Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems

John A. Keith,* Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller,* and Alexandre Tkatchenko*
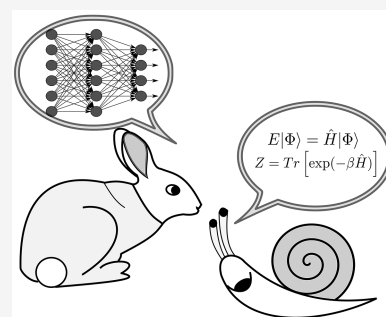
Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Machine learning models are poised to make a transformative impact on chemical sciences by dramatically accelerating computational algorithms and amplifying insights available from computational chemistry methods. However, achieving this requires a confluence and coaction of expertise in computer science and physical sciences. This Review is written for new and experienced researchers working at the intersection of both fields. We first provide concise tutorials of computational chemistry and machine learning methods, showing how insights involving both can be achieved. We follow with a critical review of noteworthy applications that demonstrate how computational chemistry and machine learning can be used together to provide insightful (and useful) predictions in molecular and materials modeling, retrosyntheses, catalysis, and drug design.

$$E|\Phi\rangle = \hat{H}|\Phi\rangle$$
$$Z = Tr\left[\exp(-\beta\hat{H})\right]$$

CONTENTS

1. INTRODUCTION

1.1. Background

A lasting challenge in applied physical and chemical sciences has been to answer the question: how can one *identify and make* chemical compounds or materials that have optimal properties for a given purpose? A substantial part of research in physics, chemistry, and materials science concerns the discovery and characterization of novel compounds that can benefit society, but most advances still are generally attributed to trial-and-error experimentation, and this requires significant time and cost. Current global challenges create greater urgency for faster, better, and less expensive research and development efforts. Computational chemistry (CompChem) methods have significantly improved over time, and they promise paradigm shifts in how compounds are fundamentally understood and designed for specific applications.

Machine learning (ML) methods have in the past decades witnessed an unprecedented technological evolution enabling a plethora of applications, some of which have become daily companions in our lives.[1−3] Applications of ML include technological fields, such as web search, translation, natural language processing, self-driving vehicles, control architectures, and in the sciences, for example, medical diagnostics,[4−8] particle physics,[9] nano sciences,[10] bioinformatics,[11,12] brain-computer interfaces,[13] social media analysis,[14] robotics,[15,16] and team, social, or board games.[17−19] These methods have also become popular for accelerating the discovery and design of new materials, chemicals, and chemical processes.[20] At the same time, we have witnessed hype, criticism, and misunderstanding about how ML tools are to be used in chemical research. From this, we see a need for researchers working at the intersection of CompChem+ML to more critically recognize the true strengths and weaknesses of each component in any given study. Specifically, we wanted to review why and how CompChem+ML can provide useful insights into the study of molecules and materials.

While developing this Review, we polled the scientific community with an anonymous online survey that asked for questions and concerns regarding the use of ML models with chemistry applications. Respondents raised excellent points including:

1. ML methods are becoming less understood while they are also more regularly used as black box tools.

2. Many publications show inadequate technical expertise in ML (e.g., inappropriate splitting of training, testing, and validation sets).

3. It can be difficult to compare different ML methods and know which is the best for a particular application or whether ML should even be used at all.

4. Data quality and context are often missing from ML modeling, and data sets need to be made freely available and clearly explained.

Additionally, when asked about the most exciting active and emerging areas of ML in the next five years, respondents mentioned a wide range of topics from catalysis discovery, drug and peptide design, "above the arrow" reaction predictions, and generative models that promise to fundamentally transform chemical discovery. When asked about challenges that ML will not surmount in the next five years, respondents mentioned modeling complex photochemical and electrochemical environments, discovering exact exchange-correlation functionals, and completely autonomous reaction discovery. This Review will give our perspective on many of these topics.

As context for this Review, Figure 1 shows a heatmap depicting the frequency of ML keywords found in scientific articles that also have keywords associated with different American Chemical Society (ACS) technical divisions. Preparing this figure required several steps. First, lists of ML keywords were chosen. Second, lists of keywords were created by perusing ACS division symposia titles from over the past five years. Third, Python scripts used Scopus Application Programming Interfaces (APIs) to identify the number of scientific publications that matched sets of ML and division symposia keywords. Figure 1 elucidates several interesting points. First, the most popular ML approaches across all divisions are clearly neural networks, followed by genetic algorithms and support vector machines/kernel methods. Second, divisions such as physical (PHYS), analytical (ANYL), and environmental (ENVR) are already using diverse sets of ML approaches, while divisions such as inorganic (INOR), nuclear (NUCL), and carbohydrate (CARB) are primarily employing more distinct subsets of approaches, while other divisions, such as educational (CHED), history (HIST), law (CHAL), and business-oriented divisions (BMGT and SCHB), that is, divisions that produce much fewer scholarly journal articles, are not linking to publications that mention ML. Third, ML has had more prevalence across practically all divisions over time. For further insight, Table 1 lists the top four keywords obtained from recent ACS symposium titles, as well as their respective contribution percentage reflected in Figure 1. There, one sees that a handful of keywords can significantly overshadow matches in some of the bins, for exampled, "electro", "sensor", "protein", and "plastic". With any ML application, there will be a risk of imperfect data or user bias, but this is a useful launch point to appreciate how and where ML is being used in chemical sciences. A key takeaway is that we are witnessing an unprecedented crescendo in interest in ML over the last ten years (e.g., Figure 1c) thanks to improved understanding of the intersectionality of traditional science and engineering disciplines with rapidly evolving disciplines such as CompChem and data science.

1.2. Motivation for This Review

The survey results and literature analysis above showed an opportunity for a tutorial reference to help readers address future research challenges that will require joint applications of
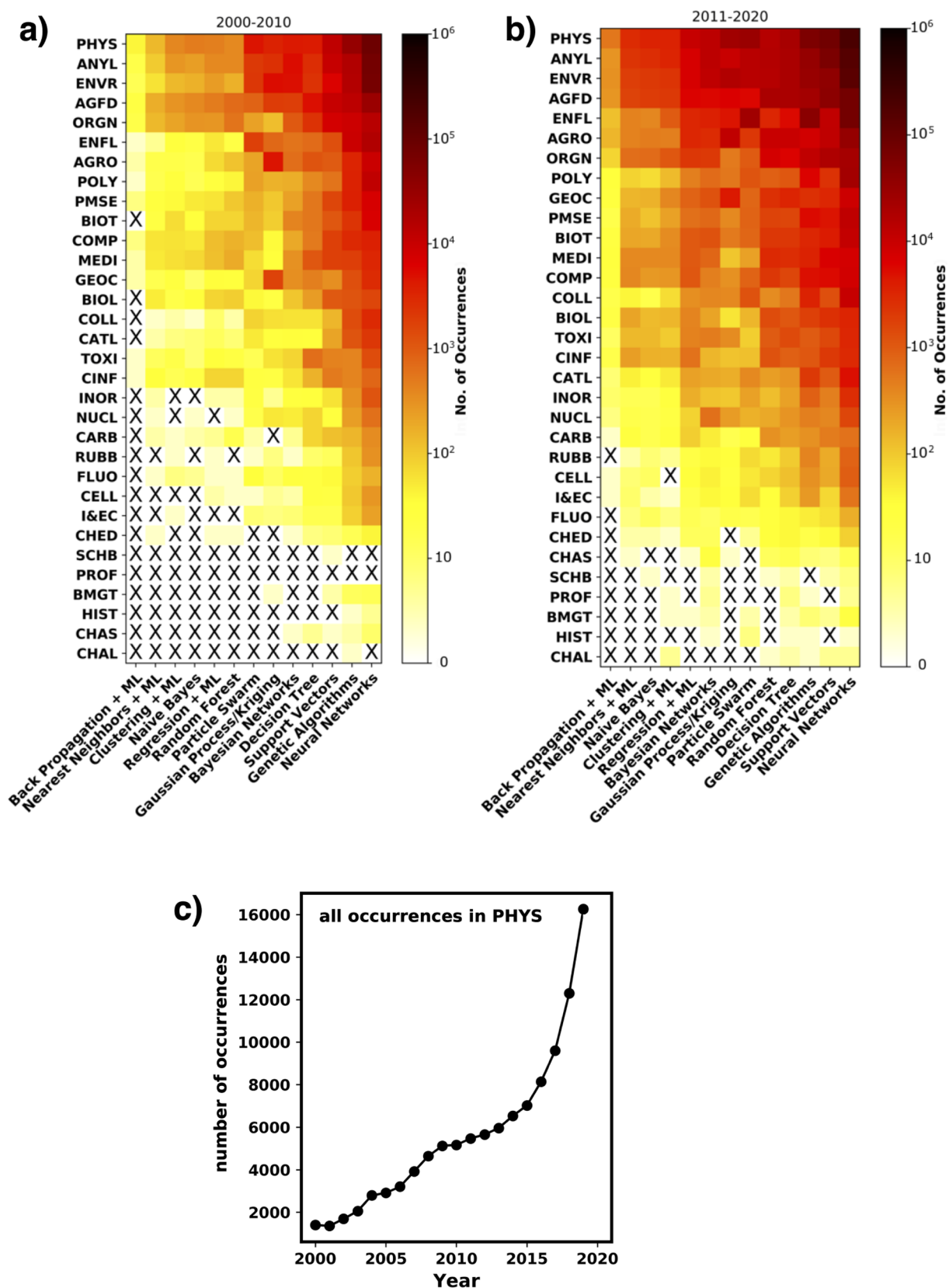
Figure 1. Heatmaps illustrating the extent that ML terms appear in scientific papers aligned by American Chemical Society (ACS) technical divisions from 2000 to 2010 (a) and from 2010−present (b). (c) Line graph showing the number of occurrences of any ML term being found in papers attributed to the ACS PHYS division, from 2000−present. Figures were made by Charles D. Griego. Python scripts used to generate these figures and corresponding Table 1 are freely available with a creative commons attribution license. Readers are welcome to use, adapt, and share these scripts with appropriate attribution: https://github.com/keithgroup/scopus_searching_ML_in_chem_literature.).

Table 1. List of Top Ranked Keywords (Per ACS Division) with Corresponding Percentage of Matches for Any ML Term

division	rank 1	rank 2	rank 3	rank 4
PHYS	electro* (56.3%)	spectroscopy (9.7%)	ion* (6.0%)	nano* (5.5%)
ANYL	sensor* (55.4%)	spectroscopy (13.1%)	characterization* (11.6%)	spectrometry (4.3%)
ENVR	*sensor* (60.9%)	soil* (14.5%)	water quality (4.2%)	environmental monitor* (3.0%)
AGFD	protein* (31.9%)	agricultur* (18.2%)	food (10.8%)	fruit* (5.7%)
ENFL	fuel* (19.2%)	petroleum (11.6%)	energy efficiency (11.1%)	batter* (10.7%)
AGRO	soil (43.3%)	crop* (25.4%)	groundwater (11.5%)	developing countr* (4.4%)
ORGN	protein* (64.6%)	amino acid* (19.7%)	peptide* (8.4%)	aromatic* (3.2%)
POLY	plastic* (51.1%)	polymer* (37.7%)	polymeriz* (5.0%)	polymeric (2.8%)
PMSE	*polymer* (50.4%)	*peptide* (30.4%)	thin film* (8.9%)	tissue engineering (4.3%)
BIOT	biochemi* (37.9%)	biophysic* (18.3%)	systems biology (10.3%)	biotechnology (9.9%)
GEOC	groundwater (33.4%)	mining (31.6%)	*geochem* (12.4%)	anthropogenic (10.9%)
MEDI	protein interaction* (25.7%)	drug discovery (19.1%)	drug design (19.1%)	antibiotic* (11.3%)
COMP	drug discovery (18.7%)	drug design (18.6%)	molecular model* (14.3%)	protein database* (13.1%)
COLL	nanoparticle* (21.2%)	adsorption (19.6%)	thin film* (14.9%)	tribolog* (9.6%)
BIOL	drug discovery (41.9%)	protein folding (19.3%)	biosynthesis (12.2%)	cytochrome* (12.2%)
TOXI	toxi* (99.2%)	chemical exposure* (0.6%)	antibody drug conjugate* (0.1%)	
CATL	cataly* (64.0%)	metal oxides (20.2%)	photocataly* (5.3%)	surface chemistry (2.8%)
CINF	drug discovery (51.7%)	computational chemistry (17.5%)	bio* modeling (7.8%)	chem* database* (7.6%)
INOR	electrochem* (67.1%)	nanomaterial* (14.9%)	organometallic* (5.8%)	metal organic framework* (4.0%)
NUCL	nuclear fuel* (28.6%)	isotope* (27.3%)	radioisotope* (15.0%)	nuclear medicine* (9.0%)
CARB	carbohydrate* (43.0%)	glycoprotein* (42.4%)	glycan* (6.6%)	oligosaccharide* (5.6%)
RUBB	rubber* (100.0%)			
CELL	cellulose (41.7%)	polysaccharide* (26.3%)	lignin (16.3%)	lignocellulos* (9.7%)
I&EC	water purification (38.1%)	industrial chem* (23.7%)	rare earth element* (11.9%)	industrial and engineering chemistry (8.3%)
FLUO	fluorine* (99.8%)	radiopharmaceutical chem* (0.2%)		
CHED	chem* class* (76.4%)	chem* communication* (8.5%)	chem* educat* (5.5%)	lab* safety (5.5%)
CHAS	chem* safety (51.5%)	lab* safety (16.7%)	environmental health and safety (16.7%)	chem* regulations (15.2%)
BMGT	chem* compan* (65.4%)	chem* enterprise* (28.8%)	chem* business* (3.8%)	chem* research and development (1.9%)
SCHB	commercial chem* (50.0%)	chem* sector* (28.6%)	academic entrepreneur* (14.3%)	science advoca* (7.1%)
HIST	chem* histor* (53.8%)	evolution of chem* (30.8%)	history of chem* (15.4%)	
PROF	chem* education (100.0%)			
CHAL	pharmaceutical patent* (60.0%)	chem* in commerce (30.0%)	chem* patent* (10.0%)	

CompChem, ML, and chemical and physical intuition (CPI). This review will classify concepts using a rendition of a "data to wisdom" hierarchy, Figure 2. Scholars have noted shortcomings with similar constructs,[21] but we use it to reflect a stepladder for scientific progress, starting from collecting data and ending with overall impact. CompChem, ML, and CPI each have different strengths and weaknesses and bring synergistic opportunities. CPI alone can be employed to climb the ladder from data to impact, but current CPI may only provide limited understanding or applicability outside of available data sets. However, CompChem is extraordinarily well-suited for generating high quality data that contain useful information (vide infra, section 2) often more easily than via traditional experimentation. ML is likewise extremely well-suited for recognizing and accurately quantifying nonlinear relationships (vide infra, section 3), a task that is especially difficult for even the most expert-level CPI alone. A key opportunity is that useful ML requires robust data sets, and these can be provided by CompChem as long as the CPI component is selecting and correctly interpreting appropriate methods for the task at hand to productively climb the ladder toward impact (vide infra, section 4). We stress that the impact generation process shown in Figure 2 is by no means a linear one — on the contrary, it contains many loops and dead ends.

Figure 2. Data−knowledge−wisdom hierarchy stepladder.

As we show later (in Section 4), within the troika of CompChem+ML+CPI, ML acts as a catalyst that accelerates explorative data-driven hypotheses generation. Automatically generated hypotheses are then validated and calibrated with CompChem and CPI to yield further improved ML modeling (enriched by more physical prior knowledge), which then loops back with improved hypotheses. This feedback loop is the key to the modern knowledge discovery leading to insight, wisdom and hopefully positive impacts to society.

2. COMPCHEM AND NOTABLE INTERSECTIONS WITH ML

2.1. Computational Modeling, Data, and Information Across Many Scales

We consider quantum mechanics as described by the nonrelativistic time-independent Schrödinger equation as our "standard model" because it accurately represents the physics of charged particles (electrons and nuclei) that make up almost all molecules and materials. Indeed, this opinion has been held by some for almost a century:

> *The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved.*

P. A. M. Dirac, 1929

Any theoretical method for predicting molecular or material phenomena must first be rooted in quantum mechanics theory and then suitably coarse-grained and approximated so that it can be applied in a practical setting. CompChem, or more precisely, computational quantum chemistry defines computationally driven numerical analyses based on quantum mechanics. In this section, we will explain how and why different CompChem methods capture different aspects of underlying physics. Specifically, this section provides a concise overview of the broad range of CompChem methods that are available for generating data sets that would be useful for ML-assisted studies of molecules and materials.

2.1.1. Models and Levels of Abstraction. Models extract information from data. The renowned statistician George Box famously discussed "good models" as those characterized as "simple", "illuminating", and "useful".[22] Good models should be *parsimonious* and describe essential relationships without overelaboration. The ideal gas equation, $PV = nRT$, exemplifies a good model. The ideal gas equation relates macroscopic pressure (P), volume (V), number of molecules (n), and temperature (T) of gases under idealized conditions, without requiring explicit knowledge of the processes occurring on an atomic scale. Its simple functional form needs just one parameter, the ideal gas constant R, and this makes it possible to formulate useful insights, such as how at constant pressure a gas expands with rising temperature. On the other hand, this elegant equation only holds for conditions where the gas behaves as an ideal gas. The derivation of more accurate models of gases requires more mathematically complicated equations of state that rely on more free parameters[23] that in turn obfuscate physical insights, require more computational effort to solve, and thus make the model less "good". This example also offers a convenient connection to ML models that will be discussed later in section 3. As mathematical models for complex phenomena become more complicated and less intuitive to derive, ML models that infer nonlinear

relationships from data become more applicable when increasing amounts of empirical data become available.

Alternatively, the conventional CompChem treatment entails first determining the system's relevant geometry and its total ground state energy, and from that physical properties of interest (e.g., pressure, volume, band gap, polarizability, etc.) can be obtained using quantum and statistical mechanics. In this section, we discuss the relevant CompChem methods for these. While the mathematical physics for these methods might occasionally be too complicated for a user to fully understand, many algorithms exist so that they can still be easily run in a "black-box" way with modern computational chemistry software and accompanying tutorials.[24−27] CompChem thus serves as an invaluable tool to generate data and information for knowledge and insights across many length and time scales. Figure 3 is an adaption of a multiscale hierarchy of different
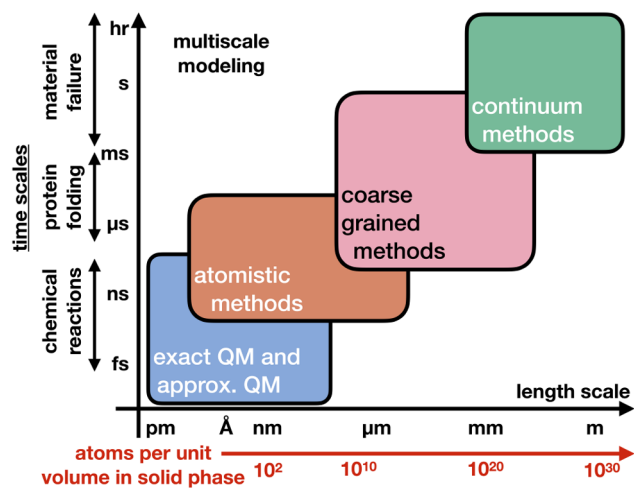
Figure 3. Hierarchy of computational methods and corresponding time and length scales. QM stands for Quantum Mechanics.

classes of CompChem methods. It shows their applicability for modeling different length and time scales and depicts how large scale models may be developed based on smaller scale theories.

2.1.2. CompChem Representations. Integral to every CompChem study is the user's *representation* for the system, that is, how the user chooses to describe the system. CompChem representations can range from simple and lucid (e.g., a precise chemical system such as a water molecule isolated in a vacuum) to complex and ambiguous (e.g., a putative but speculative depiction of a solid−liquid interface under electrochemical conditions). Approximate wavefunctions (expressed on a basis set of mathematical functions) or approximate Hamiltonians (referred to as levels of theory) as described below in this section can also be considered representations. One might then say that many representations for different components of a system will constitute an overall representation, and this is true. The point we make is that the validity of *any* computational result depends on the *overall* representation, and sometimes an incorrect representation may provide a correct result due to "fortuitous error cancellation". In CompChem studies, a valid representation is one that captures the nature of the physical phenomena of a system. For a molecular example, if one is determining the bond energy of a large biodiesel molecule using CompChem methods,[28] it may or may not be justified to approximate a nearby long-chain

alkyl group ($-C_nH_{(2n+1)}$) simply as a methyl ($-CH_3$) or even a hydrogen atom. Indeed, choosing such a representation can sometimes be a useful example of CPI since alkyl bonds usually exhibit relatively short-ranged interactions (a feature that will be discussed in the context of ML in more detail in section 4.1.3.). An atomic scale geometry with fewer atoms would reduce the computational cost of the study or allow a more accurate but more computationally expensive calculation to be run. On the other hand, it might also be a poor choice if the chemical group, for example, a substituted alkyl group participated in physical organic interactions, such as subtle steric, induction, or resonance effects.[29] For a solid-state example, a user might exercise good CPI by assuming that a relatively small unit cell under periodic boundary conditions would capture salient features of a bulk material or a material surface (as is often the case for many metals). On the other hand, subtle symmetry-breaking effects in materials (e.g., distortions arising from tilting octahedra groups in perovskites,[30] or surface reconstruction phenomena that occur on single crystals)[31] might only be observed when considering larger and more computationally expensive unit cells. Relevant to both examples, it may also be that the CompChem method itself brings errors that obfuscate phenomena that the user intends to model. In general, CompChem errors may be due to 1) errors introduced by the user in the initial set up of the CompChem application, or 2) errors in the CompChem method when treating the physics of the system. In section 3, we will discuss how the choice of ML representation also plays similarly critical roles in determining whether and to what extent an ML model is useful.

