

(41) Kitauro, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: an approximate computational method for large molecules. Chem. Phys. Lett. 1999, 313, 701.

(42) Ozaki, T. O(N) Krylov-subspace method for large-scale ab initio electronic structure calculations. Phys. Rev. B: Condens. Matter Mater. Phys. 2006, 74, 245101.

(43) Nishimoto, Y.; Fedorov, D. G.; Irls, D. Density-functional tight-binding combined with the fragment molecular orbital method. J. Chem. Theory Comput. 2014, 10, 4801.

(44) Niklasson, A. M. N.; Mniszewski, S. M.; Negre, C. F. A.; Cawkwell, M. J.; Swart, P. J.; Mohd-Yusof, J.; Germann, T. C.; Wall, M. E.; Bock, N.; Rubensson, E. H.; Djidjev, H. Graph-based linear scaling electronic structure theory. J. Chem. Phys. 2016, 144, 234101.

(45) McWeeny, R. The density matrix in self-consistent field theory I. Iterative construction of the density matrix. Proc. R. Soc. London Ser. A: Math. 1956, 235, 496.

(46) Palser, A. H. R.; Manolopoulos, D. E. Canonical purification of the density matrix in electronic-structure theory. Phys. Rev. B: Condens. Matter Mater. Phys. 1998, 58, 12704.

(47) Holas, A. Transforms for idempotency purification of density matrices in linear-scaling electronic-structure calculations. Chem. Phys. Lett. 2001, 340, 552–558.

(48) Niklasson, A. M. N. Expansion algorithm for the density matrix. Phys. Rev. B: Condens. Matter Mater. Phys. 2002, 66, 155115.

(49) Niklasson, A. M. N. Implicit purification for temperature-dependent density matrices. Phys. Rev. B: Condens. Matter Mater. Phys. 2003, 68, 233104.

(50) Liang, W. Z.; Saravanhan, C.; Shao, Y.; Baer, R.; Bell, A. T.; Head-Gordon, M. Improved Fermi operator expansion methods for fast electronic structure calculations. J. Chem. Phys. 2003, 119, 4117.

(51) Rubensson, E. H. Nonmonotonic recursive polynomial expansion for linear scaling calculation of the density matrix. J. Chem. Theory Comput. 2011, 7, 1233.

(52) Suryanarayana, P. Optimized purification for density matrix calculations. Chem. Phys. Lett. 2013, 555, 291.

(53) Rubensson, E. H.; Niklasson, A. M. N. Interior eigenvalues from density matrix expansions in quantum mechanical molecular dynamics. SIAM J. Sci. Comput. 2014, 36, B147.

(54) Trullfand, L. A.; Dianzings, R. M.; Bowler, D. R. Communication: Generalized canonical purification for density matrix minimization. J. Chem. Phys. 2016, 144, 091102.

(55) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation methods: A route to accurate calculations on large systems. Chem. Rev. 2012, 112, 632–672.

(56) Mayhall, N. J.; Head-Gordon, M. Computational quantum chemistry single for Heisenberg spin couplings made simple: Just one spin flip required. J. Chem. Phys. 2014, 141, 134111.

(57) Pokhilko, P.; Krylov, A. I. Effective Hamiltonians derived from equation-of-motion coupled-cluster wave-functions: Theory and application to the Hubbard and Heisenberg Hamiltonians. J. Chem. Phys. 2020, 152, 094108.

(58) Bowler, D. R.; Miyazaki, T. Calculations for millions of atoms with density functional theory: Linear scaling shows its potential. J. Phys.: Condens. Mater 2010, 22, 074207.

(59) VandeVondele, J.; Borstnik, U.; Hutter, J. Linear scaling self-consistent field calculations with millions of atoms in the condensed phase. J. Chem. Theory Comput. 2012, 8, 3565.

(60) Cawkwell, M. J.; Niklasson, A. M. N. Energy conserving, linear scaling Born-Oppenheimer molecular dynamics. J. Chem. Phys. 2012, 137, 134105.

(61) Arita, M.; Bowler, D. R.; Miyazaki, T. Stable and efficient linear scaling first-principles molecular dynamics for 10000+ atoms. J. Chem. Theory Comput. 2014, 10, 5419.

(62) Osei-Kuffuor, D.; Fattebert, J. L. Accurate and scalable O(N) algorithm for first-principles molecular-dynamics computations on large parallel computers. Phys. Rev. Lett. 2014, 112, 046401.
Voter, A. F. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* 1997, 78, 3908.