2.1.3. Method Accuracy. The quantitative accuracy of a CompChem model stems from its suitability in describing the system. As explained above, an observed accuracy will depend on the representation being used. High-quality CompChem calculations have traditionally been benchmarked against data sets that consist of well-controlled and relatively precise thermochemistry experiments on small, isolated molecules.[32,33] The error bars for standard calorimetry experiments are approximately 4 kJ/mol (or 1 kcal/mol or 0.04 eV), and computational methods that can provide greater accuracy than this are stated as achieving "chemical accuracy". Note that this term should be used when describing the accuracy of the method compared to the most accurate data possible; for example, if one CompChem method was found to reproduce another CompChem method within 1 kJ/mol, but both methods reproduce experimental data with errors of 20 kJ/mol, then neither method should be called chemically accurate. There are many well-established reasons why CompChem models can bring errors. For example, errors may be due to size consistency[34] or size extensivity[35] problems that are intrinsic within the CompChem method, larger systems sometimes embody significant medium and long-range interactions (e.g., van der Waals forces)[36] or self-interaction errors[37] that might not be noticeable in small test cases. The recommended path forward is to consider which fundamental interactions are in play in the system and then use a CompChem model that is adequate at describing those interactions. Besides this, users should make use of existing tutorial references that provide practical knowledge about which parameters in a CompChem calculation should be carefully noted, for example ref 38. Historically the most popular CompChem methods for molecular and materials modeling (the B3LYP[39] and PBE[40] exchange correlation

functionals, see section 2.2.3.) are often said to have an expected accuracy of about 10−15 kJ/mol (or 2−4 kcal/mol or 0.1−0.2 eV) when modeling differences between the total energies of two similar systems, and errors are expected to be somewhat larger when considering transition state energies. Though this is used as a simple rule, it is obviously an oversimplification and actual accuracy is only assessed by thoughtful benchmarking of the case being considered.[41−45]

2.1.4. Precision and Reproducibility. In CompChem, one normally assumes that any two users using the same representation for the system with the same code on the same computing architecture will obtain the exact same result within the numerical precision of the computers being used. This is not always the case, especially for molecular dynamics (MD) simulations that often rely on stochastic methods.[46] Computational precision also becomes more concerning when there are different versions of codes in circulation, errors that might arise from different compilers and libraries, and a lack of consensus in the community about which computational methods and which default settings should be used for specific application systems, for example, grid density selections,[47] or standard keywords for molecular dynamics simulations.[46,48] There have been efforts to confirm that different codes can reproduce energies for the same system representation,[48,49] but some commercial codes hold proprietary licenses that restrict publications that critically benchmark calculation accuracy and timings across different codes. A path forward to benefit the advancement of insight is the development of (open) source codes[50] that perform as well if not better than commercial codes. While increased access to computational algorithms is beneficial, it also raises the need for enforcing high standards of quality and reproducibility.[51,52] We are also glad to see active developments to more lucidly show how any set of computational data is generated, precisely with which codes, keywords, and auxiliary scripts and routines.[53−56] We are now in an era where truly massive amounts of data and information can be generated for CompChem+ML efforts. To go forward, one needs to know what constitutes good and useful data, and the next section provides an overview of how to do this using CompChem.

2.2. Hierarchies of Methods

Earlier we mentioned that a usual task in CompChem is to calculate the ground state energy of an atomic scale system. Indeed, CompChem methods can determine the energy for a hypothetical configuration of atoms, and this constitutes the potential energy surface (PES) of the system (Figure 4). The PES is a hypersurface spanning $3N$ dimensions, where N is the number of atoms in the system. Since the PES is used to analyze chemical bonding between atoms within the system, the PES can also be simplified by ignoring translational and rotational degrees of freedom for the entire system. This reduces the dimensionality of the PES from $3N$ to $3N - 5$ for linear systems (e.g., diatomic molecules or perfectly linear molecules such as acetylene) or $3N - 6$ for all other nonlinear systems. Furthermore, since visualization is difficult beyond three dimensions, PES drawings will show a 1-D or 2-D projection of this hypersurface where the z-axis is conventionally used to represent the scale for system energy.

Any arbitrary PES will contain several interesting features. Minima on the PES correspond to mechanically stable configurations of a molecule or material, for example reactant and product states of a chemical reaction or different
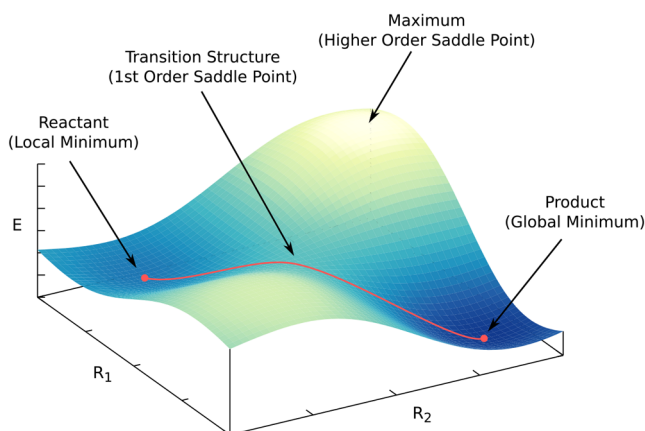
Figure 4. Potential energy surface (PES) of a fictional system with the two coordinates R_1 and R_2. The minima of the PES correspond to stable states of a system, such as equilibrium configurations and reactants or products. Minima can be connected by paths (red line), along which rearrangements and reactions can occur. The maximum along such a path is called a transition state. Transition states are first-order saddle points, a maximum in one coordinate and minima in all others. They correspond to the minimum energy required to transition between two PES minima and play a crucial role in the description of chemical transformations.

conformational isomers of a molecule. Because they are minima, the second derivative of the energy given by the PES with respect to any dimension will be positive. Minima can also be connected by pathways, which indicate chemical trans-formations (Figure 4, red line). Along such pathways, the second derivative can be positive, zero, or negative, but all other second derivatives must be positive. Transition states are first-order saddlepoints and thus represent a maximum in one coordinate and a minimum along all others. They correspond to the lowest energy barriers connecting two minima on the PES and are hence important for characterizing transitions between PES minima (e.g., chemical reactions). Second-order saddle points[57] and bifurcating pathways[58] can also exist, but these are not discussed further here.

A wide range of higher-level properties of the system can be predicted or derived using the PES, including predicted thermodynamic binding constants, kinetic rate constants for reactions, or properties based on dynamics of the system. The task is then to choose an appropriate CompChem method that can carry out energy and gradient calculations on the system's PES. Figure 5 shows several different hierarchies for CompChem methods capable of doing this. Note that all of these methods mentioned in this figure fall in the categories of the bottom two regions in the multiscale hierarchy Figure 3. All of these methods in principle could be used to develop
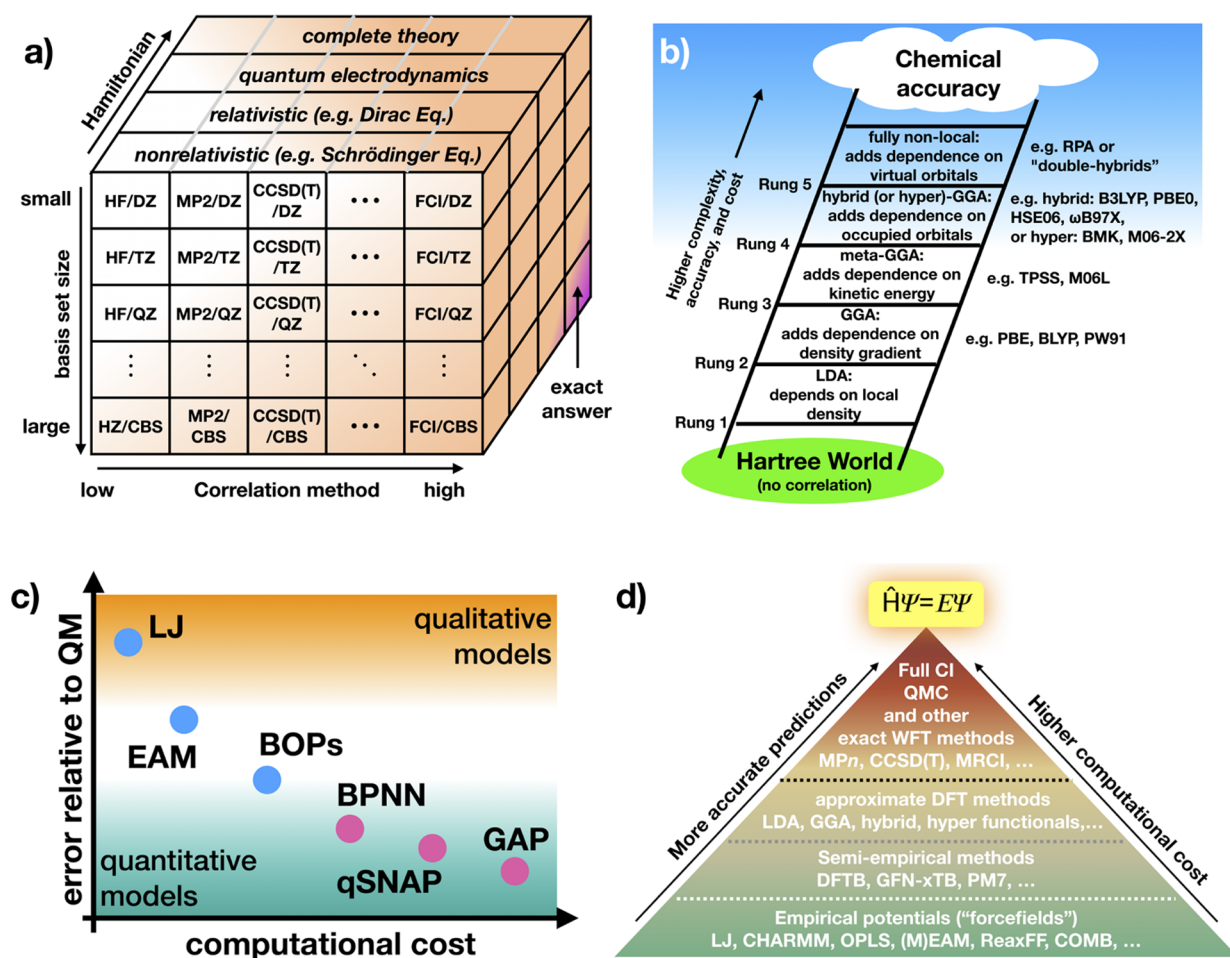
Figure 5. (a) "Magic cube"[59] depiction of hierarchies of correlated wavefunction approaches. (b) "Jacob's Ladder"[60] depiction of hierarchies of Kohn−Sham density functional theory (DFT) approaches. (c) Hierarchies of atomistic potentials. (d) Overall hierarchies in predictive atomic scale modeling methods.

coarse-grained or continuum models as well. Also note that methods in Figure 5 will bring very different computational costs and opportunities for methods involving ML.

2.2.1. Wavefunction Theory Methods. In standard computational quantum chemistry, a system's energy can be computed in terms of the Schrödinger equation.[61−63] The wavefunction that will be used to represent the positions of electrons and nuclei in the system ($\Psi(\mathbf{r}, \mathbf{R})$) is hard to intuit since it can be complex valued. However, its square describes the real probability density of the nuclear ($\mathbf{R}$) and electronic positions ($\mathbf{r}$). In a real system, the position and interactions of a single particle in the system with respect to all other particles will be *correlated*, and this makes exactly solving the Schrödinger equation impossible for almost all systems of practical interest. To make the problem more tractable, one may exploit the Born−Oppenheimer approximation;[64] since nuclei are expected to move much slower than the electrons they can be approximated as stationary at any point along the PES. This allows the energy to be calculated using the time-independent Schrödinger equation and solving the eigenvalue problem:

$$\hat{H}\Psi = (\hat{T} + \hat{V})\Psi = E\Psi \tag{1}$$

Here, the Hamiltonian operator ($\hat{H}$) is the sum of the kinetic ($\hat{T}$) and potential ($\hat{V}$) operators, Ψ is the wavefunction (i.e., an eigenfunction) that represents particles in the system, and E is the energy (i.e., an eigenvalue). In this way, nuclei can be treated as fixed point charges, and then, eq 1 can be transformed into the so-called electronic Schrödinger equation, where the Hamiltonian $\hat{H}_{el}$ and wavefunction $\Psi_{el}(\mathbf{r}; \mathbf{R})$ now only depend on the nuclear coordinates $\mathbf{R}$ in a parametric fashion:

$$\hat{H}_{el}\Psi_{el}(\mathbf{r}; \mathbf{R}) = [\hat{T}_e(\mathbf{r}) + \hat{V}_{eN}(\mathbf{r}; \mathbf{R}) + \hat{V}_{NN}(\mathbf{R}) + \hat{V}_{ee}(\mathbf{r})]$$
$$\Psi_{el}(\mathbf{r}; \mathbf{R})$$
$$= E_{el}\Psi_{el}(\mathbf{r}; \mathbf{R}) \tag{2}$$

The above expression has $\hat{H}_{el}$ composed of single electron (e) and pairwise electron−nuclear (eN), nuclear−nuclear (NN), and electron−electron (ee) terms. Here, we will now implicitly assume the Born−Oppenheimer approximation throughout and leave off the subscript indicating the electronic problem. However, we note that the Born−Oppenheimer approximation is not always sufficient and computationally intensive non-adiabatic quantum dynamics may be required.[65] In certain cases, semiclassical treatments are appropriate; for example, nonadiabatic effects between electrons and nuclei can be considered using nuclear-electronic orbital methods.[66]

A second common approximation is to expand the total electronic wavefunction in terms of one-electron wavefunctions (i.e., spin orbitals): $\phi(\mathbf{r}_i)$. Electrons are Fermions and therefore exhibit antisymmetry, which in turn results in the Pauli exclusion principle. Antisymmetry means that the interchange of any two particles within the system should bring an overall sign change to the wavefunction (i.e., from + to −, or vice versa). This property is conveniently captured mathematically by combining one electron spin orbitals into the form of a Slater determinant:

$$\Psi(\mathbf{r}_1, \cdots, \mathbf{r}_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(\mathbf{r}_1) & \cdots & \phi_n(\mathbf{r}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{r}_n) & \cdots & \phi_n(\mathbf{r}_n) \end{vmatrix} \tag{3}$$

Note that a determinant's sign changes whenever two columns or rows are interchanged, and in a Slater determinant this corresponds to interchanging electrons and thus the physically appropriate sign change for the overall wavefunction. Additionally, $\frac{1}{\sqrt{n!}}$ is a normalizing factor to ensure the wavefunction is unitary.

The spin orbitals can be treated as a mathematical expansion using a basis set of μ functions χ_μ, each having coefficients $c_{\mu i}$, which are generally Gaussian basis functions,[67−69] Slater-type hydrogenic orbitals,[70] or plane waves under periodic boundary conditions:[71−73]

$$\phi_i = \sum_\mu c_{\mu i} \chi_\mu \tag{4}$$

The different types of mathematical functions bring different strengths and weaknesses, but these will not be discussed further here. A universal point is that larger basis sets will have more basis functions and thus give a more flexible and physical representation of electrons within the system. On one hand this can be crucial for capturing subtle electronic structure effects due to electron correlation. On the other hand, larger basis sets also necessitate significantly higher computational effort. A standard technique to avoid high computational effort in electronic structure calculations is to replace nonreacting core electrons with analytic functions using effective core potentials (ECPs, i.e., pseudopotentials).[74−89] This requires reformulating the basis sets that describe the valence space of the atoms, for example see refs 90 and 91. Larger nuclei that bring higher atomic numbers and larger numbers of electrons will also exhibit relativistic effects,[92] and relativistic Hamiltonians are based on the Dirac equation[93,94] or quantum electrodynamics.[95] These methods can range from reasonably cost-effective methods[96,97] to those bringing extremely high computational cost.[98] Practical applications have traditionally used standard nonrelativistic Hamiltonian methods, along with ECPs (or pseudopotentials) that have been explicitly developed to account for compressed core orbitals that result from relativistic effects.

Using the Born−Oppenheimer approximation (eq 2) together with a Slater determinant wavefunction (eq 3) expressed in a finite basis set (eq 4) brings about the simplest wavefunction based method, the Hartree−Fock (HF) approach (for historical context see refs 99−101). The HF method is a mean field approach, where each electron is treated as if it moves within the average field generated by all other electrons. It is generally considered inaccurate when describing many chemical systems, but it continues to serve as a critical pillar for CompChem electronic structure calculations since it either establishes the foundation for all other accurate methods or provides energy contributions (i.e., exact exchange) that is not provided in some CompChem methods. CompChem methods that achieve accuracy higher than HF theory are said to contain *electron correlation*, a critical component for understanding molecules and materials (as described in more detail in section 2.2.2.). Expressing Ψ as a Slater determinant and rearranging eq 2 while temporarily neglecting nuclear−nuclear interactions allows one to define

the HF energy in terms of integrals of the electronic spin orbitals:

$$
\begin{aligned}
E_{\mathrm{HF}} = & -\sum_i \int d\mathbf{r}_1 \phi_i^*(\mathbf{r}_1) \underbrace{\frac{1}{2}\nabla^2}_{\widehat{T}_e} \phi_i(\mathbf{r}_1) \\
& -\sum_i \int d\mathbf{r}_1 \phi_i^*(\mathbf{r}_1) \underbrace{\sum_\alpha^N \frac{Z_\alpha}{|\mathbf{r}_1 - \mathbf{R}_\alpha|}}_{\widehat{V}_{\mathrm{eN}}} \phi_i(\mathbf{r}_1) \\
& +\sum_{i,j>i} \iint d\mathbf{r}_1\, d\mathbf{r}_2 \phi_i^*(\mathbf{r}_1)\phi_i(\mathbf{r}_1) \underbrace{\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}}_{\widehat{V}_{\mathrm{ee}}(\text{Coulomb})} \phi_j^*(\mathbf{r}_2)\phi_j(\mathbf{r}_2) \\
& -\sum_{i,j>i} \iint d\mathbf{r}_1\, d\mathbf{r}_2 \phi_i^*(\mathbf{r}_1)\phi_j(\mathbf{r}_1) \underbrace{\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}}_{\widehat{V}_{\mathrm{ee}}(\text{exchange})} \phi_j^*(\mathbf{r}_2)\phi_i(\mathbf{r}_2)
\end{aligned}
\tag{5}
$$

where the first two terms are referred to as one-electron integrals and represent the kinetic energy of the electrons and the potential energy contributions from electron-nuclei interactions. The remaining terms are two-electron integrals that describe the potential energy arising from electron−electron interactions and are called Coulomb and exchange integrals. Using Lagrange multipliers, one can express the HF equation in a compact matrix form, the so-called Roothan−Hall equations,[102−104] which allow for an efficient solution:

$$
\mathbf{FC} = \epsilon\mathbf{SC} \tag{6}
$$

Each matrix has a size of $\mu \times \mu$, where μ is the number of basis functions used to express the orbitals of the system. $\mathbf{C}$ is a coefficient matrix collecting the basis coefficients $c_{\mu i}$ (see eq 4), while $\mathbf{S}$ is the overlap matrix measuring the degree of overlap between individual basis functions and ϵ is a diagonal matrix of the spin orbital energies. Finally, $\mathbf{F}$ is the Fock matrix, with elements of a similar form as in eq 5, but expressed in terms of basis functions χ_μ. One important detail not readily apparent in eq 6 is that the Fock matrix depends on the orbital coefficients that must be provided before eq 6 can be solved. As such, eq 6 cannot be solved in closed form, but instead requires a so-called self-consistent field approach. Starting from an arbitrary set of trial (i.e., initial guess) functions, one iteratively solves for optimal molecular orbital coefficients, which are then used to construct a new Fock matrix, until a minimum energy is reached in accordance with the variational principle of quantum mechanics. Evaluating and transforming the two-electron integrals in eq 5 are a significant bottleneck for these calculations and thus the computational effort of the HF methods formally scales as $O(\mu^4)$ with the number of basis functions. This means that a calculation on a system twice as large will require at least $2^4 = 16$ times as much computing time. The electronic exchange interaction resulting from the antisymmetry of the wavefunction imposes a strong constraint on the mathematical form of ML models for electronic wavefunctions. Construction of efficient and reliable antisymmetric ML models for the many-body wavefunction is an important area of current research.[105,106]

2.2.2. Correlated Wavefunction Methods. The system's *correlation energy* is defined as sum of electron−electron interactions that originate beyond the mean-field approximation for electron−electron interactions that is provided by HF theory. While correlation energy makes up a rather small contribution to the overall energy of a system (usually about 1% of the total energy), because internal energies in molecular and material systems are so enormous, this contribution becomes rather significant. As an example, most molecular crystals would be unstable as solids if calculated using the HF level of theory. The missing component is attractive forces that are obtained from levels of theory that account for correlation energy. Correlation energies are obtained by calculating additional electron−electron interaction energies that arise from different arrangements of electron configurations (i.e., different possible excited states) that are not treated with the mean field approach of HF theory.

The most complete correlation treatment is the full configuration interaction (FCI) method, which is the exact numerical solution of the electronic Schrödinger equation (in the complete basis limit) that considers interactions arising from all possible excited configurations of electrons. The FCI wavefunction takes the form of a linear combination of all possible excited Slater determinants which can be generated from a single HF reference wavefunction by electron excitations:

$$
\Psi_{\mathrm{FCI}} = a_0\Psi_{\mathrm{HF}} + \sum_{\alpha,\beta} a_\beta^\alpha \Psi_\beta^\alpha + \sum_{\alpha,\beta,\gamma,\delta} a_{\beta\delta}^{\alpha\gamma}\Psi_{\beta\delta}^{\alpha\gamma} + \dots \tag{7}
$$

where Ψ_β^α represents the Slater determinant obtained by exciting an electron from orbital α into an unoccupied orbital β, and the as are expansion coefficients determining the weight of the different contributing configurations. Expectedly, FCI calculations scale extremely poorly with the number of electrons in the system ($O(n!)$), as the number of possible configurations grows rapidly, making them feasible only for small molecules. For an example of the state of the art, FCI calculations have been used to benchmark highly accurate methods on calculations on a benzene molecule.[107]

Most correlated wavefunction methods use a subset of the possible configurations in eq 7 to be computationally tractable. The configuration interaction (CI)[108] method for example only includes determinants up to a certain permutation level (e.g., single and double excitations in CISD). Alternatively, MPn[35] (e.g., MP2) recovers the correlation energy by applying different orders of perturbation theory. Coupled cluster theory, another widely used post-HF method, includes additional electron configurations via cluster operators.[109] One coupled cluster method that involves single, double, and perturbative triples excitations, CCSD(T), is referred to as the "gold-standard" approach for CompChem electronic structure methods since it brings high accuracy for molecular energies. However, there are many newer advances that improve upon CCSD(T).[107,110] Note that just because a method has a reputation for being accurate does not mean that it will be for all systems. For example, consider again the benzene molecule, which is best illustrated having dotted resonance bond depicting a planar molecule with equal C−C bond lengths. Such a geometry will not be found to be stable with many different CompChem methods, in part because of subtle chemical bonding interactions or errors that arise from specific choices of basis sets used with different levels of theory.[111,112]

A key point to reiterate is that correlated wavefunction methods are founded on the HF theory, and so they are even more computationally demanding than HF calculations, for example, $O(n^5)$ for MP2, $O(n^6)$ for CCSD and CISD and $O(n^7)$ for CCSD(T). However, this computational expense is

alleviated by continually improving computing resources (e.g., the usability of graphics processing units (GPUs))[113−116] and the development of efficiency enhancing algorithms, such as pseudospectral methods,[117−119] resolution of the identity (RI),[120] domain-based local pair natural orbital methods (DLPNO),[121] and explicitly correlated R12/F12 methods.[122] There are also ongoing efforts to develop other CompChem methods based on quantum Monte Carlo[123] and density matrix renormalization group theory (DMRG)[124] to provide high accuracy with competitive scaling with other computational methods. Efforts are beginning to become implemented that use ML to accelerate these types of calculations.[105,106,125−129]

Schemes have also been developed to exploit systematic errors between different levels of theory with different basis sets so that approximations can be extrapolated toward an exact result. Examples include the complete basis set (CBS),[130] Gaussian Gn,[131] Weizmann (W-)n[132] methods, and high accuracy extrapolated ab initio thermochemistry (HEAT)[133] methods. For a recent review on these and other methods, see ref 134. These schemes are also becoming a target of recent work using ML methods.[135]

HF determinants provide good baseline approximations of the ground state electronic structure of many molecules, but they may describe poorly more complicated bonding that arises during bond dissociation events, excited states, and conical intersections.[136−139] Some many-body wavefunctions are best described as a superposition of two or more configurations, for example, when other configurations in eq 7 can have similar or higher expansion coefficients a than the HF determinant. For this reason, high quality single reference methods like CCSD(T) fail because the theory assumes that salient electronic effects are captured by the initial single HF configuration. (In fact, methods such as CCSD(T) have been implemented with diagnostic approaches available that let users know when there may be cause for concern).[140−142] In these cases, it may no longer be trivial to find reliable black-box or automated procedures (e.g., in situations involving resonance states, chemical reactions, molecular excited states, transition metal complexes, and metallic materials, etc.).[136] So-called multiconfiguration approaches,[136] such as the generalized valence bond (GVB) method[143] or the complete active space self-consistent field (CASSCF),[144] the multireference CI (MRCI) methods,[145] complete active space perturbation theory (CASPT2),[146] or multireference coupled cluster (MRCC),[147,148] can more physically model these systems since they employ several suitable reference configurations with different degrees of correlation treatments. These methods are not black-box and should be expected to require an experienced practitioner with CPI to choose the reference states that can substantially influence the quality of results.[149] This is an area though where ML can bring progress in automating the selections of physically justified active spaces.[129]

In closing, there are a large number of available correlated wavefunction methods but many are even more costly than HF theory by virtue of requiring an HF reference energy expression shown in eq 5. Figure 5a depicts a so-called "magic cube" (that is an extension beyond a traditional "Pople diagram"[135,150]) that concisely shows a full hierarchy of computational approaches across different Hamiltonians, basis sets, and correlation treatment methods. This makes it easy to identify different wavefunction methods that should be

more accurate and more likely to provide useful atomic scale insights (as well as those that would be more computationally intensive). Another important aspect highlighted in the "magic cube" is that higher level wavefunction methods require larger basis sets to successfully model electron correlation effects. A CCSD(T) computation carried out with a small basis set for example might only offer the same accuracy as MP2 while being two orders of magnitude more expensive to evaluate.[108] As was mentioned earlier with the benzene system, spurious errors with different basis sets might still be found that indicate problems with specific combinations of levels of theory and basis sets. The deep complexity of correlated wavefunction methods makes this a promising area for continued efforts in CompChem+ML research.

2.2.3. Density Functional Theory. Density-functional theory (DFT)[151] is another method to calculate the quantum mechanical internal energy of a system using an energy expression that relies on functionals (i.e., a function of a function) of electronic density $\rho = |\Psi_{el(r; R)}|^2$:

$$E[\rho] = T[\rho] + V[\rho] \tag{8}$$

Compared to wavefunction theory, DFT should be far more efficient since the dimensionality of a density representation for electrons will always be three rather than the $3n$ dimensions for any n-electron system described by a many-body wavefunction method. DFT has an important drawback that the exact expression for the energy functional is currently unknown, all approximations bring some degree of uncontrollable error, and this has precipitated disagreeable opinions from purists in chemical physics, especially those who are developing correlated wavefunction methods. However, there is also substantial evidence that DFT approximations are reasonably reliable and accurate for many practical applications that bring information, knowledge, and sometimes insight. We now provide a bird's-eye view of DFT-based methods.

One thrust of DFT developments since its inception has focused on designing accurate expressions strictly in terms of a density representation, and these approaches are referred to as "kinetic energy (KE-)" or "orbital-free (OF-)" DFT.[152] Some energy contributions (e.g., nuclear-electron energy and classical electron−electron energy terms) can be expressed exactly, but other terms, such as the kinetic energy as a function of the density are not known and must be approximated. OF-DFT is very computationally efficient (these methods should scale linearly with system size[153,154]) but these formulations have not yet been developed to rival the accuracy or transferability of wavefunction methods, though they have been used for studying different classes of chemical and materials systems.[155−157] OF-DFT methods are also used in exciting applications modeling chemistry and materials under extreme conditions.[158−160] One should expect that once highly accurate forms are developed and matured, accurate CompChem calculations on electronic structures on systems having more than a million atoms might become commonplace. Indeed, there are efforts to use ML to develop more physical OFDFT methods.[161,162]

The most commonly used form of DFT (which is also one of the most widely used CompChem methods in use today) is called Kohn−Sham (KS-)DFT.[163] In KS-DFT, one assumes a fictitious system of noninteracting electrons with the same ground state density as the real system of interest. This makes it possible to split the energy functional in eq 8 into a new

form that involves an exact expression of the kinetic energy for noninteracting electrons:

$$E[\rho] = T_{ni}[\rho] + V_{eN}[\rho] + V_{ee}[\rho] + \Delta T_{ee}[\rho] + \Delta V_{ee}[\rho]$$

(9)

Here, $T_{ni}[\rho]$ is the kinetic energy of the noninteracting electrons, $V_{eN}[\rho]$ is the exact nuclear-electron potential, and $V_{ee}[\rho]$ is the Coulombic (classical) energy of the noninteracting electrons. The last two terms are corrections due to the interacting nature of electrons and nonclassical electron–electron repulsion. KS-DFT also expands the three-dimensional electron density into a spin orbital-basis ϕ similar to HF theory to define the one-electron kinetic energy in a straightforward manner. This allows the T_{ni}, V_{eN}, and V_{ee} expressions to be evaluated exactly and one arrives at the KS energy:

$$E_{KS}[\rho] = - \underbrace{\sum_i \int d\mathbf{r}_1 \phi_i^*(\mathbf{r}_1) \frac{1}{2}\nabla^2 \phi_i(\mathbf{r}_1)}_{T_{ni}}$$

$$- \underbrace{\sum_i \int d\mathbf{r}_1 \phi_i^*(\mathbf{r}_1) \sum_\alpha^N \frac{Z_\alpha}{|\mathbf{r}_1 - \mathbf{R}_\alpha|} \phi_i(\mathbf{r}_1)}_{V_{eN}}$$

$$+ \underbrace{\sum_i \int d\mathbf{r}_1 \phi_i^*(\mathbf{r}_1) \left[\frac{1}{2} \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r}_1 - \mathbf{r}'|} \right] \phi_i(\mathbf{r}_1)}_{\hat{V}_{ee}(\text{Coulomb})}$$

$$+ E_{xc}[\rho]$$

(10)

The last two correction terms in eq 9 arise from electron interactions, and these are combined into the so-called "exchange-correlation" term (E_{xc}), which uniquely defines which scheme of KS-DFT is being used. In theory, an exact E_{xc} term would capture all differences between the exact FCI energy and the system of noninteracting electrons for a ground state.

The KS-DFT equations can be cast in a similar form as the Roothan–Hall equations (eq 6), which allows for a computationally efficient solution. Moreover, the elements of the KS matrix (which replaces the Fock matrix $\mathbf{F}$) are easier to evaluate due to the fact that several of the computationally intensive integrals are now accounted for via E_{xc}. Hence, the formal scaling for KS-DFT is $O(n^3)$ with respect to the number of electrons. Even though this is much poorer scaling than ideally linear scaling OF-DFT, the exact treatment of noninteracting electrons makes KS-DFT more accurate. Furthermore, there are several modern exchange-correlation functionals that routinely achieve much higher accuracy than HF theory with less computational cost, and thus KS-DFT is a competitive alternative with many correlated wavefunction methods in many modern applications.

A remaining problem is constructing a practical expression for the exchange-correlation functional, as its exact functional form remains unknown. This has spawned a wealth of approximations that have been founded with different degrees of first principles and/or empirical schemes. Classes of KS-DFT functionals are defined by whether the exchange-correlation functional is based on just the homogeneous electron gas (i.e., the "local density approximation", LDA), that and its derivative (i.e., the "generalized gradient approximation", GGA), as well as other additional terms that should

result in physically improved descriptions or error cancellations. The resulting hierarchy of KS-DFT functionals is often referred to as a "Jacob's Ladder" of DFT (Figure 5b). Generally, the higher up the ladder one goes, the more accurate but more computationally demanding the calculation.[164] However, the intrinsic inexactness in DFT makes it difficult to assess which functionals are physically better than others.[165,166] Nevertheless, the Jacob's Ladder hierarchy is useful for clearly designating how and why newer methods should perform in specific applications (for perspective see refs 167–169).

Indeed, by being based on a ground-state representation for homogeneous electron gas, DFT calculations can sometimes bring more easily physical insight into some systems that are very challenging for wavefunction theory to examine (e.g., metals, where HF theory provides divergent exchange energy behaviors[170,171]). On the other hand, DFT is also generally not well-suited for studying physical phenomena involving localized orbitals or band structures such as those found in semiconducting materials with small band gaps, molecular or material excited charge transfer states, or interaction forces that can arise due to excited states, e.g. dispersion (or London) forces. The former features can normally be treated using Hubbard-corrected DFT+U models that require a system-specific $U-J$ parameter[172,173] or more generalizable but much more computationally expensive hybrid DFT approaches. Dispersion forces (i.e., van der Waals interactions) are nonexistent in semilocal DFT approximations, and it is now commonplace to introduce them into DFT calculations using a variety of different methods.[36]

There is also growing interest in using embedded CompChem calculation schemes that can partition systems into discrete regions that could be treated with highly accurate correlated wavefunction theory and computationally efficient KS-DFT schemes separately.[174−178] DFT has also been extended to the modeling of excited states in the form of time-dependent (TD-)DFT.[179] Similar to ground state DFT, TDDFT is a less computationally expensive alternative to excited state wavefunction-based methods. The approach yields reasonable results where excitations induce only small changes in the ground state density, e.g. low lying excited states.[179,180] However, due to its single reference nature, TDDFT tends to break down in situations where more than one electronic configuration contribute significantly to the excited state. Just as with correlated wavefunction methods, there are already signs of CompChem+ML efforts to improve the applicability of DFT-based methods.[181−185]

2.2.4. Semiempirical Methods. Correlated wavefunctions and, to a lesser degree, KS-DFT are still very computationally demanding and only of limited use for large scale simulations. Further approximations based on wavefunctions and DFT methods have been developed to simplify and accelerate energy calculations. These so-called semiempirical methods still explicitly consider the electronic structure of a molecule but in a more approximate way than methods described above.

Semiempirical approaches based on wavefunction theory include methods like extended Hückel theory and neglect of diatomic differential overlap (NDDO).[186] Both approaches are simplifications of the HF eqs (eq 5) by introducing approximations to the different integrals. In the NDDO approach,[187] only the two-electron integrals in eq 5 are considered, where the two orbitals on the right and left-hand side of the $\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}$ operator are located on the same atom. The

Table 2. Types of Interatomic Potentials and Their Areas of Application

potential	reactive	typical applications	examples
pairwise-distance-based	sometimes	materials, liquids	Lennard-Jones,[208,209] Morse,[210] Buckingham[211]
distance and angle-based	usually no	materials, liquids	many water potentials (e.g., SPC, TIP4P, mW),[212] Stillinger−Weber[213]
class I (nonpolarizable) force fields	no	proteins, lipids, polymers, nucleic acids, carbohydrates, organic molecules, liquids	AMBER,[214,215] GAFF,[216] CHARMM,[217] GROMOS,[218−220] OPLS,[221,222] DREIDING,[223] MMFF94,[224] UFF,[225] COMPASS,[226] INTERFACE,[227] interatomic potentials for ionic systems[228]
class II (polarizable)	no	proteins, lipids, polymers, nucleic acids, carbohydrates, organic molecules, liquids	AMOEBA,[229] classical Drude oscillator models,[230] fluctuating charge (FQ) models,[231] MB-Pol,[232] distributed point polarizable models (DPP2),[233] and many more[234]
embedded atom method (EAM)-like	yes	reactions within solid materials	EAM,[235] MEAM,[236] Finnis−Sinclair,[237] Sutton−Chen[238]
bond-order potentials (BOPs)	yes	reactions within solids, liquids, gases	Brenner,[239] Tersoff,[240,241] REBO,[239,242] COMB,[243,244] ReaxFF,[245,246] APT[247]
other quantum mechanics-derived force fields	yes	reactions within liquids and gases	EVB[248] and related models[249,250]

remaining two-center (and one-center) integrals are then approximated by introducing a set of empirical functions, one for each unique type of integral. Moreover, the overlap matrix in eq 6 is assumed to be diagonal, which greatly simplifies the energy evaluation. This reduces the required computational effort tremendously and allows the scaling of these approaches to be reduced to $O(N^2)$. NDDO serves as a basis for more sophisticated semiempirical schemes, such as AM1,[188] PM7,[189] and MNDO,[190] where the energy is usually determined self-consistently using a minimally sized basis set. Inadequacies in theory can be compensated by different empirical parametrization schemes that can allow these calculations to rival the accuracy of higher level theory for some systems. For example Dral et al.[191] provided a recent "big-data" analysis of the performance of several semiempirical methods with large data sets.

Semiempirical schemes are also carried over to approximate KS-DFT with so-called density functional tight binding (DFTB).[192] DFTB simplifies the KS eqs (eq 10) by decomposing the total electron density ρ into a density of free and neutral atoms ρ_0 and a small perturbation term $\delta\rho_0$ ($\rho = \rho_0 + \delta\rho_0$). Expanding eq 10 in the perturbation $\delta\rho_0$ makes it possible to partition the total energy into three terms amendable to different approximation schemes:

$$E[\delta\rho_0] = E_{rep} + E_{Coul}[\delta\rho_0] + E_{BS}[\delta\rho_0] \quad (11)$$

E_{rep} is a repulsive potential containing interactions between the nuclei and contributions from the exchange correlation functional (these are typically approximated via pairwise potentials). The charge fluctuation term E_{Coul} is modeled as a Coulomb potential of Gaussian charge distributions computed from the approximate density. Finally, E_{BS} refers to the "band structure" term, which considers the electronic structure and contains contributions from T_{ni}, V_{eN}, and the exchange correlation functional (see eq 10). To compute E_{BS}, the density is expressed in a minimal basis of atomic orbitals, similar as in NDDO. The necessary Hamiltonian and overlap integrals are then evaluated via an approximate scheme based on Slater−Koster transformations. In addition to the energy, atomic partial charges are also computed in this step, which are then used in E_{Coul}. As a consequence, DFTB equations can also be solved self-consistently. DFTB methods are parametrized by finding suitable forms for the repulsive potential and adjusting the parameters used in the Slater−Koster integrals. Non-self-consistent and self-consistent tight-binding DFT meth-

ods[193,194] have been developed for simulating large scale systems. Semiempirical methods have also been a target of different ML schemes, yielding improved parametrization schemes and more accurate functional approximations.[195−198]

2.2.5. Nuclear Quantum Effects. The quantum nature of lighter elements, such as H−Li, and even heavier elements that form strong chemical bonds (C−C bond in graphene for example[199]) gives rise to significant nuclear quantum effects (NQEs). Such effects are responsible for large differences from the Dulong−Petit limit of the heat capacity of solids, isotope effects, and the deviations of the particle momentum distribution from the Maxwell−Boltzmann equation.[200] To capture NQEs, path-integral molecular dynamics (PIMD)[201,202] or centroid molecular dynamics (CMD)[203,204] can be used, but these methods are associated with much higher computational costs (usually about 30 times higher) compared with classical MD simulations using point nuclei. Moreover, because systems may be influenced by competing NQEs, the extent of NQEs is sensitive to the potential energy surface assumed. (Semi)local DFT approaches may not even qualitatively predict isotope fractionation ratios, and usually hybrid DFT is needed to reach quantitative accuracy.[205] However, employing hybrid DFT calculations or other high level methods in PIMD/CMD simulations can accrue extremely high computational costs. For this reason, ML force fields have been proposed as efficient means to carry out PIMD simulations, enabling essentially exact quantum-mechanical treatment of both electronic and nuclear degrees of freedom, at least for small molecules with dozens of atoms.[206,207]

2.2.6. Interatomic Potentials. Interatomic potentials introduce an additional level of abstraction compared to methods described above. Instead of using exact quantum mechanical expressions to create the PES for the system, analytic functions are used to model a presupposed PES that contains explicit interactions between atoms, while electrons are treated in an implicit manner (sometimes using partial charge schemes).[251−256] Interatomic potentials thus are (oftentimes dramatically) more computationally efficient than correlated wavefunction, DFT, and semiempirical approaches. This efficiency makes it possible to study even larger systems of atoms (e.g., biomolecules, surfaces, and materials) than is possible with other computational methods. Note that different empirical potentials bring substantially different computational efficiencies; for example Lennard-Jones (LJ) potentials are

more efficient than classical forcefields (FFs) like AMBER and CHARMM, while those are more efficient than most bond-order potentials, such as ReaxFF.[245,246] The degree of efficiency arises from the balance of using accurate or physically justified functional forms, approximations, and model parametrizations. There are many different formulations (see Figure 5c), and we will discuss the most general classes. An overview of the different types of potentials and their features is provided in Table 2. For extensive discussions on these methods including semiempirical approaches, we refer to the extensive review by Akimov and Prezhdo (ref 257). An excellent review for interatomic potentials is provided by Harrison et al. (ref 258), and an excellent overview of modern methods can be found in a special issue of *J. Chem. Phys.*[259]

The distinctions between different types of FFs can be blurry sometimes, and we will differentiate categories in ascending complexity. One of the simplest interatomic potentials is the LJ potential:[260]

$$E_{LJ} = \sum_{i,j>i} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{12}$$

It models the total energy as the sum of all pairwise interaction between atoms i and j using an attractive and repulsive term depending on the interatomic distance r_{ij}. ε_{ij} modulates the strength of the interaction function, while σ_{ij} defines where it reaches its minimum. The LJ potential is a prototypical "good model" of interatomic potentials, as it has a sufficiently simple physical form with only two parameters while still yielding useful results.

For covalent systems, such as bulk carbon or silicon, just pairwise distances are not sufficient to capture the local coordination of the atoms, and many empirical potentials[212,213,261] for these systems were expressed as a function of the pairwise distances and three-body terms within a certain cutoff distance. The pairwise term can take the form of LJ-type, electrostatic, or harmonic potentials, and the three-body term is usually a function of the angles formed by sets of three atoms.

So-called class I classical FFs introduce a more complicated energy expression:

$$E_{tot} = \sum_{bonds} k_{ij}(r_{ij} - \overline{r}_{ij})^2 + \sum_{angles} k_{ijk}(\theta_{ijk} - \overline{\theta}_{ijk})^2$$
$$+ \sum_{dihedral} \sum_{\gamma} k_{ijkl}^{(\gamma)}[1 - \cos(\gamma\phi_{ijkl} - \overline{\phi}_{ijkl}^{(\gamma)})] + \sum_{i,j>i} \frac{q_i q_j}{r_{ij}} + E_{LJ} \tag{13}$$

The first three terms are the energy contributions of the distances (r_{ij}), angles (θ_{ijk}) and dihedral angles (ϕ_{ijkl}) between bonded atoms. Because of this, they are also referred to as bonded contributions. Bond and angle energies are modeled via harmonic potentials, with the k_{ij} and k_{ijk} parameters modulating the potential strength and $\overline{r}_{ij}$ and $\overline{\theta}_{ijk}$ are the equilibrium distances and angles. The dihedral term is modeled with a Fourier series to capture the periodicity of dihedral angles, with k_{ijkl} and ϕ_{ijkl} as free parameters. The last two terms account for nonbonded interactions. The long-range electrostatics are modeled as the Coulomb energy between charges q_i and q_j, and the van der Waals energy is treated via a LJ potential (eq 12). In Class I/II FFs, empirical parameters are tabulated for a variety of elements in wide ranges of

chemical environments (for example ref 262). Parameters for any one system should not necessarily be assumed to transfer well to other systems, and reparametrizations may be needed depending on the application. Different sets of parametrization schemes give rise to different types of classical FFs, with CHARMM,[217] Amber,[214,215] GROMOS,[218−220] and OPLS[221,222] being a few of many examples.