Voter, A. F.; Montalenti, F.; Germann, T. C. Extending the time scale in atomistic simulation of materials. *Annu. Rev. Mater. Res.* 2002, 32, 321.

Perez, D.; Uberuaga, B. P.; Shim, Y.; Amar, J. G.; Voter, A. F. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Ed.; Elsevier, 2009; p 70.

Niklasson, A. M. N. Extended Lagrangian Born-Oppenheimer molecular dynamics for orbital-free density functional theory and polarizable charge equilibration models. *J. Chem. Phys.* 2021, 154, 054101.

Wang, J.; Olsson, S.; Wehmeyer, C.; Perez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.* 2019, 5, 755–767.

Noe, F.; Tkatchenko, A.; Muller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* 2020, 71, 361–390.

Husic, B. E.; Charron, N. E.; Lemm, D.; Wang, J.; Perez, A.; Krämer, A.; Chen, Y.; Olsson, S.; de Fabritiis, G.; Noé, F.; Clementi, C. Coarse graining molecular dynamics with graph neural networks Technical Report. *ArXiv* 2020, arXiv:2007.11412.

Behler, J.; Parrinello, M. Michele Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* 2007, 98, 146401.

Bartok, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* 2010, 104, 136403.

Rupp, M.; Tkatchenko, A.; Muller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* 2012, 108, 058301.

Thompson, A. P.; Swiler, L. P.; Trotz, C. R.; Foiles, S. M.; Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* 2015, 285, 316–330.

Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Intertatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B: Condens. Matter Mater. Phys.* 2015, 92, 045131.

Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* 2016, 145, 170901.

Schutt, K. T.; Arbabzadah, F.; Chmiela, S.; Muller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* 2017, 8, 13890.

Han, J.; Zhang, L.; Car, R.; Weinan, E. Deep potential: A general representation of a many-body potential energy surface. *Commun. Comput. Phys.* 2018, 23, 629–639.

Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N. S.; Devereux, C.; Baros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* 2019, 10, 2903.

Noe, F.; Tkatchenko, A.; Muller, K.-R.; Clementi, C. Cecilia Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* 2020, 71, 361–390.

Daw, M. S.; Foiles, S. M.; Baskes, M. I. The embedded-atom method: a review of theory and applications. *Mater. Sci. Rep.* 1993, 9, 251–310.

Mardirossian, N.; Head-Gordon, M. oB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* 2014, 16, 9904–9924.

Mardirossian, N.; Head-Gordon, M. Survival of the most transferable at the top of Jacob’s ladder: Defining and testing the oB97M2 double hybrid density functional. *J. Chem. Phys.* 2018, 148, 241736.

Dick, S.; Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. *Nat. Commun.* 2020, 11, 3509.

Yoon, H.; Sim, J.-H.; Han, M. J. Analytic continuation via domain knowledge free machine learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* 2018, 98, 245101.

Fournier, R.; Wang, L.; Yazyev, O. V.; Wu, Q. Artificial neural network approach to the analytic continuation problem. *Phys. Rev. Lett.* 2020, 124, 056401.

Xie, X.; Bao, F.; Maier, T.; Webster, C. Analytic continuation of noisy data using Adams Bashforth ResNet Technical Report 1905. *ArXiv* 2019, 10430.

Ramakrishnan, R.; Drit, P. O.; Rupp, M.; von Lilienfeld, O. A. Anatole Big data meets quantum chemistry approximations: The delta-machine learning approach. *J. Chem. Theory Comput.* 2015, 11, 2087–2096.

Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F. Orbitnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* 2020, 153, 124111.

Anisimov, V. I.; Zaamen, J.; Andersen, O. K. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Phys. Rev. B: Condens. Matter Mater. Phys.* 1991, 44, 943.

Kulik, H. Treating electron over-delocalization with the DFT +U method. *J. Chem. Phys.* 2015, 142, 240901.

Behler, J. Four generations of high-dimensional neural network potentials *Chem. Rev. 2021, DOI: 10.1021/acs.chemrev.0c00868*.

Hu, Y.; Ounkham, P.; Marsalek, O.; Markland, T. E.; Krishmoorthy, B.; Clark, A. E. Persistent homology metrics reveal quantum fluctuations and reactive atoms in path integral dynamics. *Front. Chem.* 2021, 9, 57.