An extension beyond these FFs are class II (i.e., "polarizable") FFs, where the static charges are replaced by environment dependent functions (e.g., AMOEBA[263]). A significant advantage to the class I and II types of FFs is that they are computationally efficient, which makes them well suited for MD simulations of complex and extended (bio)-molecules, such as proteins, lipids, or polymers. Implementations of FF calculations on GPUs makes these simulations extremely productive.[264−268] A disadvantage of Class I and II types of interatomic potentials is that they rely on predefined bonding patterns to compute the total energy, and this limits their transferability. In general, bonds between atoms are defined at the beginning of the simulation run and cannot change. Furthermore, bonding terms make use of harmonic potentials that are not suitable for modeling bond dissociation.

Reactive potentials, which eschew harmonic potential dependencies and thus can describe the formation and breaking of chemical bonds, include the embedded atom method (EAM, Figure 5c), which is used widely in materials science.[235] EAM is a type of many-body potential primarily used for metals, where each atom is embedded in the environment of all others. The total energy is given by

$$E_{tot} = \sum_{i}^{N} \left[F_i(\tilde{\rho}_i) + \frac{1}{2} \sum_{j \neq i} V_{ij}(r_{ij}) \right] \tag{14}$$

F_i is an embedding function and $\tilde{\rho}_i$ an approximation to the local electron density based on the environment of atom i. $F_i(\tilde{\rho}_i)$ can be seen as a contribution due to nonlocalized electrons in a metal. V_{ij} is a term describing to the core−core repulsion between atoms. An EAM potential is determined by the functional forms used for F_i and V_{ij}, as well as how the density is expressed. Its dependence on the local environment without the need for predefined bonds make EAM well suited for modeling material properties of metals. An extension of EAM is modified EAM (MEAM),[236] which includes directional dependence in the description of the local density $\tilde{\rho}_i$, but this brings greater computational cost. EAMs also form the conceptual basis of the embedded atom neural network (EANN) machine learning potentials (MLPs).[269]

Another common type of reactive potentials are bond-order potentials (BOPs). In general, BOPs model the total energy of a system as interactions between the neighboring atoms:

$$E_{tot} = \sum_{i,j>i} [V_{rep}(r_{ij}) - b_{ij(k)}V_{att}(r_{ij})]f_{cut}(r_{ij}) \tag{15}$$

V_{rep} and V_{att} are repulsive and attractive potentials depending on the interatomic distance r_{ij}. A cutoff function f_{cut} restricts all interactions to the local atomic environment. $b_{ij(k)}$ is the bond order term, from which the potential takes its name. This term measures the bond order between atoms i and j (i.e., "1" for a single bond, "2" for a double bond, and "0.6" for a partially dissociated bond). Bond orders can also depend on neighboring atoms k in some implementations. BOPs are typically used for covalently bound systems, such as bulk solids

and liquids containing hydrogen, carbon or silicon (e.g., carbon nanotubes and graphene). Depending on the exact form of the expressions in eq 15, different types of BOPs are obtained, such as Tersoff[240,241] and REBO[239,242] potentials. BOPs can also be extended to incorporate dynamically assigned charges, yielding potentials like COMB[243,270] or ReaxFF.[245,246] As with EAMs, BOPs have also been used as a starting point for constructing more elaborate MLPs[271−273] that will also be discussed in more detail in section 3.

While efficient and versatile, all interatomic potentials described above are inherently constrained by their functional forms. A different approach is pursued by MLPs, such as Behler−Parinello Neural Networks,[274] q-SNAP,[275] and GAP potentials[276] (Figure 5c). In MLPs, suitable functional expressions for interactions and energy are determined in a fully data-driven manner and ultimately only limited by the amount and quality of available reference data. One can then use substantially more data to generate a much more accurate MLP than would be possible when using, for instance, a ReaxFF potential trained on similar data sets.[277]

For the sake of completeness, we note that all approaches described here are fully atomistic—each atom is modeled as an individual entity. It is also possible to combine groups of atoms into pseudoparticles giving rise to so-called coarse grained methods. On an even higher level of abstraction, whole environments can be modeled as a single continuum. As such approaches are not subject of the present review, we refer the interested reader, for example, to refs 278 and 279.

2.3. Response Properties

Once an energy calculation is completed by one of the CompChem methods above, many other interesting molecular properties can be calculated. Most of these properties can be obtained as the response of the energy to a perturbation, for example, changes in nuclear coordinates $\mathbf{R}$, external electric (ϵ) or magnetic ($\mathbf{B}$) fields or the nuclear magnetic moments $\{\mathbf{I}_i\}$. Given an expression for the energy, which depends on the above quantities, so-called response properties can be computed via the corresponding partial derivatives of the energy. A general response property $\mathbf{\Pi}$ then takes the form

$$\mathbf{\Pi}(n_{\mathbf{R}}, n_{\epsilon}, n_{\mathbf{B}}, n_{\mathbf{I}_i}) = \frac{\partial^{n_{\mathbf{R}}+n_{\epsilon}+n_{\mathbf{B}}+n_{\mathbf{I}_i}}E(\mathbf{R}, \epsilon, \mathbf{B}, \mathbf{I}_i)}{\partial\mathbf{R}^{n_{\mathbf{R}}}\partial\epsilon^{n_{\epsilon}}\partial\mathbf{B}^{n_{\mathbf{B}}}\partial\mathbf{I}_i^{n_{\mathbf{I}_i}}} \quad (16)$$

where the ns indicate the n-th order partial derivative with respect to the quantity in the subscript.[102]

A common response property is nuclear forces $\mathbf{F} = -\mathbf{\Pi}$ (1, 0, 0, 0) that are the negative first derivatives of the energy with respect to the nuclear positions. Such calculations allow a plethora of different geometry optimization schemes for chemical structures on the PES. Hessian calculations corresponding to the second derivative of energy with respect to nuclear positions are necessary to confirm the location of first-order saddle points on the PES and identify normal modes and their frequencies for vibrational partition functions that are useful for modeling temperature dependencies based on statistical thermodynamics. Hessian calculations are computationally costly, since they normally involve calculations based on finite differences methods involving many nuclear force calculations. Many methods have been developed to allow CompChem algorithms to sample minimum energy regions of the PES[280−284] or precisely locate points of interest.[285,286] Historically, many of these techniques have relied on approximate or full Hessian calculations,[287] but other

approaches, such as the nudged-elastic band[288,289] and string[290−292] methods, are popular alternatives that do not require a Hessian calculation. There have also been efforts using different forms of ML to accelerate procedures or overcome long-standing challenges in efficient sampling of and optimization on the PES.[293−298]

The general expression above can provide a wealth of other quantities, some of which are relevant for molecular spectroscopy or provide a direct connection to experiment (see Table 3). Infrared spectra can be simulated based on dipole moments

Table 3. Response Properties of the Potential Energy

n_{R}	n_{ϵ}	n_{B}	n_{I}	property
0	0	0	0	energy
1	0	0	0	forces
2	0	0	0	Hessian (harmonic frequencies)
0	1	0	0	dipole moment (IR)
1	1	0	0	infrared absorption intensities (IR)
0	2	0	0	polarizability (Raman)
0	0	1	1	nuclear magnetic shielding (NMR)
0	0	0	2	nuclear spin−spin coupling (NMR)
0	1	1	0	optical rotation (circular dichroism)

$\boldsymbol{\mu} = -\mathbf{\Pi}$ (0, 1, 0, 0), while molecular polariziabilities $\boldsymbol{\alpha} = -\mathbf{\Pi}$ (0, 2, 0, 0) offer access to polarized and depolarized Raman spectra. Nuclear magnetic shielding tensors $\sigma = \Pi$ (0,0,1,1) are a central response property of a magnetic field. These allow the computation of chemical shifts recorded in nuclear magnetic resonance (NMR) spectroscopy via their trace $\sigma_i = \frac{1}{3}\mathrm{Tr}[\boldsymbol{\sigma}_i]$. The beauty of this formalism lies in the fact that a single energy calculation method provides access to a wide range of quantum chemical properties in a highly systematic manner. A large number of modern MLPs use the response of the potential energy with respect to nuclear positions to obtain energy conserving forces. However, far fewer applications model perturbations with respect to electric and magnetic fields. Ref 299 extends the descriptor used in the Faber−Christensen−Huang−Lilienfeld (FCHL) Kernel by adding an explicit field dependent term that makes it possible to predict dipole moments across chemical compound space. Ref 300 introduces a general neural network (NN) framework to model interactions of a system with vector fields, which was then used to predict dipole moments, polarizabilities and nuclear magnetic shielding tensors as response properties.

2.4. Solvation Models

An important aspect of CompChem is molecular descriptions from within a solution environment. Simulating a dynamical environment composed of many surrounding molecules is usually not feasible with electronic-structure methods. To circumvent this problem, solvation modeling schemes have been devised (see refs 301−306 for discussions on this topic).

The most popular approaches are so-called polarizable continuum solvent models (PCM).[279] They model the electrostatic interaction of a solute molecule with its environment by representing the charge distribution of the solvent molecules as a continuous electric field, the reaction field. This dielectric continuum can be interpreted as a thermally averaged representation of the environment and is typically assigned a constant permittivity depending on the particular solvent to be modeled ($\varepsilon = 80.4$ for water). The solute is placed inside a cavity embedded in this continuum.

The charge distribution of the molecule then polarizes the continuous medium, which in turn acts back on the molecule. To compute the electrostatic interactions arising from this mutual polarization with electronic structure theory, a self-consistent scheme is employed. After constructing a suitable molecular cavity, a Poisson problem of the following form is solved:

$$-\nabla[\epsilon(\mathbf{r})\nabla V(\mathbf{r})] = 4\pi\rho_m(\mathbf{r}) \tag{17}$$

Here, $\rho_m(\mathbf{r})$ is the charge distribution of the solute and $\epsilon(\mathbf{r})$ is the position dependent permittivity, which usually is set to one within the cavity and the ϵ of the solvent on the outside. $V(\mathbf{r})$ is the electrostatic potential composed of the two terms

$$V(\mathbf{r}) = V_m(\mathbf{r}) + V_s(\mathbf{r}) \tag{18}$$

where $V_m(\mathbf{r})$ is the solute potential and $V_s(\mathbf{r})$ is the apparent potential due to the surface charge distribution $\sigma(s)$

$$V_s(\mathbf{r}) = \int_\Gamma \frac{\sigma(\mathbf{s})}{|\mathbf{r} - \mathbf{s}|} d\mathbf{s} \tag{19}$$

Γ indicates the surface of the cavity. Eq 17 is solved numerically to obtain the surface charge distribution $\sigma(\mathbf{s})$. Once $\sigma(\mathbf{s})$ has been determined in this fashion, the potential is computed according to eq 19 and used to construct an effective Hamiltonian of the form

$$\hat{H}_{eff} = \hat{H} + V_s(\mathbf{r}) \tag{20}$$

where $\hat{H}$ is the vacuum Hamiltonian. These equations are then solved self-consistently in a Roothan−Hall or KS approach, yielding the electrostatic solvent−solute interaction energy. This scheme is also called the self-consistent reaction field approach (SCRF).

Continuum models differ in how the cavities are constructed and how eq 17 is solved to obtain the surface charge distribution. Variants include the original PCM model, also referred to as dielectric PCM (D-PCM),[307] the integral equation formulation of PCM (IEFPCM),[308] SMD,[309] conductor PCM (C-PCM),[310] or the conductor-like screening model (COSMO).[311] The latter two approaches replace the dielectric medium by a perfect conductor to allow for a particularly efficient computation of $\sigma(s)$. PCMs can be further extended with statistical thermodynamics treatments to account for solutes having different size and concentration effects, and this leads to models such as COSMO-RS.[312]

A drawback of most PCM-like approaches is that they neglect local solvent structures. Thus, they cannot reliably account for situations where explicit solvent interactions are important, for example, when for stabilizing specific sites for a transition state through hydrogen bonding.[301] Furthermore, while implicit models might be parametrized to fit bulk-like properties of mixed or ionic solvents (e.g., ref 313.), the complex local solvent environment presented by these systems are treatable by other means. For mixed solvent systems a range of hybrid schemes such as COSMO-RS,[305] reference interaction site models (RISMs)[314,315] or QM/MM[316−318] approaches have been developed. As an in-depth discussion of these alternative schemes exceeds the scope of this Review, we instead refer to other references.[319,320]

ML models are becoming used to describe solvent effects. Ref 300 introduces a continuum ML model based on a reaction field that can predict energies and response properties for continuum solvents, it can extrapolate to solvents not seen during training, and it can be extended to operate in a QM/MM fashion to account for explicit solvents effects in a Claisen rearrangement reaction. Ref 321 implemented automatable calculation schemes and unsupervised ML to allow predictions of single ion solvation energies for monovalent and divalent cations and anions based on physically rigorous quasi-chemical theory.[322,323] Ref 324 used convolutional NNs and MD simulations to carry out high-throughput screening of mixed solvent systems. Ref 325 implemented efficient ways to carry out ML-based QM/MM MD simulations.

2.5. Insightful Predictions for Molecular and Material Properties

By solving for electronic structures, by whatever means is appropriate, one obtains molecular energies and energy spectrum (typically corresponding to *quasiparticles* given by KS or HF orbitals). From these, one can then compute molecular or material properties that arise from quantum mechanical and statistical operators, for example, thermodynamic energies, response properties, highest and lowest occupied molecular orbital energies, and band gaps, among other properties. Many properties are defined by the characters of the orbitals, and having knowledge of these should always be helpful and aid in deriving useful insight into designing molecules and materials for a particular function. Furthermore, one is often interested in how these molecules behave over time (i.e., the dynamics given some statistical ensemble that depends on temperature, pressure, etc) over all possible degrees of freedom. By understanding how energies and forces change over time, one can predict thermal and pressure dependencies as well as spectroscopic properties for advanced knowledge that builds toward insightful predictions.

Molecular and materials chemistry is vastly complex and variable, and one often faces a question of whether to span wider chemical spaces versus take deeper explorations of a specific phenomenon. A key problem is that even after the effort of either approach, it is also not as clear how information for one system might be related to another to provide more knowledge. For instance, one may decide to calculate all possible properties of ethanol with a CompChem method, but understanding how any calculated property would be correlated to an analogous property of isopropanol is still usually difficult to do. There is great interest in understanding chemical and materials space through applications of quantitative structure activity/property relationships,[326,327] cheminformatics,[328] conceptual DFT,[329] and alchemical perturbation DFT.[330] All these applications benefit from greater access to CompChem data, and all have promise as being interfaced with ML for transformative applications to catalyze wisdom and impact.

3. MACHINE LEARNING TUTORIAL AND INTERSECTIONS WITH CHEMISTRY

ML has had a dramatic impact on many aspects of our daily lives and has arguably become one of the most far-reaching technologies of our era. It is hard to overstate its importance in solving long-standing computer science challenges, such as image classification[331−334] or natural language processing,[335−339] tasks that require knowledge that is hard to capture in a traditional computer program.[340−342] Previous classical artificial intelligence (AI) approaches relied on very large sets of rules and heuristics, but these were unable to cover the full scope of these complex problems. Over the past decade,

BIAS-VARIANCE TRADEOFF
What is a good ML model?
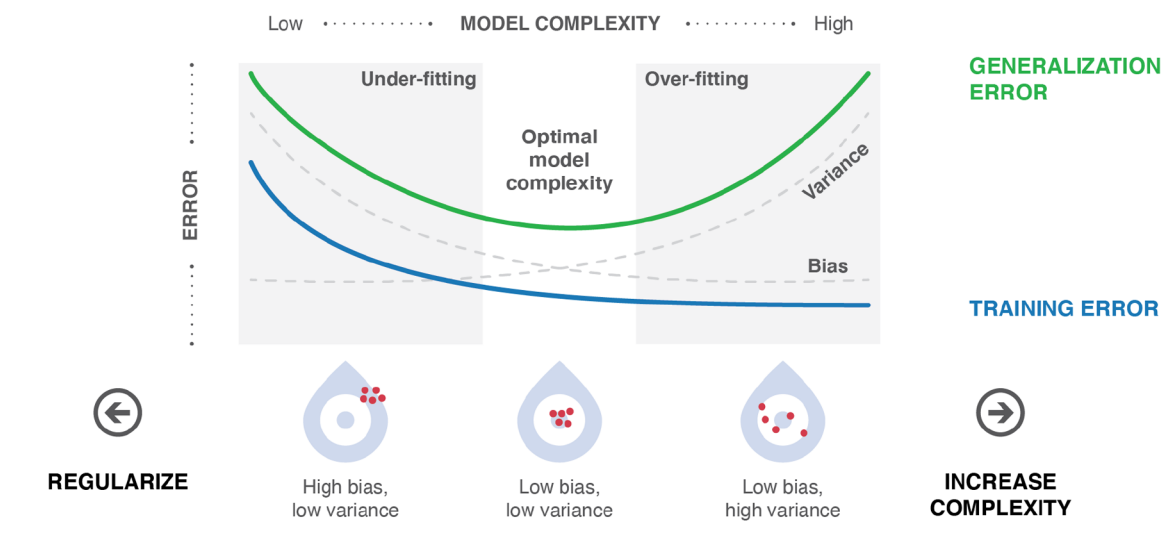
CROSS-VALIDATION
How to find a good ML model?

Figure 6. Supervised learning algorithms have to balance two sources of error during training: the bias and variance of the mo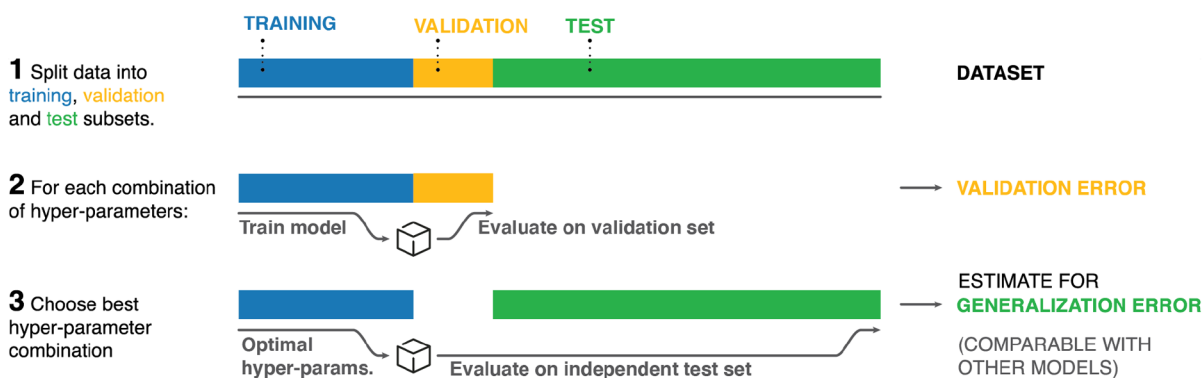del. A highly biased model is based on flawed assumptions about the problem at hand (under-fitting). Conversely, a high variance causes a model to follow small variations in the data too closely, therefore making it susceptible to picking up random noise (overfitting). The optimal bias-variance trade-off minimizes the generalization error of the model, for example, how well it performs on unknown data. It can be estimated with cross-validation techniques.

advances in ML algorithms and computer technology made it possible to learn underlying regularities and relevant patterns from massive data sets that enable automatic constructions of powerful models that can sometimes even outperform humans at those tasks.

This development inspired researchers to approach challenges in science with the same tools, driven by the hope that ML would revolutionize their respective fields in a similar way. Here, we give an overview of these developments in chemistry and physics to serve as an orientation for newcomers to ML. We will first explain what tasks ML is good at and when it might not be the best solution to a problem. We will start by introducing the field of ML in general terms and dissect its strengths and weaknesses.

3.1. What is ML?

In the most general sense, ML algorithms estimate functional relationships without being given any explicit instructions of how to analyze or draw conclusions from the data. Learning algorithms can recover mappings between a set of inputs and corresponding outputs or just from the inputs alone. Without output labels, the algorithm is left on its own to discover structure in the data.

Universal approximators[343,344] are commonly used for that purpose. These reconstruct any function that fulfills a few basic properties, such as continuity and smoothness, as long as enough data is available. Smoothness is a crucial ingredient that makes a function learnable, because it implies that neighboring points are correlated in similar ways. That property means that one can draw successful conclusions about unknown points as long as they are close to the training data (coming from the same underlying probability distribution).[341] In contrast, completely random processes in the above sense allow no predictions.

An association that immediately springs to mind is traditional regression analysis, but ML goes a step further.