Zhou, T.; Martinez-Baez, E.; Schenter, G.; Clark, A. E. PageRank as a collective variable to study complex chemical transformations and their energy landscapes. *J. Chem. Phys.* 2019, 150, 134102.

Chen, S. H.; Rouch, J.; Sciortino, F.; Tartaglia, P. Static and dynamic properties of water-in-oil microemulsions near the critical and percolation points. *J. Phys.: Condens. Matter* 1994, 6, 10855–10883.

Mallamace, F.; Beneduci, R.; Gambaddeo, P.; Lombardo, D.; Chen, S. H. Glass and percolation transitions in dense attractive micellar system. *Phys. A* 2001, 302, 202–219.

Xia, K. Elina Persistent homology analysis of ion aggregations and hydrogen-bonding networks. *Phys. Chem. Chem. Phys.* 2018, 20, 13448–13460.

Luzanov, A. V.; Sukhorukov, A. A.; Umanskii, V. E. Application of transition density matrix for analysis of excited states. *Theor. Exp. Chem.* 1976, 10, 354–361. *Russian original: Teor. Eksp. Khim., 197410, 456.*

Luzanov, A. V.; Zhikol, O. A. In *Practical Aspects of Computational Chemistry I: An Overview of the Last Two Decades and Current Trends*; Leszczynski, J., Shukla, M. K., Eds.; Springer, 2012; p 415.

Plasser, F.; Wormit, M.; Dreyer, A. New tools for the systematic analysis and visualization of electronic excitations. I. Formalism. *J. Chem. Phys.* 2014, 141, 024106–13.

Krylov, A. I. From orbitals to observables and back. *J. Chem. Phys.* 2020, 153, 080901.

Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* 2021, 12, 398.

Zubatyiuk, T.; Isayev, O. Development of multimodal machine learning potentials: Toward a physics-aware artificial intelligence. *Acc. Chem. Res.* 2021, 54, 1575–1585.

Li, H.; Collins, C.; Tanha, M.; Gordon, J. G.; Yaron, D. J. A density functional tight binding layer for deep learning of chemical hamiltonians. *J. Chem. Theory Comput.* 2018, 14, 5764–5776.

Li, L.; Hoyer, S.; Pederson, R.; Sun, R.; Cubuk, E. D.; Riley, P.; Burke, K. Kieron Kohn-Sham equations as regularizer: Building
prior knowledge into machine-learned physics. *Phys. Rev. Lett.* 2021, 126, 036401.

(105) Pople, J. A.; Beveridge, D. L. *Approximate Molecular Orbital Theory*; McGraw-Hill Series in Advanced Chemistry, 1970.

(106) Clark, A. E. *In Annual Reports in Computational Chemistry; Dixon, D. A., Ed.; Elsevier, 2015; p 313.

(107) Mirth, J.; Zhai, Y.; Bush, J.; Alvarado, E. G.; Jordan, H.; Heim, M.; Krishnamoorthy, B.; Pilaua, M.; Clark, A.; Z; Y; Adams, H. Henry Representations of energy landscapes by sublevelset persistent homology: An example with n-alkanes. *J. Chem. Phys.* 2021, 154, 114114.

(108) Galamba, N. Water's structure around hydrophobic solutes and the iceberg model. *J. Phys. Chem. B* 2013, 117, 2153–2159.

(109) Chau, P.-L.; Hardwick, A. J. A new order parameter for tetrahedral configurations. *Mol. Phys.* 1998, 93, 511–518.

(110) Giovambattista, N.; Debenedetti, P. G.; Sciortino, F.; Stanley, H. E. Eugene Structural order in glassy water. *J. Phys. Chem. A* 2008, 112, 13066–13073.

(111) Paolantoni, L.; Lago, N. F.; Alberti, M.; Lagana, A. Tetrahedral ordering in water: Raman profiles and their temperature dependence. *J. Phys. Chem. A* 2009, 113, 15100–15105.

(112) Nayar, D.; Chakravarty, C. Charusita Water and water-like liquids: Relationships between structure, entropy and mobility. *Phys. Chem. Chem. Phys.* 2013, 15, 14162–14177.

(113) Sedlmeyer, F.; Horinek, D.; Netz, R. R. Spatial correlations of density and structural fluctuations in liquid water: A comparative simulation study. *J. Am. Chem. Soc.* 2011, 133, 1391–1398.

(115) Bernal, J. D. A geometrical approach to the structure of liquids. *Nature* 1959, 183, 141–147.