Regression analyses aim to reconstruct the function that goes through a set of known data points with the lowest error, but ML techniques aim to identify functions to predict *interpolations between data points* and thus minimize the prediction error for new data points that might later appear.[345] Those contrasting objectives are mirrored in the different optimization targets. In traditional regression, the optimization task

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left[\sum_{i}^{M} \mathcal{L}(f(\mathbf{x}_i), y_i) \right] \quad (21)$$

only measures the fit to the data, but learning algorithms typically aim to find models $\hat{f}$ that satisfy

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left[\sum_{i}^{M} \mathcal{L}(f(\mathbf{x}_i), y_i) + \|\Gamma\Theta\|^2 \right] \quad (22)$$

Both optimization targets reward a close fit, often using the squared loss $\mathcal{L}(\hat{f}(\mathbf{x}), \mathbf{y}) = (\hat{f}(\mathbf{x}) - \mathbf{y})^2$. However, the key difference is an additional *regularization* term in eq 22, which influences the selection of candidate models by introducing additional properties that promote generalization. To understand why this is necessary, it is helpful to consider that eq 22 is only a proxy for the optimization problem

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left[\int \mathcal{L}(f(\mathbf{x}_i), y_i) \mathrm{d}p(\mathbf{x}, \mathbf{y}) \right] \quad (23)$$

that we would actually like to solve. In an ideal world, we would minimize the loss function over the *complete* distribution of inputs and labels $p(\mathbf{x}, \mathbf{y})$. However, this is obviously impossible in practice, so we apply the principle of Occam's razor that presumes that simpler (parsimonious) hypotheses are more likely to be correct. With this additional consideration we hope to be able to recover a reasonably general model, despite only having seen a finite training set. A common way to favor simpler models is via an additional term in the cost function, which is what $\|\Gamma\Theta\|^2$ in eq 22 expresses. Here, Γ is a matrix that defines "simplicity" with regard to the model parameters Θ. Usually, $\Gamma = \lambda I$ (where I is the identity matrix and $\lambda > 0$) is chosen to simply favor a small L^2-norm on the parameters, such that the solution does not rely on individual input features too strongly. This particular approach is called Tikhonov regularization,[346−348] but other regularization techniques also exist.[349,350]

A model that is heavily regularized (i.e., using a large λ) will eventually become *biased* in that it is too simplistic to fit the data well. In contrast, a lack of regularization might yield an overly complex model with high *variance*. Such an "overly fit" model will follow the data exactly to the point that it also models the noise components and consequently fails to generalize (see Figure 6). Finding the appropriate amount of regularization λ to manage under- and overfitting is known as attaining a good *bias-variance trade-off*.[351] We will introduce a process called *cross-validation* to address this challenge further below (see section 3.4.3).

3.1.1. What Does ML Do Well? *Implicit Knowledge from Data.* ML algorithms can infer functional relationships from data in a statistically rigorous way without detailed knowledge about the problem at hand. ML thus captures implicit knowledge from a data set—even aspects where CPI might not be available. Traditional modeling approaches, such as the

classical force fields discussed in section 2.2.6, rely on preconceived notions about the PES that is being modeled and, thus, the way the physical system behaves. In contrast, ML algorithms start from a loss function and a much more general model class. Within the limits permitted by the noise inherent to the data, generalization can be improved to arbitrary accuracy given increasingly larger informative training data sets. This process allows us to explore a problem even before there is a reasonably full understanding. An ML predictor can serve as a starting point for theory building and be regarded as a versatile tool in the modeling loop: building predictive models, improving them, enriching them by formal insight, and improving further and ultimately extracting a formal understanding. More and more research efforts start to combine data-driven learning algorithms with rigorous scientific or engineering theory to yield novel insights and applications.[10,16,352]

Redundancy in CompChem Calculations. For a quantum chemical property for compounds in a data set, CompChem calculations need to be repeated independently for each input, even if they are very similar. No formally rigorous method exists to exploit redundancies in the calculations in such a scenario. The empiricism of learning algorithms however does provide a pathway to extract information based on compound structure similarity. A data-driven angle allows one to ask questions in new ways that give rise to new perspectives on established problems. For example, unsupervised algorithms like clustering or projection methods group objects according to latent structural patterns and provide insights that would remain hidden when only looking at individual compounds.

3.1.2. What Does ML Do Poorly? *Lack of Generality and Precision.* Some difficult problems in chemistry and physics can be solved accurately with CompChem, but doing so would require significant resources. For example, enumerating all pairwise interactions in a many-body system will inevitably scale quadratically, and there is no obvious path around this. One might ask if empirical approaches can address such fundamental problems more efficiently, but this is unfortunately not possible since ML is more suited for finding solutions in general function spaces rather than in deterministic algorithms where constraints guide the solution process. However, if we were not as interested in finding a full solution but rather some aspect of it, the stochastic nature of ML can be beneficial. For instance, a traditional ML approach might not be the best tool for explicitly calculating the Schrödinger equation, but it might be a far more useful tool for developing a force field that returns the energy of a system without the need for a cumbersome wavefunction and a self-consistent algorithm. As an example, Hermann et al.[105] used deep NNs to show how ML methods may be suitable for overcoming challenges faced by traditional CompChem approaches.

Reliance on High-Quality Data. ML algorithms require a large amount of high quality data, and it is hard to decide a priori when a data set is sufficient. Sometimes, a data set may be large, but it does not adequately sample all the relevant systems one intends to model. For example, an MD simulation might generate many thousands of molecular confirmations used to train an ML force field, but perhaps that sampling only occurred in a local region of the PES. In this case, the ML force field would be effective at modeling regions of the PES it was trained to but useless in other regions until more data and broader sampling occurred. This feature is general to all

empirical models that are generally limited in their extrapolation abilities.

Inability to Derive High-Level Concepts. Standard ML algorithms cannot conceptualize knowledge from a data set. Two main reasons are the nonlinearity and excessive parametric complexity of most models that allow many equally viable solutions for the same problem.[353,354] It can be hard to gain insight into the modeled relationship because it is not based on a small set of simple rules. Techniques have emerged to make ML models interpretable (explainable AI—XAI[355]). While helpful, drawing scientific insight clearly still requires human expertise.[352,355−361] Furthermore, the path from an ML model back to a physical set of equations is being explored, but it is far from being fully established automatically.[362−368]

Prone to Artifacts. Despite following the rules of best practice, ML algorithms can give unexpected and undesired results. Instead of extracting meaningful relationships, they may occasionally exploit nuisance patterns within the underlying experimental design, like the model architecture, the loss function or artifacts in the data set. This results in a "clever Hans" predictor,[360] which technically manages the learning problem but uses a trivial solution that is only applicable within the narrow scope of the particular experimental setup at hand. The predictor will appear to be performing well, while actually harvesting the wrong information and, therefore, not allowing any generalization or transferable insights.

For example, a recently proposed random forest predictor for the success of Buchwald−Hartwig coupling reactions[369] was later revealed to give almost the same performance when the original inputs were replaced by Gaussian noise.[370,371] This finding strongly suggested that the ML algorithm exploited some hidden underlying structure in the input data, irrespective of the chemical knowledge that was provided through the descriptor. Even though the model might appear quite useful, any conclusions that rely on the importance of the chemical features used in the model were thus rendered questionable at best. This example demonstrates that out-of-sample validation alone is often not sufficient to establish that a proposed model has indeed learned something meaningful. Therefore, the hypothesis described by the model must be challenged in extensive testing in practically relevant scenarios like actual physical simulations. In other words the ML model needs to lead to a better understanding of the modeling itself and the underlying chemistry.

3.2. Types of Learning

ML models are classified by the type of learning problem they solve. Consider for instance a data scientist who develops an ML model that can predict acidity constants (pK_a values) for any molecule. A researcher with knowledge of physical organic chemistry might be aware of the empirical Taft equation[29] that provides a linear free energy relationship between molecules on the basis of empirical parameters that account for a molecule's fundamental field, inductive, resonance, and steric effects (e.g., values related to Hammett ρ and σ values). There are several ways the data scientist might develop an ML model for this or another application. Examples mentioned here include supervised, unsupervised, and reinforcement learning.

3.2.1. Supervised Learning. *Supervised learning* addresses learning problems where the ML model $\hat{f} : \mathcal{X} \xrightarrow{\text{ML}} \mathcal{Y}$ connects a set of known inputs $\mathcal{X}$ and outputs $\mathcal{Y}$, either to perform a regression or classification task. While the former maps onto a continuous space (e.g., energy, polarizability), the latter

outputs a categorical value (e.g., acid or base; metal or insulator) for each data point.

Using the pK_a predictor example, a supervised learning algorithm could be trained to correlate recognizable chemical patterns or structures to experimentally known pK_a values. The goal would be to deduce the relationship between these inputs and outputs, such that the model is able to generalize beyond the known training set. A standard universal approximator has to accomplish this learning task without any preconceived notion about the problem at hand and will, therefore, likely require many examples before it can make accurate predictions. Recently, a lot of research is being carried out that investigates ways to incorporate high-level concepts into the learning algorithm in the form of prior knowledge.[207,372] In this vein, one could take into account chemically relevant parameters, such as Hammett constants so that the parametrized ML model incorporates the modified Hammett or Taft equation. An example of a classification problem in materials science is the categorization of materials, where identifying characteristics of the electronic structure can be used to distinguish between insulators and metals.[373]

3.2.2. Unsupervised Learning. *Unsupervised learning* describes problems in which only the inputs are known, with no corresponding labels. In this setting, the goal is to recover some of the underlying structure of the data to gain a higher-level understanding. Unsupervised learning problems are not as rigorously defined as supervised problems in the sense that there can be multiple correct answers, depending on the model and objective function that is applied.

For example, one might be interested in separating conformers of a molecule from an MD trajectory, given exclusively the positions of the atoms. A *clustering* algorithm (like the k-means algorithm) could identify those conformers by grouping the data based on common patterns.[374,375] Alternatively, a *projection* technique could reveal a low-dimensional representation of the data set.[376] Often data is represented in high dimension, despite being intrinsically low-dimensional. With the right projection technique, it is possible to retain the meaningful properties in a representation with fewer degrees of freedom. A conceptually simple embedding method is principal component analysis (PCA) in which the relationship that is sought to be preserved is the scalar product between the data points.[340] There are many other linear and nonlinear projection methods, such as multidimensional scaling,[377] kernel PCA (KPCA),[378,379] *t*-distributed stochastic neighbor embedding (t-SNE),[380] sketch-map,[381] and the uniform manifold approximation and projection (UMAP).[382] Finally, anomaly detection is another extension of unsupervised learning, where 'outliers' to the available data can be discovered.[383] However, without knowing the labels (in this example, the potential energy associated with each geometry), there is no way to conclusively verify that the result is correct. The literature is gradually seeing more instances of unsupervised learning, particular to reveal important chemical properties to efficiently explore chemical/materials spaces.

3.2.3. Reinforcement Learning. *Reinforcement learning* (RL) describes problems that combine aspects of supervised and unsupervised learning. RL problems often involve defining an *agent* within an environment that learns by receiving feedback in the form of punishments and rewards. The progress of the agent is characterized by a combination of explorative activity and exploitation of already gathered knowledge.[384] For chemistry applications, RL techniques are

being increasingly used for finding molecules with desired properties in large chemical spaces.[10]

3.3. Universal Approximators

Universal approximators have their origins in the 1960s, where the hope was to construct "learning machines" that have similar capabilities as the human brain. An early mathematical model of a single simplified neuron emerged that was called a *perceptron* (eq 24).[385,386]

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} w_i x_i - b\right) \tag{24}$$

Here, $\mathbf{x}$ denotes the N-dimensional input to the perceptron. It has $N + 1$ parameters consisting of w_i (so-called weights) and a single b (a so-called threshold) that are adapted to the data. This adaption process is typically called "learning" (vide infra), and it amounts to minimizing a predefined loss function.

In the 1960s, this simple NN had very limited use, as it was only able to model a linear separating hyperplane. Even simple nonlinear functions like the XOR were out of reach.[387] Thus, excitement waned but then reappeared two decades later with the emergence of novel models consisting of more neurons and their arrangement in multilayer NN structures[388] (see eq 25). Recent algorithmic and hardware advances now allow deep and increasingly complex architectures.[1,2]

$$f_k(\mathbf{x}) = g\left(\sum_{j=1}^{H} w_{kj} g\left[\sum_{i=1}^{N} w_{ji} x_i - b\right]\right) \tag{25}$$

In eq 25, $g(\cdot)$ denotes an activation function that is a nonlinear transformation that allows complex mappings between input and output. As with the perceptron, the parameters of multilayer NNs can be learned efficiently using iterative algorithms that compute the gradient of the loss-function using the so-called back-propagation (BP) algorithm.[388−390] In the late 1980s, artificial NNs were then proven to be universal approximators of smooth nonlinear functions,[343,391,392] and so they gained broad interest even outside the ML community that then was still relatively small.

In 1995, a novel technique called Support Vector Machine (SVM)[345,393] and kernel-based learning were then proposed,[379,394−396] which came with some useful theoretical guarantees. SVMs implement a nonlinear predictor:

$$f(\mathbf{x}) = \sum_{j=1}^{N} y_j \alpha_j K(\mathbf{x}_j, \mathbf{x}) - b \tag{26}$$

where K is the so-called kernel. The kernel implicitly defines an inner product in some feature space and thus avoids an explicit mapping of the inputs. This "kernel trick"[397] makes it possible to introduce nonlinearity into any learning algorithm that can be expressed in terms of inner products of the input.[379] It has since been applied to many other algorithms beyond SVMs,[394] such as Gaussian Processes (GP),[348] PCA,[378,379] and independent component analysis (ICA).[398]

The most effective kernels are tailored to the specific learning task at hand, but there are many generic choices, such as the polynomial kernel $K(\mathbf{x}_j, \mathbf{x}) = (\langle \mathbf{x}_j, \mathbf{x}\rangle - b)^d$, which describes inner products between degree d polynomials. Another popular choice is the Gaussian kernel $K(\mathbf{x}_j, \mathbf{x}) = \exp(-(\mathbf{x}_j - \mathbf{x})^2/(2\sigma^2))$. It is one of the most versatile kernels

because it only imposes smoothness assumptions on the solution depending on the width parameter σ.[347,395]

As seen in eq 26, an SVM can also be understood as a shallow NN with a fixed set of nonlinearities. In other words, the kernel explicitly defines a similarity metric to compare data points, whereas NNs have more freedom to shape this transformation during training because they nest parametrizable nonlinear transformations on multiple scales. This difference gives both techniques unique strengths and drawbacks. Despite that, there exists a duality between both approaches that allows NNs to be translated into kernel machines and analyzed more formally (see refs 399−401).

In the context of CompChem, both NNs and kernel-based methods are the most used ML approaches. Simpler learners, such as nearest neighbor models or decision trees can still be surprisingly effective. Those have also been successfully used to solve a wide spectrum of problems including drug design, chemical synthesis planning, and crystal structure classification.[402−407]

3.4. ML Workflow

In the following, we summarize the overall ML process, starting from a data set all the way to a trained and tested model. The ML workflow typically includes the following stages:

1 Gathering and preparing the data

2 Choosing a representation

3 Training the model
    3a Train model candidates
    3b Evaluate model accuracy
    3c Tune hyperparameters

4 Testing the model out of sample

Note, that the progression to a good ML model is not necessarily linear and some steps (except the out of sample test) may require reiteration as we learn about the problem at hand.

3.4.1. Data Sets. On a fundamental level, ML models could be simply regarded as sophisticated parametrizations of data sets. While the architectural details of the model matter, the reference data set forms the backbone that ultimately determines the model's effectiveness. If the data set is not representative of the problem at hand, the model will be incomplete and behave unpredictably in situations that have been improperly captured. The same applies to any other shortcomings of the data set, such as biases or noise artifacts that will also be reflected in the model. Some of these data set issues are likely to remain unnoticed when following the standard model selection protocol since training and test data sets are usually sampled from the same distribution. If the sampling method is too narrow, errors seen during the cross-validation procedure may appear to be encouragingly small, but the ML model will fail catastrophically when applied to a real problem. If the training and test sets come from different distributions, then techniques to compensate this covariate shift can be used.[408,409]

Robust models can generally only be constructed from comprehensive data sets, but it is possible to incorporate certain patterns into models to make them more data-efficient. Prior scientific knowledge or intuition about specific problems can be used to reduce the function space from which an ML algorithm has to select a solution. If some of the unphysical solutions are removed a priori, less data are necessary to identify a good model. This is why NNs and kernel methods,

despite both being broad universal function classes, bring different scaling behaviors. The choice of the kernel function provides a direct way to include prior knowledge such as invariances, symmetries, or conservation laws, whereas NNs are typically used if the learning problem cannot be characterized as specifically.[207,372,410] In general, without prior knowledge, NNs often require larger data sets to produce the same accuracy as well-constrained kernel methods that embody problem knowledge. This consideration is particularly important if the data is expensive, for example, if it comes from high quality experiments or expensive computations.

3.4.2. Descriptors. To apply ML, the data set needs to be encoded into a numerical representation (i.e., features/descriptors) that allows the learning algorithm to extract meaningful patterns and regularities.[411−419] This is particularly challenging for unstructured data like molecular graphs that have well-defined invariable or equivariable characteristics that are hard to capture in a vectorial representation. For example, atoms of the same type are indistinguishable from each other, but it is hard to represent them without imposing some kind of order (which inevitably assigns an identity to each atom). Furthermore, physical systems can be translated and rotated in space without affecting many attributes. Only a representation that is adapted to those transformations can solve the learning problem efficiently.

It turned out to be a major challenge to reconcile all invariances of molecular systems in a descriptor without sacrificing its uniqueness or computability. Some representations cannot avoid collisions, where multiple geometries map onto the same representation. Others are unique, but prohibitively expensive to generate. Many solutions to this problem have been proposed, based on general strategies such as invariant integration,[207] parameter sharing,[352,421−423] density representations,[276] or finger printing techniques.[424−433] Alternatively, an NN model infers the representation from data.[352,424,434,435] To date, none of the proposed approaches are without compromise, which is why the optimal choice of descriptor depends on the learning task at hand.

3.4.3. Training. The training process is the key step that ties together the data set and model architecture. Through the choice of the model architecture, we implicitly define a function space of possible solutions, which is then conditioned on the training data set by selecting suitable parameters. This optimization task is guided by a loss function that encodes our two somewhat opposing objectives: (1) achieving a good fit to the data, while (2) keeping the parametrization general enough such that the trained model becomes applicable to data that is not covered in the training set (see the two terms in eq 22). Satisfying the latter objectives involves a process called *model selection* in which a suitable model is chosen from a set of variants that have been trained with exclusive focus on the first objective. Depending on the model architecture, more or less sophisticated optimization algorithms can be applied to train the set of model candidates.

Kernel-based learning algorithms are typically linear in their parameters $\vec{\alpha}$ (see eq 26). Coupled with a quadratic loss function, $\mathcal{L}(\hat{f}(\mathbf{x}), y) = (\hat{f}(\mathbf{x}) - y)^2$, they yield a convex optimization problem. Convex problems can be solved quickly and reliably due to only having a single solution that is guaranteed to be globally optimal. This solution can be found algebraically by taking the derivative of the loss function and

setting it to zero. For example, kernel ridge regression (KRR) and GPs then yield a linear system of the form

$$\nabla_{\boldsymbol{\alpha}} \mathcal{L}(\hat{f}(\mathbf{x}), y) = (\mathbf{K} + \lambda \mathrm{I})\boldsymbol{\alpha} - y = 0 \qquad (27)$$

which is typically solved in a numerically robust way by factorizing the kernel matrix **K**. There exist a broad spectrum of matrix factorization algorithms, such as the *Cholesky decomposition*, that exploit the symmetry and positive definiteness properties of kernel matrices.[436−440] Factorization approaches are, however, only feasible if enough memory is available to store the matrix factors, and this can be a limitation for large-scale problems. In that case, numerical optimization algorithms provide an alternative: they take a multistep approach to solve the optimization problem iteratively by following the gradient:

$$\boldsymbol{\alpha}^t = \boldsymbol{\alpha}^{t-1} - \gamma \underbrace{\nabla_{\boldsymbol{\alpha}} \mathcal{L}(\hat{f}(\mathbf{x}), y)}_{\text{e.g.,} (\mathbf{K}+\lambda \mathrm{I})\boldsymbol{\alpha}^{t-1}-y} \qquad (28)$$

where γ is the step size (or learning rate). Iterative solvers follow the gradient of the loss function until it vanishes at a minimum, which is much less computationally demanding per step, because it only requires the evaluation of the model f. In particular, kernel models can be evaluated without storing **K** (see eq 28).

NNs are constructed by nesting nonlinear functions in multiple layers, which yields nonconvex optimization problems. Closed-form solutions similar to eq 27 do not exist, which means that NNs can only be trained iteratively, that is, analogous to eq 28. Several variants of this standard gradient descent algorithm exist including stochastic or mini-batch gradient descent, where only an n-sized portion of the training data $(\mathbf{x}, y)_{i:i+n}$ is considered in every step. Because of multiple local minima and saddle points on the loss surface, the global minimum is exponentially hard to obtain (since these algorithms usually converge to a local minimum). However, thanks to the strong modeling power of NNs, local solutions are usually good enough.[441]

Hyperparameters. In addition to the parameters that are determined when fitting an ML model to the data set (i.e., the node weights/biases or regression coefficients), many models contain so-called hyperparameters that need to be fixed before training. Two types of hyperparameters can be distinguished: ones that influence the model, such as the type of kernel or the NN architecture, and ones that affect the optimization algorithm, for example, the choice of regularization scheme or the aforementioned learning rate. Both tune a given model to the prior beliefs about the data set and thus play a significant role in model effectiveness. Hyperparameters can be used to gauge the generalization behavior of a model.

Hyperparameter spaces are often rather complex: certain parameters might need to be selected from unbounded value spaces, others could be restricted to integers or have interdependencies. This is why they are usually optimized using primitive exhaustive search schemes like grid or random searches in combination with educated guesses for suitable search ranges. Common gradient-based optimization methods typically cannot be applied for this task. Instead, the performance of a given set of hyperparameters is measured by evaluating the respective model on another training data set called the validation data set (see Figure 6). This process is also referred to as model selection.

Table 4. ML Descriptors Found in the Literature[a]

descriptors	comp. efficiency[b]	periodic[c]	unique	invariances[d]			global	smooth[e]
				T	R	P		
atom-centered symmetry functions (ASCF)[411]	Ⓑ 1,2,3-body terms, cutoff	√	X	√	√	√	X	√
smooth overlap of atomic positions (SOAP)[412]	Ⓑ density based, SO(3) rotational group integration	√	X	√	√	√	X	√
Coulomb matrix (CM)[413]	Ⓐ 1,2-body terms	X	√	√	√	X	√	√
sine matrix[414]	Ⓐ 1,2-body terms	√	√	√	√	X	√	√
Ewald sum matrix[414]	Ⓐ 1,2-body terms	√	√	√	√	X	√	√
bag of bonds (BoB)[415]	Ⓐ 1,2-body terms	X	X	√	√	○	√	X
Faber−Christensen−Huang−Lilienfeld (FCHL)[416]	Ⓒ 1,2,3-body terms	√	X	√	√	√	X	√
spectrum of London and Axilrod−Teller−Muto potential (SLATM)[417]	Ⓓ 1,2,3,4-body terms	√	X	√	√	√	X	√
many-body tensor representation (MBTR)[418]	Ⓒ 1,2,3-body terms	X	X	√	√	√	√	√
atomic cluster expansion[420]	Ⓐ 1,2-body terms	√	X	√	√	√	√	√
invariant many-body interaction descriptor (MBI)[460]	Ⓑ 1,2,3-body terms	X	X	√	√	√	X	√
neural network architectures								
deep potential—smooth edition (DeepPot-SE)[461,462]	Ⓑ 1,2,3-body terms, cutoff	√	X	√	√	√	X	√
MPNN, SchNet[352,434]	Ⓐ/Ⓑ 1,2-body terms, hierarchical	√	X	√	√	√	X	√
Cormorant[463]	Ⓑ 1,2-body terms, hierarchical	X	X	√	√	√	X	√
tensor field networks[464]	Ⓑ 1,2-body terms	√	X	√	√	√	X	√
similarity metrics								
root mean square deviation of atomic positions (RMSD)[454]	Ⓐ 1,2-body terms, input matching	X	X	○	○	X	√	X
overlap matrix[454]	Ⓐ 1,2-body terms, input matching	X	X	√	√	√	√	X
REMatch[459]	Ⓒ 1,2-body terms, input matching	X	X	√	√	√	√	X
sGDML[207]	Ⓐ 1,2-body terms	√	√	√	√	○[f]	√	√

[a]"√" = satisfies condition; "○" = partially satisfies condition; "X" = does not satisfy condition. [b]Computational efficiency ranks with grades Ⓐ−Ⓓ in descending order. The efficiency class reflects the extent that the descriptor requires expensive operations (e.g., a hierarchical processing or matching of inputs). [c]Descriptor has been used within periodic boundary conditions. [d]"T" = translational; "R" = rotational; "P" = permutational. [e]In this context, a descriptor is referred to as smooth if its first derivative with respect to nuclear positions is continuous. [f]Only invariant to permutations represented in the training data.

Model Selection. Cross-validation or out-of-sample testing is a technique to assess how a trained ML model will generalize to previously unseen data.[340,395] For a reasonably complex model, it is typically not challenging to generate the right responses for the data known from the training set. This is why the training error is not indicative of how the model will fulfill its ultimate purpose of predicting responses for new inputs. Alas, since the probability distribution of the data is typically unknown, it is not possible to determine this so-called *generalization error* exactly. Instead, this error is often estimated using an independent test subset that is held back and later passed through the trained model to compare its responses to the known test labels. If the model suffers from overfitting on the training data, this test will yield large errors. It is important to remember not to tweak any parameters in response to these test results, as this will skew this assessment of the model performance and will lead to overfitting on the test set.[442]

Besides cross-validation, there are alternative ways to estimate the generalization error, for example via maximization of the marginal likelihood in Bayesian inference.[443−445] Some well-defined learning scenarios even allow the computation of rigorous upper bounds for the generalization error.[345,446−448]

4. APPLICATIONS OF MACHINE LEARNING TO CHEMICAL SYSTEMS

We now discuss ways that CompChem methods described in section 2 and ML methods in section 3 can be implemented as CompChem+ML approaches for insights into chemical systems. We often notice the lack of details about *why* an ML model is used and *how* it actually contributes to

worthwhile and scientific insights. Thus, we will summarize the underlying attributes of conventional CompChem+ML efforts and then explain why these attributes are important for specific applications.

To begin, consider molecules or materials in a data set, and any entry will be related to another based on an abstract concept of "similarity". While similarity is an application-dependent concept, it should go hand in hand with CPI. For instance, physical properties of chemical systems can be attributed to the structure or composition of the chemical fragments within those systems. Thus, if chemical structures and compositions of two entries in the database were similar, then their physical properties would also likely be similar.

For CompChem+ML using a supervised algorithm, a CompChem prediction might be made on a hypothetical system, pinpointed by an ML model that was trained to identify chemical fragments that correlate with labeled physical properties. This would be a direct exploitation of chemical similarity. Alternatively, for CompChem+ML using an unsupervised algorithm, the ML model would identify an underlying distribution or key features based on the similarity between pairs of entries in the data set without labels. This would be a more nuanced leveraging of chemical similarity. In both cases the accuracy, efficiency and reliability of the ML models depend strongly on how similarity is defined and measured.

In this section, we will first describe state-of-the-art descriptors and kernels for atomic systems that can be used to quantify the similarity between chemical systems. We will then explain the essential attributes of good atomic descriptors.

Lastly for this section, we will elucidate why and how specific combinations of these descriptors and ML algorithms are beginning to revolutionize the field of CompChem.

4.1. Representing Chemical Systems

In CompChem, molecules and materials are usually represented by the Cartesian coordinates and the chemical elements of all the atoms. Thus, the size of the vector representation containing the coordinates and charges will be $\{\mathcal{R}^{3N} \text{ and } \mathcal{Z}^N\}$, respectively, for a system of size N. Even though these atomic coordinates provide a complete description of the system, they are hardly ever used as the input of a ML model because this vector would introduce substantial superfluous redundancy. For instance, an ML model might treat two identical molecules that are rotated or translated as different molecules, and that in turn might cause the ML model to predict different physical properties for the two otherwise indistinguishable molecules. There are further difficulties when comparing molecules having different numbers of atoms. To work around these problems, atomic coordinates are usually converted into an appropriate representation ψ that is suitable for a particular task. Such conversions are useful because they allow the incorporation of physical invariances. Mathematically speaking, the representation fulfills

$$\psi(\{\mathcal{R}^{3N}, \mathcal{Z}^N\}) = \psi(S(\{\mathcal{R}^{3N}, \mathcal{Z}^N\})) \tag{29}$$

where S indicates a symmetry operation, for example, a rigid rotation about an axis C_i, an exchange of two identical atoms, or a translation of the whole system in the Cartesian space, etc. It can also be advantageous to adopt a coarse-grained representation of the system.[449,450] For example, dihedral angles of a peptide might be accounted for without the positions of the side-chains, positions of ions in a solution might be accounted for without the explicit coordinates of solvents, or just the center of mass for a water molecule might be accounted for in place of the full three-centered atomistic representation. The choice of these coarse-grained representations provides a way to incorporate prior knowledge of the data, or such representations can be learned from an unsupervised learning step.[451]

4.1.1. Descriptors. Atomistic systems can be represented in a myriad of ways. Some descriptions are designed to emphasize particular aspects of a system, while others aim to disambiguate similar chemical or physical principles across a wide range of molecules or materials. The set of desirable properties in a representation thus depends on the task at hand. All adhere to the aforementioned physical symmetries and invariances needed for chemical systems. Many have similar theoretical foundations that can be understood as the basis onto which the atomic density is projected,[452] and the connection between them has been summarized in a recent review.[453]

Table 4 gives a coarse characterization of popular representations.[276,411,412,415,417,418,454,455] To create this overview, we had to adopt a reductionist perspective, which inevitably hides the complexities involved in developing robust atomistic representations. Whether a representation satisfies a particular property can sometimes not be answered unequivocally. For example, is a descriptor unique if the ML model showed pathologically erroneous results? Should a symmetry be perfectly satisfied, even if it is a bad ML feature? We therefore stress that the table simply presents representations

and their attributes. A representation that satisfies more attributes is not necessarily better if it also lacks another important attribute. We kindly refer the reader to the respective original publications for more information.

The descriptors in Table 4 can be classified into two categories: global and atomic (i.e., not global). Traditional descriptors used in cheminformatics are global descriptors based on the covalent connectivity of atoms. These include simple valence counting and common neighbor analysis,[456] the presence or absence of predefined atomic fragments (e.g., the Morgan fingerprints[427]), pairwise distances between atoms (e.g., Coulomb Matrix,[413] Sine Matrix,[414] Ewald Sum Matrix,[414] Bag of Bonds (BoB)[415]), etc. Coulomb matrices have known problems because of lack of smoothness, but these are partly addressed by employing the Wasserstein norm, rather than Euclidean or Manhattan norms.[457] However, atomic descriptors[411,412,416−420,458] are generally more popular than the global ones in ML and CompChem. In atomic descriptors, a chemical system is described as a set of atomic environments, $\mathcal{X}_1, ... \mathcal{X}_i ... \mathcal{X}_N$, and each consists of the atoms (chemical species and position) within a sphere of radius r_{cut} centered at a specific atom i. One needs to combine the set of atomic descriptors of all environments to construct a descriptor for the entire atomic structure. The most straightforward way to do this is to average the atomic descriptors,

$$\Phi(A) = \frac{1}{N_A} \sum_{i \in A}^{N_A} \psi(\mathcal{X}_i) \tag{30}$$

where the sum runs over all N_A atoms i in structure A and $\mathcal{X}_i$ is the environment around atom i. When there are multiple chemical species, the descriptors for the local environments of different species can either be included in the single sum, or the averaging can be performed for the environments of each species separately and the species-specific averaged local descriptors can be concatenated. This can be done by considering the root mean square displacement (RMSD),[454] the best match between the environments of the two structures (best-match),[459] or by combining local descriptors using a regularized entropy match (RE-Match).[459]

4.1.2. Representing Local Environments. We will now describe the Smooth Overlap of Atomic Positions (SOAP) descriptors[412] since many other descriptors based on the atomic density are similar and differ mainly by how the density is projected onto basis functions.[420,452] To construct SOAP descriptors, one first considers an atomic environment $\mathcal{X}$ that contains only one atomic species, and a Gaussian function of width σ is then placed on each atom i in $\mathcal{X}$ to make an atomic density function:

$$\rho_{\mathcal{X}_i}(\mathbf{r}) = \sum_{i \in \mathcal{X}} \exp\left[-\frac{|\mathbf{r} - \mathbf{r}_i|^2}{2\sigma^2}\right] f_{cut}(|\mathbf{r}|) \tag{31}$$

Here, $\mathbf{r}$ denotes a point in Cartesian space, $\mathbf{r}_i$ is the position of atom i relative to the central atom of $\mathcal{X}$, and the cutoff function f_{cut} smoothly decays to zero beyond the cutoff radius r_{cut}. This density representation ensures invariance with respect to translations and permutations of atoms of the same species but not rotations. To obtain a rotationally invariant descriptor, one expands the density in a basis of spherical harmonics, $Y_{lm}(\hat{\mathbf{r}})$, and a set of orthogonal radial functions, $g_n(|\mathbf{r}|)$, as

```
asap gen_desc -f *.xyz --no-periodic soap -u minimal -pa
asap map -f ASAP-desc.xyz -dm '[*]' -ua --only_use_species 6 pca --no-scale
```
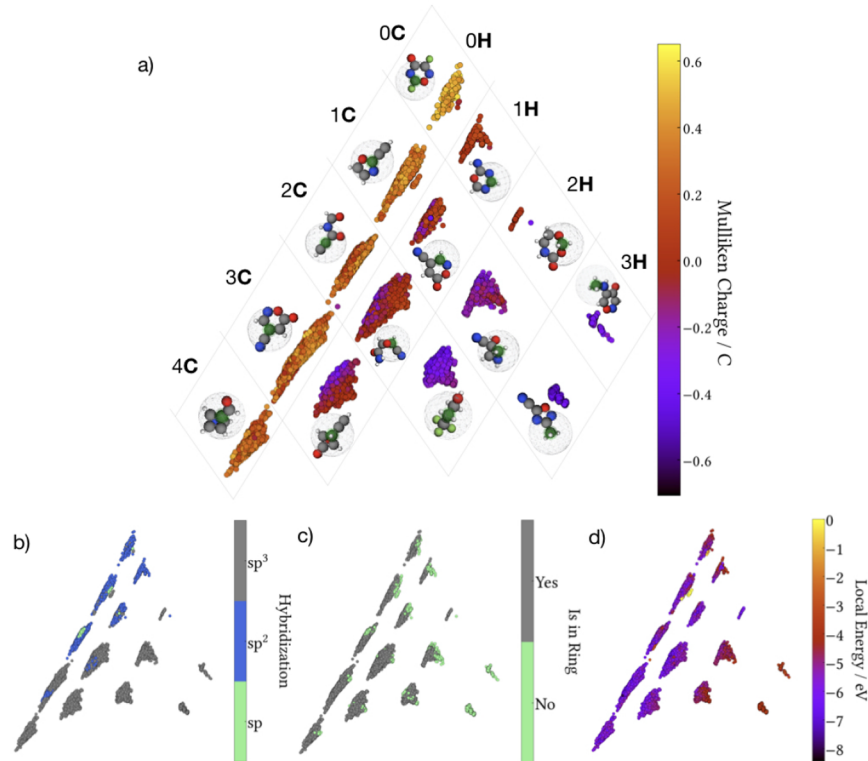

Figure 7. KPCA maps of carbon atom environments in the QM9 database. Maps are color-coded according to Mulliken charges (a), hybridization (b), whether atoms are found in rings (c) and according to local energies predicted by a machine learning potential (d). Reprinted with permission from ref 374. Copyright 2020 American Chemical Society.

$$\rho_\chi(\mathbf{r}) = \sum_{nlm} c_{nlm}\, g_n(|\mathbf{r}|)\, Y_{lm}(\hat{\mathbf{r}}) \tag{32}$$

to construct the *power spectrum* of the density using the expansion coefficients:

$$\psi_{nn'l}(\mathcal{X}) = \sqrt{\frac{8}{2l+1}} \sum_m (c_{nlm})^* c_{n'lm} \tag{33}$$

One then obtains a vector of descriptors $\psi = \{\psi_{nn'l}\}$ by considering all components $l \leq l_{max}$ and $n, n' \leq n_{max}$ that act as band limits that control the spatial resolution of the atomic density. The generalization to more than one chemical species is straightforward:[459] one constructs separate densities for each species α and then computes the power spectra $\psi_{nn'l}^{\alpha\alpha'}(\mathcal{X})$ for each pair of elements α and α', where the two species indices correspond to the c^* and c coefficients, respectively. The resulting vectors corresponding to each of the α and α' pairs are then concatenated to obtain the descriptor vector of the complete environment.

Atom-centered symmetry functions (ACSFs), or sometimes called Behler−Parrinello symmetry functions,[411] descriptors differ from SOAP in that they project the atomic densities over selected 2-body or 3-body symmetry functions. FCHL[417] descriptors follow similar principles while also considering the correlations between the atomic densities coming from different chemical species. The many-body tensor representation (MBTR)[418] approach involves taking the histograms of atom counts, inverse pairwise distances, and angles. Atomic

cluster expansion (ACE) descriptors[420] first express atomic densities using spherical harmonics and then generate invariant products by contracting the spherical harmonics with the Clebsch−Gordan coefficients.

Length-Scale Hyperparameters. Most atomic descriptors use length-scale hyperparameters specifically chosen for a given problem and system.[276,411,412,415,417,418,454,455] There are several ways to automate hyperparameter selections. Ref 374 introduced general heuristics for choosing the SOAP hyperparameters for a system with arbitrary chemical composition based on characteristic bond lengths. Ref 465 adopts the strategy to first generate a comprehensive set of ACSFs and then select a subset using the sparsification methods such as farthest point sampling (FPS)[466] and CUR matrix decomposition.[467]

Incompleteness of Atomic Descriptors. A structural descriptor is complete when there is no pair of configurations that produces the same descriptor.[468] For atomic descriptors, this means that different atomic environments—after considering the invariances of rotation, translation, and permutation of identical atoms—should adopt distinct descriptors. Without completeness, any ML model using the descriptors as input will give identical predictions of physically different systems. Ensuring completeness while preserving the invariances is nontrivial, however. One of the simplest descriptors is based on permutationally invariant pairwise atomic distances (2-body descriptors), and ref 412 demonstrated that these are generally not complete since one can construct two distinct tetrahedra using the same set of distances. Many have assumed that

permutationally invariant 3-body atomic descriptors uniquely specify atomic environments because of the tremendous success of ML models for chemical systems and particularly MLPs. However, refs 469 and 468 exemplify that structural degeneracies can be found even when using 3- or 4-body descriptors. This underscores an important shortcoming of state-of-the-art 3-body descriptors, such as ACSF,[411] SOAP,[412] FCHL,[417] and MBTR.[418] ACE[420] should be a complete descriptor of local environments, but its reliance on spherical harmonic expansion and the subsequent contraction makes their evaluations expensive. Hence, there are still opportunities to develop improved atomic descriptors.

4.1.3. Locality Approximation. Representing a many-body chemical system in terms of atomic environments brings physical significance since certain extensive physical properties (e.g., the total energy, total electrostatic charge, and polarizability of a system) can be approximated by the sum of the atomic contributions coming from each atomic environment, for example, $\Theta = \sum_i \theta(\mathcal{X}_i)$. This approximation is valid because the atomic contribution associated with a central atom is largely determined by its neighbors, and long-range interactions can be approximated in a mean-field manner without explicitly considering distant atoms. Such "locality" is tacitly assumed in many ML models for CompChem, and it is a crucial necessity for most common atomistic potentials and MLPs (section 2.2.6.). Most MLPs (e.g., BPNN,[274] GAP,[276] and DeepMD[462]) approximate the total energy of a system as sums of local atomic energies.

Figure 7 illustrates locality by showing a KPCA map of the atom environments of carbon in the QM9 set (see section 3.3 for more detailed descriptions of the data set). By color-coding the KPCA plot with the local energies from a SOAP-based GAP model trained on QM9 energies,[470] one observes a systematic and smooth trend in energies across clusters. The total molecular energy can then be accurately predicted by the sum of local energies, which means the total energy can be approximated on the basis of all the local environments contained in the molecule. For example, an NN potential trained on liquid water simulations can predict the densities, lattice energies, and vibrational properties of diverse ice phases because the local atomic environments found in liquid water span the similar environments as those observed in ice phases.[471] Another GAP potential of carbon trained on amorphous structures and other crystalline phases predicted novel carbon structures in random structure searches as well as approximate reaction barriers.[472,473]

The locality approximation is typically rationalized based on the multiscale nature of interatomic interactions in chemical systems. It is generally expected that shorter interatomic distances correspond to stronger interactions, such that a cutoff may be imposed after a certain radial distance d given a certain energy accuracy threshold ϵ. The multiscale nature of physical interactions underlies the usual classification of chemical interactions, from strong covalent bonds and ionic interactions to weaker noncovalent hydrogen bonds and van der Waals interactions. However, our understanding of noncovalent interactions in large molecules and materials is still emerging,[36] and no general rules-of-thumb exist to define the cutoff distance d corresponding to a defined ϵ. Moreover, the sufficiency of the locality argument also depends on the phase of the system and whether the system is extended or not.[474] Hence, for systems having long-range interactions

(which includes most chemical systems), the locality assumption needs revision. There are currently three schools of approaches handling the long-range interactions. The first is to use global ML models, such as (s-)GDML,[207,372] which learn global interactions directly. Global models tend to be more data-efficient because they focus on learning a full molecular or material PES, but this significantly limits transferability since the ML model alone can only be used on the system it was trained upon. The second is to learn the charges[475,476] and multipoles[477] for each atom, and then the long-range electrostatic interactions based on environment-dependent charges or multipoles can be explicitly included using Coulomb's law. To ensure that the sum of the atomic charges reaches neutrality, charge equilibration schemes can be used.[478] The third is to capture the long-range electrostatic effects by introducing a nonlocal long-distance equivariant (LODE) representation,[479,480] which is dependent on the electrostatic field generated by the decorated atom density.

4.1.4. Advantages of Built-In Symmetries. Built-in symmetry in ML models substantially compresses the dimensionality of atomic representations and ensures that physically equivalent systems are predicted to have identical properties. One of the most rigorous ways of imposing symmetry onto a model f is via the invariant integration over the relevant group $\mathcal{S}$

$$f_{sym}(\mathbf{x}) = \int_{\pi \in \mathcal{S}} f(\mathbf{P}_\pi \mathbf{x}) \tag{34}$$

where $\mathbf{P}_\pi \mathbf{x}$ is a permutation of the input. However, the cardinality of even basic symmetry groups is exceedingly high, which makes this operation prohibitively expensive. This combinatorial challenge can be solved by limiting the invariant integral to the physical point group and fluxional symmetries that actually occur in the training data set, as done in sGDML.[207] Alternative approaches, such as parameter sharing[352,421−423] or density representations,[276] have also proven effective. For example, the DeepMD potential has two versions, the Smooth Edition (DeepPot-SE) explicitly preserves all the natural symmetries of the molecular system, and the other version that does not.[462] The DeepPot-SE offers much improved stability and accuracy.[207,462]

For ML predictions of scalar properties, the rotationally invariant atomic descriptor framework described earlier is appropriate. One may wish to predict vectorial or tensorial properties including dipole moments, polarizability, and elasticity. A covariant version of descriptors may be advantageous, and this can be expressed as

$$S(\psi(\{\mathcal{R}^{3N}\})) = \psi(S(\{\mathcal{R}^{3N}\})) \tag{35}$$

where S indicates a symmetry operation such as a rigid rotation about an axis. Ref 481 proposed a general method for transforming a standard kernel for fitting scalar properties into a covariant one. Ref 482 derived a rotational-symmetry-adapted SOAP kernel, which can be understood as using the angular-dependent SOAP vectors based on spherical harmonics expansions as the descriptors. Note that the SOAP kernels for learning scalar properties introduced in ref 412 remove angular dependencies by summing up the SOAP vectors in separate spherical harmonics channels.

Symmetry can be further exploited into "alchemical" representations that incorporate similarity between chemical species that are relatable by changing one atom into another. The FCHL[417] representation considers the similarity between

elements in the same row and columns of the periodic table and performs very well on chemical compounds across chemical space. Ref 483 compiled a data-driven periodic table of the elements by fitting to an elpasolite data set using an alchemical representation.