(116) Medvedev, N. N.; Voloshin, V. P.; Naberkhun, Yu I Structure of simple liquids as a problem of percolation in the Voronoi grid. *J. Struct. Chem.* 1989, 30, 253–260.

(117) Naberkhun, Yu I; Voloshin, V. P.; Medvedev, N. N. Geometrical analysis of the structure of simple liquids: percolation approach. *Mol. Phys.* 1991, 73, 917–936.

(118) Idriiss, A.; Damay, P.; Yukichi, K.; Jedlovsky, P. Pal Self-association of urea in aqueous solutions: A Voronoi polyhedron analysis study. *J. Chem. Phys.* 2008, 129, 164512.

(119) Duboue-Dijon, E.; Laage, D. Characterization of the local structure in liquid water by various order parameters. *J. Phys. Chem. B* 2015, 119, 8406–8418.

(120) Bartok, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Matter. Phys.* 2013, 87, 184115.

(121) De, S.; Bartok, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* 2016, 18, 13754–13769.

(122) Behler. Jörg Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* 2011, 134, 074106.

(123) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* 2018, 148, 241717.

(124) Trinajstič, N. *Chemical Graph Theory, 2nd ed.*; CRC Press, 1992.

(125) Balaban, A. T. Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* 1985, 25, 334–343.

(126) Amigó, J. M.; Gálvez, J.; Villar, V. M. A review on molecular topology: applying graph theory to drug discovery and design. *Naturwissenschaften* 2009, 96, 749–761.

(127) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in ai-driven drug discovery: a review and practical guide. *Cheminfo.* 2020, 12, 56.
between structure and function. Henry Cliques of neurons bound into cavities provide a missing link as a network invariant.

The topology of the directed clique complex as a network invariant. Rudolph, J.

153. Aggarwal, P. K.; Edelsbrunner, H.; Harer, J.; Wang, Y. Yusu Extreme elevation on a 2-manifold. Discrete Comput. Geom. 2006, 36, 553−572. (154) Wang, Y.; Aggarwal, P. K.; Brown, P.; Edelsbrunner, H.; Rudolph, J. Pac. Symp. Biocomput. 2005, 10, 64−76. DOI: 10.1142/9789812702456_0007

155. Alvarado, E.; Liu, Z.; Servis, M. J.; Krishnamoorthy, B.; Clark, A. E. A geometric measure theory approach to identify complex structural features on soft matter surfaces. J. Chem. Theory Comput. 2020, 16, 4579−4587.

156. Xia, K.; Wei, G. W.P. Persistent homology analysis of protein structure, flexibility, and folding. Int. J. Numer Method Biomed Eng. 2014, 30, 814−844.

157. Chen, C.; Freedman, D. Hardness results for homology localization. Discrete Comput. Geometry 2011, 45 (3), 425−448.

158. Chen, C.; Freedman, D. Measuring and computing natural generators for homology groups. Comput. Geom. 2010, 43 (2), 169−181.

159. Erickson, J.; Whittlesey, K. In SODA ’05: Proc. 16th Ann. ACM-SIAM Sympos.; Discrete Algorithms, 2005.

160. Dey, T. K.; Hou, T.; Mandal, Computing minimal persistent cycles: Polynomial and hard cases CoRR 2019, abs/1907.04889.

161. Servis, M. J.; Clark, A. E. Cluster identification using modularity optimization to uncover chemical heterogeneity in complex solutions. J. Phys. Chem. A 2021, 125, 3986.

162. Carlsson, G.; Mémoli, F. Characterization, stability and convergence of hierarchical clustering methods. J. Mach. Learn. Res. 2010, 11, 1425−1470.

163. Kleinberg, J. M. Adv. Neural Information Processing Syst. 2002, 15, 463−470.

164. Sheehy, D. R. Linear-size approximations to the Vietoris-Rips filtration. Discrete Comput. Geom. 2013, 49, 778−796.

165. Dey, T. K.; Shi, D.; Wang, Y. SimBia: An efficient tool for approximating Rips-filtration persistence via simplicial batch collapse. ACM J. Exp. Algorithmics 2019, 24, 1−16.

166. Dey, T. K.; Fan, F.; Wang, Yusu In Proc. 29th. Annu. Symp. Comput. Geom.; SoCG, 2013.

167. Choi, S. H. C.; Cha, S.-H.; Tappert, C. A survey of binary similarity and distance measures. Systemics, Cybernetics, Informatics 2010, 8, 4348.

168. Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity — a review. QSR Comb. Sci. 2003, 22, 1006−1026.

169. Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S. Materials cartography: Representing and mining materials space using structural and electronic fingerprints. Chem. Mater. 2015, 27, 735−743.

170. Bille, P. A survey on tree edit distance and related problems. Theor. Comput. Sci. 2005, 337, 217−239.

171. Cohen-Steiner, D.; Edelsbrunner, H.; Harer, J. Stability of persistence diagrams. Discrete Comput. Geom. 2007, 37, 103−120.

172. Milevsky, Y.; Mukherjee, S.; Harer, J. Probability measures on the space of persistence diagrams. Inverse Problems 2011, 27, 124007.

173. Bubenik, P. Statistical topological data analysis using persistence landscapes. J. Mach. Learn. Res. 2015, 16, 77−102.

174. Reininghaus, J.; Huber, S.; Bauer, U.; Kwitt, R. In Computer Vision and Pattern Recognition, 2015.

175. Adams, H.; Emerson, T.; Kirby, M.; Neville, R.; Peterson, C.; Shipman, P.; Chepushanova, S.; Hanson, E.; Motta, F.; Ziegelmeier, L. Persistence images: A stable vector representation of persistent homology. J. Mach. Learn. Res. 2017, 18, 1−35.

176. Kusano, G.; Fukumizu, K.; Hiraoka, Y. Kernel method for persistence diagrams via kernel embedding and weight factor. J. Mach. Learn. Res. 2018, 18, 1−41.

177. Carrière, M.; Cuturi, M.; Oudot, S. Y. Sliced Wasserstein kernel for persistence diagrams. Proceedings of the 34th International Conference on Machine Learning, PMLR 2017, 70, 664−673.

178. Le, T.; Yamada, M. In Advances in Neural Information Processing Systems; NeurIPS, 2018.

179. Hagy, M. C.; Hernandez, R. Dynamical simulation of dipolar Janus colloids: Dynamical properties. J. Chem. Phys. 2013, 138, 184903.

180. Li, Z.; Bian, X.; Li, X.; Karniadakis, G. E. Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism. J. Chem. Phys. 2015, 143, 243128.

181. Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. J. Chem. Phys. 2008, 129, 144108.

182. Davtyan, A.; Dama, J. F.; Voth, G. A.; Andersen, H. C. Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence. J. Chem. Phys. 2015, 142, 154104.

183. Minarik, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: Coarse grained model for biomolecular simulations. J. Phys. Chem. B 2007, 111, 7812−7824.

184. Cui, Q.; Hernandez, R.; Mason, S. E.; Frauenheim, T.; Pedersen, J. A.; Geiger, F. Mini-review. Sustainable nanotechnology: Opportunities and challenges for theoretical/computational studies. J. Phys. Chem. B 2016, 120, 7297−7306.

185. Noe, F.; Tkatchenko, A.; Muller, K.-R.; Clementi, C. Machine learning for molecular simulation. Annu. Rev. Phys. Chem. 2020, 71, 361−390.

186. Jia, W.; Wang, H.; Chen, M.; Lu, D.; Lin, L.; Car, R.; Weinan, E.; Zhang. L. In SC 20: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE Press, 2020, 5, 1−14.

187. Pant, S.; Smith, Z.; Wang, Y.; Tajkhorshid, E.; Tiwary, P. Pratyush Confronting pitfalls of AI-augmented molecular dynamics using statistical physics. J. Chem. Phys. 2020, 153, 234118.

188. Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A.; Ong, S. P. A performance and cost assessment of machine learning interatomic potentials. J. Phys. Chem. A 2020, 124, 731−745.

189. Chen, W.; Sidky, H.; Ferguson, A. L. Nonlinear discovery of slow molecular modes using state-free reversible vannpets. J. Chem. Phys. 2019, 150, 214114.

190. Mardt, A.; Pasquali, L.; Wu, H.; Noe, F. VAMPnets for deep learning of molecular kinetics. Nat. Commun. 2018, 9, 1−11.

191. Xie, T.; France-Lanord, A.; Wang, Y.; Shao-Horn, Y.; Grossman, J. C. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. Nat. Commun. 2019, 10, 2667.

192. Wang, W.; Gomez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics npj. Computational Materials 2019, 5, 1−9.

193. Jackson, N. E.; Webb, M. A.; de Pablo, J. Recent advances in machine learning towards multiscale soft materials design. Carr. Opin. Chem. Eng. 2019, 23, 106−114.