4.1.5. End-to-End NN Representations. All descriptors introduced above rely on a suitable set of hyperparameters (e.g., length scales, radial and angular resolution). Determining an optimal set of hyperparameters can be a tedious process, especially when heuristics are unavailable or fail due to the structural and compositional complexity of the system. A poor choice of descriptors can limit the accuracy of the final ML model, for example, when certain interatomic distances can not be resolved.

End-to-end NN representations follow a different strategy to learn a representation directly from reference data. Using atom types and positions of a system as inputs, end-to-end NNs construct a set of atom-wise features $\mathbf{x}_i$. These features are then used to predict the property of interest, for example, the energy as a sum of atom-wise contributions. Unlike static descriptors, the representation is also optimized as part of the overall training process. This way end-to-end NNs can adapt to structural features in the data and the target properties in a fully automatic fashion to eliminate the need for extensive feature engineering from the practitioner.

The deep tensor NN framework (DTNN)[352] introduced a procedure to iteratively refine a set of atom-wise features $\{\mathbf{x}_i\}$ based on interactions with neighboring atoms. Higher-order interactions can then be captured in an hierarchical fashion. For example, a first information pass would only capture radial information, but further interactions would recover angular relations and so on. In DTNN, a learnable representation depending only on atom types $\mathbf{x}_i^0 = \mathbf{e}_{z_i}$ serves as an initial set of features. These are then refined by successive applications of an update function depending on the atomic environment that takes the general form

$$\mathbf{x}_i^{l+1} = F^l\left(\mathbf{x}_i^l + \sum_j G^l[\mathbf{x}_i^l, \mathbf{x}_j^l, \mathbf{g}^l(|\mathbf{r}_i - \mathbf{r}_j|)]f_{\mathrm{cut}}(|\mathbf{r}_i - \mathbf{r}_j|)\right)$$

(36)

Here, l indicates the number of overall update steps. The sum runs over all atoms j in the local environment, and a cutoff function f_{cut} ensures smoothness of the representation. Each feature is updated with information from all neighboring atoms through the interaction function G. Apart from the neighbor features $\mathbf{x}_j$, G also depends on the interatomic distance $|\mathbf{r}_i - \mathbf{r}_j|$, which is usually expressed in the form of a radial basis vector $\mathbf{g}$. After the update, an atom-wise transformation F can be applied to further modulate the features. Since each update depends only on the interatomic distances and the summation over neighboring atoms is commutative, end-to-end NNs of this type automatically achieve a representation that is invariant to rotation, translation and permutations of atoms. Using these atom-type dependent embeddings compactly encodes elemental information. This is advantageous for systems comprised of many different chemical elements. Such multicomponent systems can be problematic to treat with predefined descriptors (e.g., ACSFs or SOAP), as these typically introduce additional entries for each possible combination of atom types, resulting in a large number of descriptor dimensions.

Since the introduction of DTNN, many different types of end-to-end NNs have been developed, and these vary by the choice for the functions F and G. For example, SchNet[434] uses continuous convolutions inspired by convolutional neural networks (CNNs) to describe the interatomic interactions. In this case, the update in eq 36 takes the form

$$\mathbf{x}_i^{l+1} = \mathbf{x}_i^l + \mathrm{NN}_{\mathrm{tr}}^l\left(\sum_j \mathbf{x}_j^l \mathrm{NN}_{\mathrm{rad}}^l[\mathbf{g}^l(|\mathbf{r}_i - \mathbf{r}_j|)]f_{\mathrm{cut}}(|\mathbf{r}_i - \mathbf{r}_j|)\right)$$

(37)

where the feature transformation ($\mathrm{NN}_{\mathrm{tr}}$) and the radial dependence ($\mathrm{NN}_{\mathrm{rad}}$) are both modeled as trainable NNs.

Other ML models introduce additional physical information. The hierarchical interacting particle NN (HIP-NN)[484] enforces a physically motivated partitioning of the overall energy between the different refinement steps, while the PhysNet architecture[485] introduces explicit terms for long-range electrostatic and dispersion interactions. In ref 421, Gilmer et al. categorize graph networks of this general type as message-passing NNs (MPNNs) and introduce the concept of edge updates. These make it possible to use interatomic information beside the radial distance metric in the refinement procedure, and they have since been adapted for other architectures.[486] Another interesting extension are end-to-end NNs incorporating higher-order features beside the scalar $\mathbf{x}_i$ used in the original DTNN framework. These are equivariant features that encode rotational symmetry and can be based on angles, dipole moment vectors, or features that can be expressed as spherical harmonics with $l > 0$. This enables the exchange of only radial information between atoms in each interaction pass and instead include higher structural information, such as dipole−dipole interactions or angular information. In addition, equivariant end-to-end NNs can also be used to predict vectorial or tensorial properties in a manner similar to the rotational-symmetry-adapted SOAP kernel. Examples include TensorField networks,[464] Cormorant,[463] DimeNet,[487] PiNet,[488] and FieldSchNet.[300]

4.2. From Descriptors to Predictions

After a descriptor vector for each chemical structure is defined, one can then construct the design matrix and the kernel matrix for a set of structures. These matrices can then be used as the input of ML models. As described in section 2, supervised ML methods, such as NNs and GPs, can be used to approximate nonlinear and high-dimensional functions, particularly when massive amounts of training data become available. Thus, one should expect that using CompChem would be very useful for generating a large amount of almost noise-free training data of specific systems or atomic configurations, as long as a physically accurate method is being applied in the right way with appropriate computational resources. In contrast, experimental observations can be difficult to measure and reproduce precisely. Note that the aim of most CompChem+ML efforts have a similar scope as decades-old quantitative structure activity/property relationship (QSAR/QSPR) models that are often based on experiments or CompChem modeling.[326,327,489] Thus, researchers in CompChem+ML should be aware of potentially relatable work done by the QSAR/QSPR communities, and to what extent questions being posed have been sufficiently answered. On the other hand, ML usually provides higher accuracy than non-ML

Table 5. ML Databases for CompChem

| database	description	location
AFLOWLIB	databases containing calculated properties of over 625k materials[510]	http://www.aflowlib.org
ANI-1	large computational DFT database, which consists of more than 20 M off equilibrium conformations for 57.5k small organic molecules[511,512]	https://github.com/isayev/ANI1_dataset
ANI-1x/ANI-1ccx	ANI-1x contains multiple QM properties from 5 M DFT calculations, while ANI-1ccx contains 500k data points obtained with an accurate CCSD(T)/CBS extrapolation[513]	https://github.com/aiqm/ANI1x_datasets
BindingDB	measured binding affinities focusing on interactions of proteins considered to be candidates as drug-targets; 1 200 000 binding data for 5500 proteins and over 520 000 drug-like molecules[514]	http://www.bindingdb.org
Clean Energy Project	contains ~10 000 000 molecular motifs of potential interest which cover small molecule organic photovoltaics and oligomer sequences for polymeric materials[515]	http://cepdb.molecularspace.org
CoRE MOF	database containing over 4700 porous structures of metal−organic frameworks with publicly available atomic coordinates; includes important physical and chemical properties[516]	10.11578/1118280
FreeSolv	experimental and calculated hydration free energies for neutral molecules in water[517]	http://www.escholarship.org/uc/item/6sd403pz
GDB	GDB-11, GDB-13, and GDB-17; together these databases contain billions of small organic molecules following simple chemical stability and synthetic feasibility rules[518]	http://gdb.unibe.ch/downloads/
Hypothetical Zeolites	contains approximately 1 M zeolite structures[519]	http://www.hypotheticalzeolites.net/
Materials Project	contains computed structural, electronic, and energetic data for over 500k compounds[520]	https://www.materialsproject.org
MD17	data sets in this package range in size from 150k to nearly 1 M conformational geometries; all trajectories are calculated at a temperature of 500 K and a resolution of 0.5 fs[372]	http://www.sgdml.org
MoleculeNet	contains data on the properties of over 700k compounds[521]	http://moleculenet.ai
Open Catalyst Project	1.2 M molecular relaxations with results from over 250 M DFT calculations relevant for renewable energy storage[522]	https://opencatalystproject.org/index.html
OQMD	consists of DFT predicted crystallographic parameters and formation energies for over 200k experimentally observed crystal structures[523]	http://oqmd.org
PubChemQC PM6	provides 221 million molecular structures optimized with the PM6 method and several electronic properties computed at the same level of theory[524]	http://pubchemqc.riken.jp/pm6_datasets.html
PubChemQC	provides ~3 million molecular structures optimized by DFT and excited states for over 2 million molecules using TD-DFT[525]	http://pubchemqc.riken.jp/
QM7-X	comprehensive data set of 42 physicochemical properties for ~4.2 M equilibrium and nonequilibrium structures of small organic molecules with up to seven non-hydrogen (C, N, O, S, Cl) atoms[526]	https://zenodo.org/record/4288677#.X9jHNC2ZNTY
QM9	geometric, energetic, electronic, and thermodynamic properties for 134k stable small organic molecules out of GDB-17[527]	https://figshare.com/collections/Quantum_chemistry_structures_and_properties_of_134_kilo_molecules/978904
Synthesis Project	collection of aggregated synthesis parameters computed using the text contained within over 640 000 journal articles[528]	www.synthesisproject.org
quantum-machine.org	a repository of diverse data sets, including valence electron densities, chemical reactions, solvated protein fragments, and molecular Hamiltonians	http://quantum-machine.org/datasets/

statistical models, and so QSAR/QSPR efforts have been turning toward ML models as well.[490]

We have explained how data from different CompChem methods, each containing different degrees of physical rigor, can be used to train ML models. ML models in turn can be created to approximate underlying high-dimensional functions intrinsic to physical systems. For example, research efforts are toward learning electron densities,[491] density functionals,[162] and molecular polarizabilities.[492]

Besides these direct learning strategies, ML has been used to enhance the performance and suitability of CompChem models. As mentioned in section 1, the Δ-ML[493] approach is now a common technique for adapting an ML model that improves the quality of a theoretically insufficient but computationally affordable method. This approach has been used to learn many body corrections for water molecules to allow a relatively inexpensive KS-DFT approach like BLYP to more accurately reproduce CCSD(T) data.[494] Along similar lines, Shaw and co-workers used CompChem features along with an NN to reweight terms from an MP2 interaction energy to provide ML-enhanced methods with increased performance.[126] Miller and co-workers have developed ML-models where molecular orbitals themselves are learned to generate a density matrix functional that provides CCSD(T)-quality PESs

with a single reference calculation.[495] von Lilienfeld and co-workers have investigated how the choice of regressors and molecular representations for ML models impacts accuracy, and their findings suggest ways that ML models may be trained to be more accurate and less computationally expensive than hybrid DFT methods.[496] Burke and co-workers have studied how ML methods can result in improved understanding and more physical exact KS-DFT[181,497−499] and OFDFT functionals.[161] Brockherde et al. have presented an approach, where ML models can directly learn the Hohenberg−Kohn map from the one-body potential efficiently to find the functional and its derivative.[162,184] Akashi and co-workers have also reported the out-of-training transferability of NNs that capture total energies, which shows a path forward to generalizable methods.[500]

Toward predictive insights, there are many other approaches that are broadly useful. One can exploit the "universal approximator" nature of ML architectures to find a function that gives the best solution in a variational setting. For instance, using restricted Boltzmann machines[501] or deep NNs as a basis representation of wavefunctions[105,106,502] in Quantum Monte Carlo calculations. Alternatively, the use of active learning might increase the efficiency, accuracy, scalability, and transferability of ML models.[503−505]
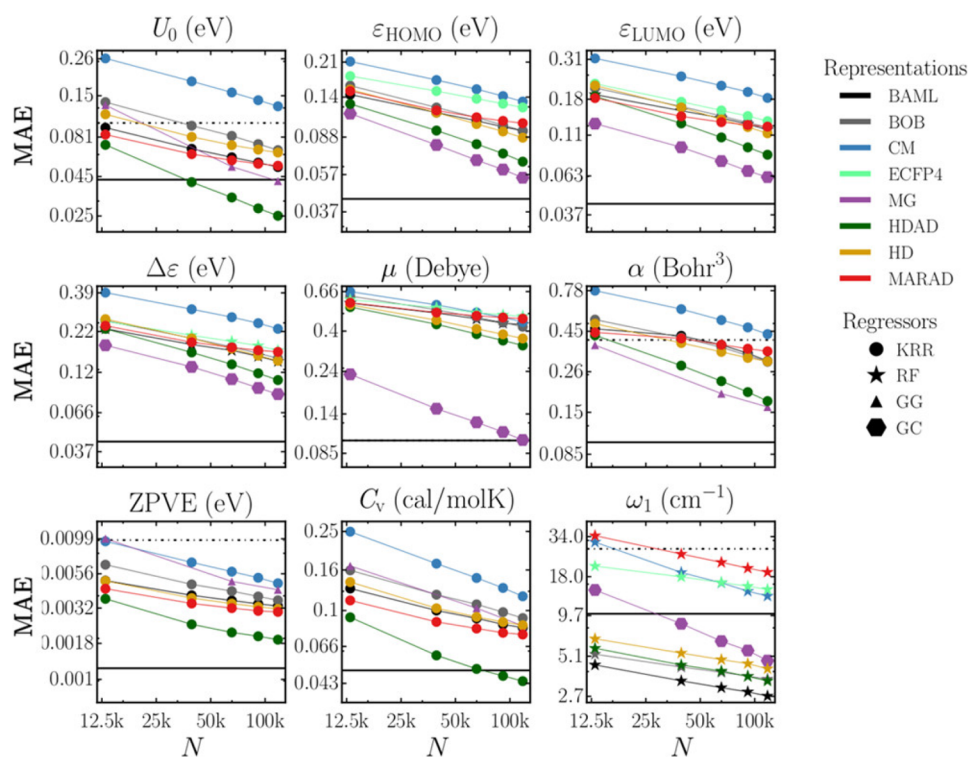
Figure 8. Learning curves of various properties contained in the QM9 database, reporting the predictive accuracy of various models as a function of training set sizes. Each curve represents an individual model based on a different architecture and descriptor. Shown are learning curves for the internal energy (U_0), HOMO and LUMO energies (ϵ_{HOMO}, ϵ_{LUMO}), the HOMO−LUMO gap ($\Delta\epsilon$), the length of the dipole moment vector (μ), the isotropic polarizability (α), the zero point vibrational energy (ZPVE), heat capacity (C_V), and the highest fundamental vibrational frequency (ω_1). Reprinted with permission from ref 496. Copyright 2020 American Chemical Society.
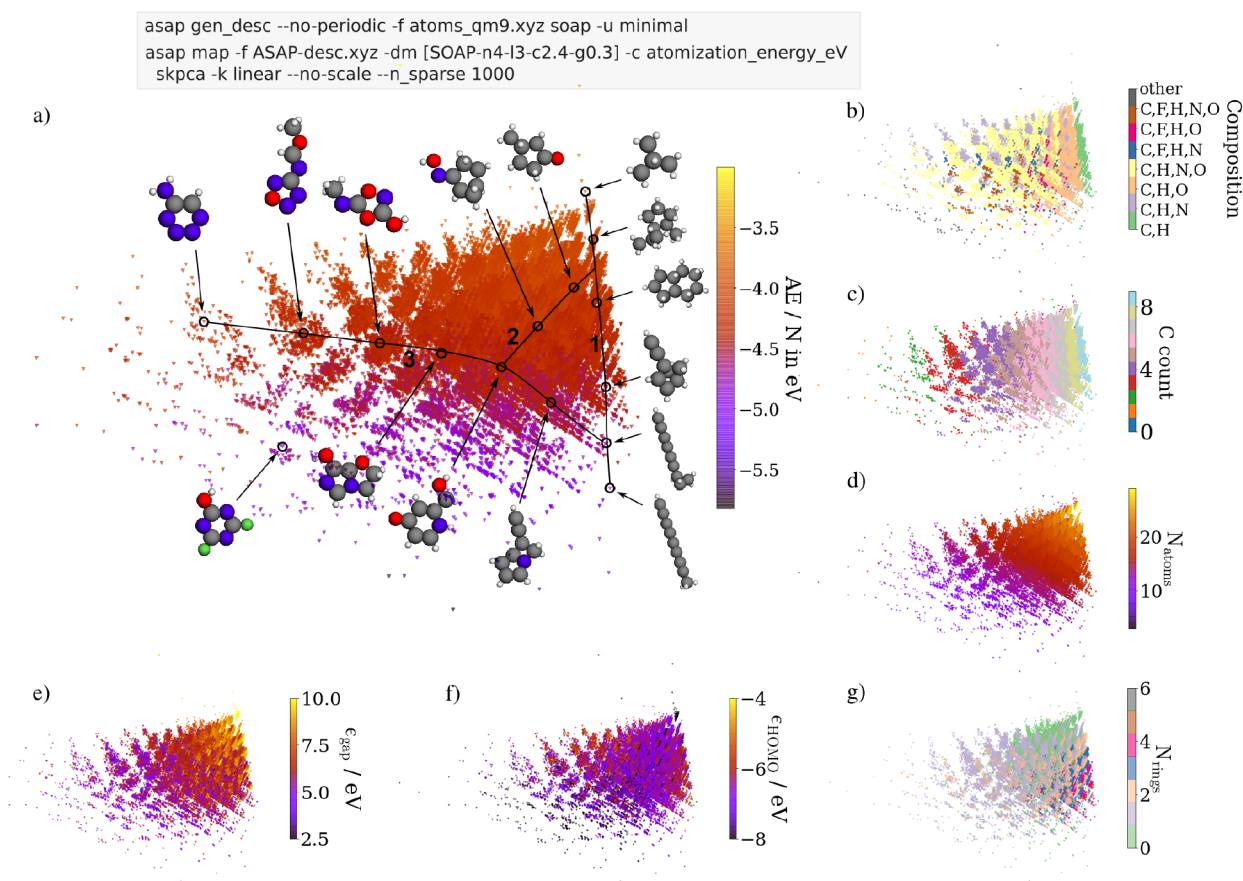
4.3. CompChem Data

We have laid the general framework for CompChem+ML studies, but this direction would not be complete without more details about training data (i.e., garbage in, garbage out). We now review the landscape of data sets in CompChem and how they will likely evolve over time. The past decade has seen continually increasing usefulness and availabilty of "big data" from CompChem that include community-wide data repositories comprised of millions of atomistic structures along with diverse physical and chemical properties.[506−509] Such repositories are becoming the norm, and it is more customary for different users to deposit raw or processed simulation data there for the benefit of the research community. This brings the possibility of robust validation tests for ML models, but it also necessitates approaches that are well-equipped to handle large and complex data sets. Typical data sets may come from diverse origins such as MD trajectories from *ab initio* simulations, data sets of small molecules and molecular conformers, or other training sets used for developing ML and non-ML FFs for specific applications. As the data sets grow, so do the scope of publications that involve ML as shown in Figure 1.

4.3.1. Benchmark Data Sets. ML models must be validated before they can be trusted for predictions. Validations of descriptors or model trainings are performed on benchmark data sets, and several popular ones are summarized in Table 5. These allow ML models to be compared on the same ground and provide large amounts of data for robust training. Their availability to the public also ensures that the data sets can evolve with time and be extended as a part of community efforts.[529]

Among the entries in Table 5, the most often used one is the QM9 set, which consists of approximately 134 000 of the smallest organic molecules that contain up to 9 heavy atoms (C, O, N, or F; excluding H) along with their CompChem-computed molecular properties such as total energies, dipole moments, HOMO−LUMO gaps, etc. Several ML studies have already been published using this data set (see Figure 8, ref 496). A popular challenge associated with QM9 is to develop a next-generation ML model that learns the electronic energies of random assortments of organic molecules with higher accuracy and less required training data than other existing models. Doing so tests next generation molecular representations and training algorithms. Figure 8 illustrates how the choice of architecture and descriptors can influence the predictive performance and data efficiency of ML models using different properties of the QM9 data set as examples. The next significant advance will potentially be due to a combination of supervised and unsupervised learning models.

4.3.2. Visualization of Data Sets. As the structural data sets grow it becomes infeasible to manually identify hidden patterns or curate the data. Data-driven and automated frameworks for visualizing these data sets become increasingly popular.[530−533] Dimensionality reduction effectively translates the high dimensional data (i.e., the *xyz*-coordinates for molecules or materials in different atomic configurations) into a low-dimensional space easily visualized on paper or a computer screen. In this way, entries such as those in the QM9 set can be shown (see Figure 9). The KPCA maps in Figure 9 are based on the dimensionality reduction of the global SOAP descriptors, which are constructed by combining all the atomic SOAP descriptors using eq 30. Each dot represents a small

```
asap gen_desc --no-periodic -f atoms_qm9.xyz soap -u minimal
asap map -f ASAP-desc.xyz -dm [SOAP-n4-l3-c2.4-g0.3] -c atomization_energy_eV
  skpca -k linear --no-scale --n_sparse 1000
```

Figure 9. KPCA maps of the QM9 database using a global SOAP kernel. The maps are color-coded according to atomization energy per atom (a), composition (b), number of carbon atoms in the molecule (c), total number of atoms in the molecule (d), HOMO−LUMO gap ϵ_{gap} (e), HOMO energies ϵ_{HOMO} (f), and the number of atoms in a ring (g). Examples of molecules along various "paths" in panel a are illustrated. Reprinted with permission from ref 374. Copyright 2020 American Chemical Society.

molecule in the QM9 set, and the maps, thus, illustrate the similarity between the molecules, instead of the relations between the carbon atomic environments in Figure 7. The maps in Figure 9 are color-coded using different molecular properties, such as the atomization energies, composition, and optical properties, and these properties are strongly correlated with the principal axes. These KPCA maps are, therefore, an intuitive and condensed way to help navigate the QM9 set. Similarly, ref 321 used SOAP-sketchmaps in conjunction with quasi-chemical theory to visualize similarities in local solvation structures and thus show an unsupervised learning procedure to identify structures that significantly impact solvation energies of small ions.

Generally speaking, these data-driven maps are generated by processing the design matrix (or kernel matrix) associated with a data set using dimensionality reduction techniques introduced in section 3.2. A simple option is to use the ASAP code,[374] a Python-based command line tool, that automates analysis and mapping. Figures 7 and 9 were generated using ASAP using only two commands that are displayed in the figure. Data sets can also be explored in an intuitive manner using interactive visualizers[534] that run in a web browser and display 3D-structures corresponding to each atomistic structure in the data set.