194. Chakraborty, M.; Xu, C.; White, A. D. Andrew D Encoding and selecting coarse-grain mapping operators with hierarchical graphs. J. Chem. Phys. 2018, 149, 134106.

195. Chen, E.; Yang, Q.; Dufour-Decieux, V.; Sing-Long, C. A.; Freitas, R.; Reed, E. J. Transferable kinetic Monte Carlo models with thousands of reactions learned from molecular dynamics simulations. J. Phys. Chem. A 2019, 123, 1874.
(196) Yang, Q.; Sing-Long, C. A.; Reed, E. J. Rapid data-driven model reduction of nonlinear dynamical systems including chemical reaction networks using l1-regularization. *Chaos* 2020, 30, 053122.

(197) Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U. S. A.* 2016, 113, 3932–3937.

(198) Raissi, M.; Perdikaris, P.; Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 2019, 378, 686–707.

(199) Wiewel, S.; Becher, M.; Thuerey, N. Latent space physics: Towards learning the temporal evolution of fluid flow. *Computer Graphics Forum* 2019, 38, 71–82.

(200) Ford, J.; Seritan, S.; Zhu, X.; Sakano, M. N.; Islam, M. M.; Strachan, A.; Martinez, T. J. Nitromethane decomposition via automated reaction discovery and an ab initio corrected kinetic model. *J. Phys. Chem. A* 2021, 125, 1447–1460.

(201) Dontgen, M.; Przybylski-Freund, M.-D.; Kroger, L. C.; Kopp, W. A.; Ismail, A. E.; Leonhard, K. Automated discovery of reaction pathways, rate constants, and transition states using reactive molecular dynamics simulations. *J. Chem. Theory Comput.* 2015, 11, 2517–2524.

(202) Jung, H.; Covino, R.; Hummer, G. Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations *arXiv* 2019, 1901.04595.

(203) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martinez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* 2014, 6, 1044–1048.

(204) Xu, L.; Henkelman, G. Adaptive kinetic Monte Carlo for first-principles accelerated dynamics. *J. Chem. Phys.* 2008, 129, 114104.

(205) Xu, L.; Zhu, F. XX. A new way to develop reaction network automatically via DFT-based adaptive Kinetic Monte Carlo. *Chem. Eng. Sci.* 2020, 224, 115746.

(206) Hernandez, R.; Uzer, T.; Bartsch, T. Transition state theory in liquids beyond planar dividing surfaces. *Chem. Phys.* 2010, 370, 270–276.

(207) Schraft, P.; Junginger, A.; Feldmaier, M.; Bardakcioglu, R.; Main, J.; Wunner, G.; Hernandez, R. Neural network approach to time-dependent dividing surfaces in classical reaction dynamics. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* 2018, 97, 042309.

(208) Tschöpe, M.; Feldmaier, M.; Main, J.; Hernandez, R. Neural network approach for the dynamics on the normally hyperbolic invariant manifold of periodically driven systems. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* 2020, 101, 022219.

(209) Moqadam, M.; Lervik, A.; Riccardi, E.; Venkatraman, V.; Alsberg, B. K.; van Erp, T. S. Local initiation conditions for water autoionization. *Proc. Natl. Acad. Sci. U. S. A.* 2018, 115, E4569–E4576.

(210) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. Transition path sampling. *Adv. Chem. Phys.* 2003, 123, 1–78.

(211) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. Simulating rare events in equilibrium or nonequilibrium stochastic systems. *J. Chem. Phys.* 2006, 124, 024102.

(212) Van Erp, T. S.; Bolhuis, P. G. Elaborating transition interface sampling methods. *J. Comput. Phys.* 2005, 205, 157–181.

(213) Bolhuis, P. G. Rare events via multiple reaction channels sampled by path replica exchange. *J. Chem. Phys.* 2008, 129, 114108.

(214) DeFever, R. S.; Sarupria, S. Contour forward flux sampling: Sampling rare events along multiple collective variables. *J. Chem. Phys.* 2019, 150, 024103.

(215) Galib, M.; Limmer, D. T. Reactive uptake of n2o5 by atmospheric aerosol is dominated by interfacial processes. *Science* 2021, 371, 921–925.

(216) Dufour-Decieux, V.; Freitas, R.; Reed, E. J. Atomic-level features for statistical learning of the chemical kinetics from molecular dynamics simulations. *J. Phys. Chem. A* 2021, 125, 4233.