4.3.3. Text and Data Mining for Chemistry. Conventional publications are an essential part of any CompChem knowledge base, and ML is becoming useful at accelerating information extraction from the scientific literature via text

mining.[535−537] This topic was previously comprehensively reviewed in the context of cheminformatics.[538,539] Natural language processing has already driven text-mining efforts for materials science discovery[538] and experimental synthesis conditions of oxides.[528,540] CompChem+ML can also amplify existing efforts in chemometrics,[541] the science of data-driven extraction of chemical information.[542] This area has also branched into related disciplines of data mining for specific classes of materials[543] and catalysis informatics.[544] These approaches have great promise, especially for deriving information and knowledge from data, but it remains challenging to implement these in ways that achieve insight (and true impact).

Some have shown paths forward for doing so. For example, ML models can obtain knowledge from failed experimental data more reliably than humans who are more susceptible to survivor bias,[545] and it can also be used to distill physical laws and fundamental equations using experimental[363] and computational data.[546] ML models can also be used to reliably predict SMILES representations (a string-based representation of molecular graphs) that allow encoded information to be derived from low-resolution images found in the literature.[547] ML models can interpret experimental X-ray absorption near edge structure (XANES) data and predict real space information about coordination environments.[548] Likewise, scanning tunneling microscopy (STM) data can be used to classify structural and rotational states on surfaces,[549] and name indicators can be used to predict in tandem mass

spectrometry (MS/MS) properties.[550] In closing, we see exciting opportunities for future applications that complement data and text mining to chemometrics through chemical space.

4.4. Transforming Atomistic Modeling

We previously mentioned that ML can handle large data sets and extract insights while circumventing the high cost of quantum-mechanical calculations by statistical learning. CompChem+ML also has great potential in developing MLPs. Car and Parrinello proposed running MD using electronic-structure methods in 1985.[551] These are now mainstream but also quite computationally demanding and normally restricted to small system sizes ($\sim$100 atoms) and short simulation times ($\sim 10^{-12}$ s). Alternatively, accurate atomistic potentials introduced in section 2.2.6 have been developed to allow Monte Carlo and MD simulations, but sufficiently accurate potentials are sometimes not available. MLPs have emerged as way to achieve as high accuracy as KS-DFT or correlated wavefunction methods but with a fraction of the cost. MLPs have been constructed for far-reaching systems from small organic molecules to bulk condensed materials and interfaces.[433,552,553] Several of the coauthors of the current review have also written separate review focused more narrowly on this topic,[554] and so, we only provide a brief overview here.

Training an MLP to reproduce a system's PES usually requires generating diverse and high quality CompChem data points that cover the relevant temperature and pressure conditions, reaction pathways, polymorphs, defects, compositions, etc.[555−562] After data points comprised of atomic configurations, system energies, and forces are obtained, different methods for constructing MLPs employ either different descriptors (see a list of examples in Table 4) or different ML architectures to perform interpolations of the full PES. Again, smoothness is an essential feature for any PES, so special considerations are needed to avoid numerical noise that would result in discontinuities.[563,564] Kernel method-based MLPs, such as GAP[276,565] and sGDML,[207,372,566] ensure smoothness by relying on smoothly varying basis functions, but the scaling of kernel-based methods with respect to the number of training points is challenged without reduction mechanisms.[396,567] As a much more efficient but somewhat less accurate alternative to GAP, SNAP[568] uses the coefficients of the SOAP descriptors and assumes a linear or quadratic relation between energies and the SOAP bispectrum components.[569] The most popular MLPs are currently NN-based due to their flexibility and capacity to train based on large amounts data. Among these, ANI[511,513] and BPNN[274,433,570] potentials use ACSF descriptors as inputs, while Deep NNs, such as SchNet[422,434,571] and DeepMD[572] use the coordinates and nuclear charges of atoms. We now focus on a few example applications.

4.4.1. Predicting Thermodynamic Properties. Many CompChem efforts focus on predicting thermodynamic properties at finite temperatures, such as heat capacity, density, and chemical potential. Although many physical properties are already accessible from MD simulations, doing estimations of free energies that establish the relative stability of different states using electronic structure methods remains difficult. The configurational part of the Gibbs free energy of a bulk system that has N distinguishable particles with atomic coordinates $\mathbf{r} = \{\mathbf{r}_{1...N}\}$, and the associated potential energy $U(\mathbf{r})$ can be expressed as

$$G(N, P, T) = -k_{\mathrm{B}}T\ln \int d\mathbf{r} \exp\left[-\frac{U(\mathbf{r}) + PV}{k_{\mathrm{B}}T}\right] \quad (38)$$

integrated over all possible coordinates $\mathbf{r}$, where k_{B} is the Boltzmann constant. In order to rigorously determine G, one must exhaustively sample the configuration space that has relatively high weight arising from the $\exp\left[-\frac{U(\mathbf{r}) + PV}{k_{\mathrm{B}}T}\right]$. This normally requires thermodynamic integration or enhanced sampling methods (e.g., umbrella sampling,[573] metadynamics,[574] or transition path sampling[575]), that require simulation times and scales far beyond what is accessible with MD simulations based on KS-DFT or correlated wavefunction methods.

However, MLPs have unleashed both limits on the time scale and system size. An early example,[576] used an MLP with umbrella sampling[573] and the free energy perturbation method[577] to reveal the influence of van der Waals corrections on the thermodynamic properties of liquid water. Later, the combination of an MLP trained from hybrid DFT data and free energy methods reproduced several thermodynamic properties of water from quantum mechanics, including the density of ice and water, the difference in melting temperature for normal and heavy water, and the stability of different forms of ice.[578,579] Ref 580 employed the DeepMD approach to study the relatively long time-scale nucleation of gallium. MLPs for high-pressure hydrogen provided evidence on how hydrogen gradually turns into a metal in giant planets.[581] In all these examples, high accuracy and long time scales were required to model the specific phenomena and reveal physical insights, and it is precisely the combination of CompChem +ML that enables both.

4.4.2. Nuclear Quantum Effects. As mentioned in section 2.2.5, NQEs of chemical systems having light elements bring challenges for atomistic modeling because the added mobility of lighter atoms in dynamics simulations requires higher computational cost to treat. To make the matter even more complicated, many atomistic potentials (see section 2.2.6), particularly the ones for water or organic molecules, cannot be used to model NQEs, because they often describe colavent bonds as rigid and thus cannot describe the fluctuations of the bond lengths and angles. As a remedy, several studies have been performed by training an MLP using higher rungs of KS-DFT (e.g., hybrid-DFT or meta-GGA) and then using this potential in PIMD simulations.[578,582−584] The study of water mentioned in the previous section, which used MLPs trained from hybrid DFT, revealed that NQEs were critical for promoting the hexagonal packing of molecules inside ice that ultimately lead to the 6-fold symmetry of snowflakes.[578] Highly data efficient ML potentials can even be trained on reference data at the computationally very expensive quantum-chemical CCSD(T) level of accuracy. For example, the sGDML[206,207,585] approach has been shown to faithfully reproduce such FFs for small molecules, which were then used to perform simulations with effectively fully quantized electrons and nuclei.

4.5. ML for Structure Search, Sampling, and Generation

Locating stationary points on the PES is a frequent task in CompChem, since these are needed for explaining reaction kinetics. Explorations for stationary points normally require many energy and force evaluations. ML approaches are being implemented to dramatically accelerate minimum energy as

well as saddle-point optimizations.[293−295,565,586−588] Bernstein et al. proposed an automated protocol that iteratively explores structural space using a GAP potential.[565] Bisbo and Hammer employed an actively learned surrogate model of the PES to perform local relaxations while only performing single-point quantum-mechanical calculations for selected structures with high values of acquisition.[586] Work in refs 293 and 295−297 accelerated nudged elastic band (NEB) calculations by incorporating a surrogate ML models.

ML can also dramatically accelerate the challenge of efficiently sampling equilibrium or transition states by accelerating enhanced sampling methods such as umbrella sampling[573] and metadynamics.[574] These procedures make use of collective variables (CVs) that define a reaction coordinate, and computing the associated free energy surface (FES) amounts to generating the marginal probability distribution in these CVs. Unfortunately, the choice of the CVs is not always clear for specific systems, and ML has shown some promise in guiding their determination.[589−591] Another direction is to exploit that ML models can be considered as universal approximators of FESs.[592] For example, there are reports of adaptive enhanced sampling methods using a Gaussian Mixture model,[593] using an NN architecture to represent the FES[594] or the bias function in variational sampling simulations.[595]

ML methods also offer fundamentally new ways to explore chemical compound and configuration space. Generative models can learn the structural and elemental distribution underlying chemical systems, and once trained, these models can then be used to directly sample from this distribution. It is furthermore possible to bias the generated structures toward exhibiting desired properties, for example, drug activity or thermal conductivity. As a consequence, generative models offer exciting new avenues in drug and materials design.[596,597] Generative methods in CompChem include recurrent neural networks (RNNs), which can be used for the sequential generation of molecules encoded as SMILES strings.[598−600] Segler et al. demonstrated how such a recurrent model can first learn general molecular motifs and then be fine-tuned to sample molecules exhibiting activity against a variety of medical targets.[599] Autoencoders (AE) are another frequently used ML method for molecular generation. AEs learn to transform molecular graphs or SMILES into a low-dimensional feature space and backward. The resulting feature vector represents a smooth encoding of the molecular distribution and can be used to effectively sample chemical space.[601−606] By applying a variational AE to the QM9 and ZINC databases, Gomez-Bombarelli et al. could generate several optimized functional compounds.[607] An interesting extension to AEs are conditional AEs, which not only capture the distribution of molecular structures but also dependencies on various properties.[426,608] This makes it possible to directly generate structures exhibiting certain property ranges or combinations without the need for biasing or additional optimization steps. AEs can also form the basis of another approach for exploring chemical space called generative adversarial networks (GANs).[609,610] In a GAN, a generator model (often an AE) attempts to create samples that closely match the underlying data, while a discriminator tries to distinguish true from generated samples. These architectures can be enhanced by using RL objectives. RL learns an optimal sequence of actions (e.g., placement of atoms) leading to a desired outcome (e.g., molecule with certain property). This makes it possible to

drive generative processes toward certain objectives, allowing for the targeted generation of molecules with particular properties.[611−614] RL in general is a promising alternative strategy for generative models,[615,616] and they offer the possibility for tight integration into drug design cycles.[617] Alternative approaches combine autoregressive models with graph convolution networks.[618,619]

While these methods use SMILES or graphs to encode molecular structures, generative models have recently been extended to operate on 3D coordinates of molecules and materials.[620,621] Gebauer et al. proposed an autoregressive generative model based on the SchNet architecture, called g-SchNet.[622] Once trained on the QM9 data set, g-SchNet was able to generate equilibrium structures without the need for optimization procedures. It was further found, that the model could be biased toward certain properties. In another promising approach, Noé et al. used an invertible NN based on normalizing flows to learn the distribution of atomic positions (e.g., sampled from an MD trajectory). This network can then be used to directly sample molecular configurations by sampling from this distribution without performing costly simulations.[298]

4.6. Multiscale Modeling

Multiscale modeling is a term for including simulation or information from different scales (see Figure 3). ML has been introduced into QM/MM-like schemes that enable improved multiscale simulations,[300,325,623] and on the side of coarse-graining.[624] Different coarse-graining potentials have been developed,[625] but the inherent functional form for these potentials relies on CPI as well as trial-and-error procedures. Several works used ML for constructing coarse-grained potentials by matching mean forces.[449,450,626,627] In closing, we see promise for incorporating experimental priors into ML models, for instance, using experimental measurements to improve an ML PES by complementing them with experimental data. We are not aware of such efforts for developing highly accurate MLPs beyond the atomic scale, although much work has been done along this line to refine FFs of RNAs and proteins, often incorporating methods from ML, including the maximum entropy approach.[628]

5. SELECTED APPLICATIONS AND PATHS TOWARD INSIGHTS

The central challenge posed at the beginning of this review was how to identify and make chemical compounds or materials having optimal properties for a given purpose. To do so would help address critical and broad issues from pollution to global warming to human diseases. Traditional developments are often slow, expensive, and restricted by nontransferable empirical optimizations, and so efforts have turned to CompChem+ML to alleviate this.[515,525,629,630]

CompChem+ML are enabling searches through larger areas of chemical space much faster than before.[20,631−634] This section is not to extensively review the large amount of work using CompChem+ML in these different areas, but rather to highlight examples of applications that have resulted in notable insights so that others might use these notable works as templates for future efforts.

5.1. Molecular and Material Design

Molecules and materials design is usually considered to be an optimization problem.[270,426,602,607,635] Thus, a comprehensive understanding of chemical space is needed to identify

compounds with desired properties that are subject to certain required constraints (e.g., a specific thermal stability or a suitable optical gap for absorbing sunlight). Those properties will also depend on many key variables (e.g., constitutive elements, crystal forms, geometrical and electronic characteristics, among others), which make the property prediction complex.[531] CompChem calculations as explained in section 2 should provide a continuous description of properties across a continuous representation (i.e., a descriptor or fingerprint) of molecules that is used to map molecular configurations to target properties, and vice versa. ML methods then can be implemented to search large databases to extract structure−property relationships for designing compounds with specific characteristics.[531,635−637] Optimizations would then be performed on the structure-based function learned from training configurations, and the composition of the chemical compound would then be recovered back from the continuous representation.

As a protoypical example of molecular design via high-throughput screening, Gomez-Bombarelli et al.[632] showed a computation-driven search for novel thermally activated delayed fluorescence organic light-emitting diode (OLED) emitters. That work first filtered a search space of 1.6 million molecules down to approximately 400 000 candidates using ML to anticipate criteria for desirable OLEDs. For the purpose of evaluating candidates, they estimated an upper bound on the delayed fluorescence rate constant (k_{TADF}). TD-DFT calculations were then used to provide refined predictions of specific properties of thousands of promising novel OLED molecules across the visible spectrum so that synthetic chemists, device scientists, and industry partners would be able to choose the most promising molecules for experimental validation and implementation. Notably, this example of CompChem+ML resulted in new devices that exhibited an external quantum efficiency of over 22%. Figure 10 shows the high accuracy of ML in predicting useful properties for high-throughput
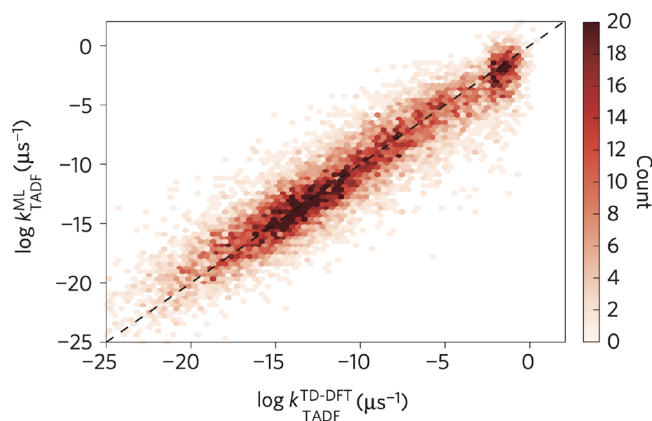
Figure 10. NN predictions compared to TD-DFT derived data of log k_{TADF} ($R^2 = 0.94$). ML models computed molecular properties needed for screening with an accuracy comparable to CompChem calculations, but at a fraction of the computational cost. Reprinted by permission from Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T., et al. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, 15, 1120−1127.[632] Copyright 2016 Springer Nature, Nature Materials.

screening of molecules and materials based on k_{TADF} calculations. This work exemplifies how ML can accelerate the design of novel compounds in such a way that could not be possible using traditional CompChem methods alone.

Integrations of features relevant to learning tasks allow one to improve the accuracy of ML predictions for a given target property. Park and Wolverton[638] improved the performance of the crystal graph convolution neural network (CGCNN)[639] by adding to the original framework information about the Voronoi tessellated crystal structures, which are explicit 3-body correlations of neighboring constituent atoms, and an optimized representation of interatomic bonds. The new approach that was labeled as iCGCNN achieved a predictive accuracy 20% higher than that of the original CGCNN when determining thermodynamic stabilities of compounds (i.e., predictions of hull distances). When used for high-throughput searches, iCGCNN exhibited a success rate higher than an undirected high-throughput search and higher than that of CGCNN. Figure 11 shows the improvement in predictions of nearly stable compounds after using more appropriate descriptors. This study showcases how descriptors can be tailored to further enhance the success of ML-aided high-throughput screening.

5.2. Retrosynthetic Technologies

A grand challenge in chemistry is to understand synthetic pathways to desired molecules.[640,641] Retrosynthesis involves the design of chemical steps to produce molecules and materials that would be crucial to drug discovery, medicinal chemistry, and materials science. As a different kind of optimization problem, the general tactic is to analyze atomic scale compounds recursively, map them onto synthetically achievable building blocks, and then assemble those blocks into the desired compound.[642−644]

Three main issues make retrosynthesis a formidable intellectual challenge.[645] First, simple combinatorics make the space of possible reactions greater than the space of possible molecules. Second, reactants seldom contain only one reactive functional group, and thus require predictions of multiple functional groups. Third, one failed step in the route can invalidate the entire synthesis because organic synthesis is a multistep process.

Given these challenges, ML is becoming more established in determining reaction rules from CompChem data.[641] Computer-aided synthesis planning was actually first attempted in the 1960s.[646] Many have since attempted to formalize chemical perception and synthetic thinking using computer programs.[647−649] These programs are typically based on one of three possible algorithms:[649]

1. Algorithms that use reaction rules (manually encoded or automatically derived from databases).
2. Algorithms that use principles of physical chemistry based on ab initio calculations to predict energy barriers.
3. Algorithms based on ML techniques.

ML approaches are used to try to overcome the generalization issues of rule-based algorithms (that normally suffer from incompleteness, infeasible suggestions, and human bias) while also avoiding the high cost of CompChem calculations. It is now possible to obtain purely data-driven approaches for synthesis planning, which are promoting a rapid advancement in the field. For example, Coley and co-workers[650] designed a data-driven metric, SCScore, for describing a real synthesis modeled after the idea that products are, on average, more
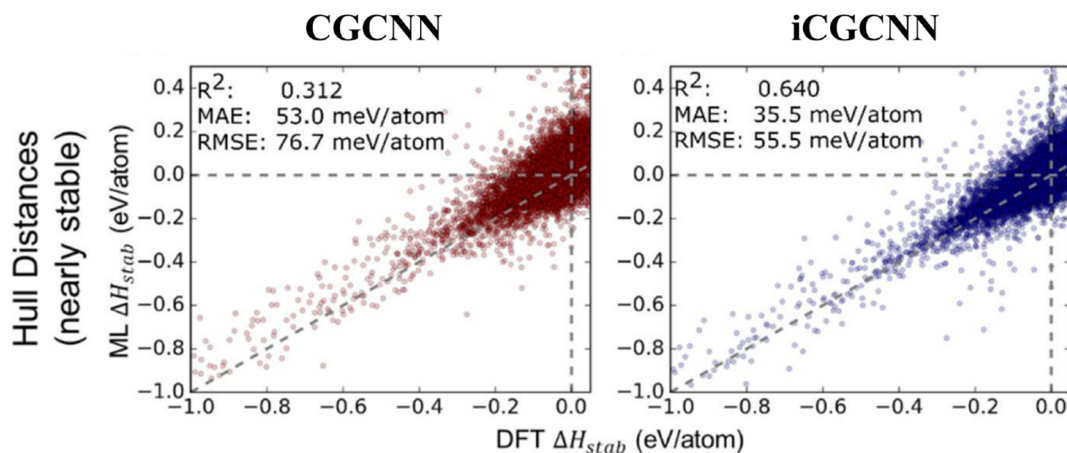
Figure 11. DFT vs ML predicted hull distances of nearly stable compounds (hull distances smaller than 50 meV/atom) for CGCNN and iCGCNN. The flexibility of ML approaches enable constructions of robust models tailored for specific target properties. See ref 638.

synthetically complex than each of their reactants. The definition of a metric for selecting the most promising disconnections that produce easily synthesizable compounds is crucial for avoiding combinatorial explosions. Figure 12 shows that a data-driven metric, the SCScore, is more suitable than other heuristic metrics to perceive the complexity of each step in a given synthesis. This work offered a valuable contribution to the retrosynthesis working pipeline by providing a method that implicitly learns what structures and motifs are more prevalent as reactants.

Apart from isolated approaches or algorithms to deal with specific tasks within retrosynthesis, there is already software available to advance this field. One example is the Chematica program,[651] which has implemented a new module that combines network theory, modern high-power computing, AI, and expert chemical knowledge to design synthetic pathways. A scoring function is used to promote synthetic brevity and penalize any reactivity conflicts or nonselectivities, thus allowing it to find solutions that might be hard for a human to identify. Figure 13A shows the decision tree for one of the almost 50 000 reaction rules used in Chematica. Reaction rules can be considered as the allowed moves from which the synthetic pathways are built, and such moves lead to an enormous synthetic space (the number of possibilities within n steps scales as 100^n) as the one shown by the graph in Figure 13B. Chematica explores this large synthetic space by truncating and reverting from unpromising connections and drives its searches to the most efficient sequences of steps. Moreover, in the pathways presented to the user, each substance can be further analyzed with molecular mechanics tools. This software was used to obtain insights into the synthetic pathways to eight targets (seven bioactive substances and one natural product). All of the computer-planned routes were not only successfully carried out in the laboratory, but they also resulted in improved yields and cost savings over previous known paths. This work opened an avenue for chemists to finally obtain reliable pathways from in silico retrosynthesis. For further reading we recommend the two-part reviews of Coley and co-workers.[652,653]

5.3. Catalysis

Catalysis research involves understanding how to impact chemical product yields and selectivities.[654] Traditional catalysis is normally discussed in textbooks in terms of
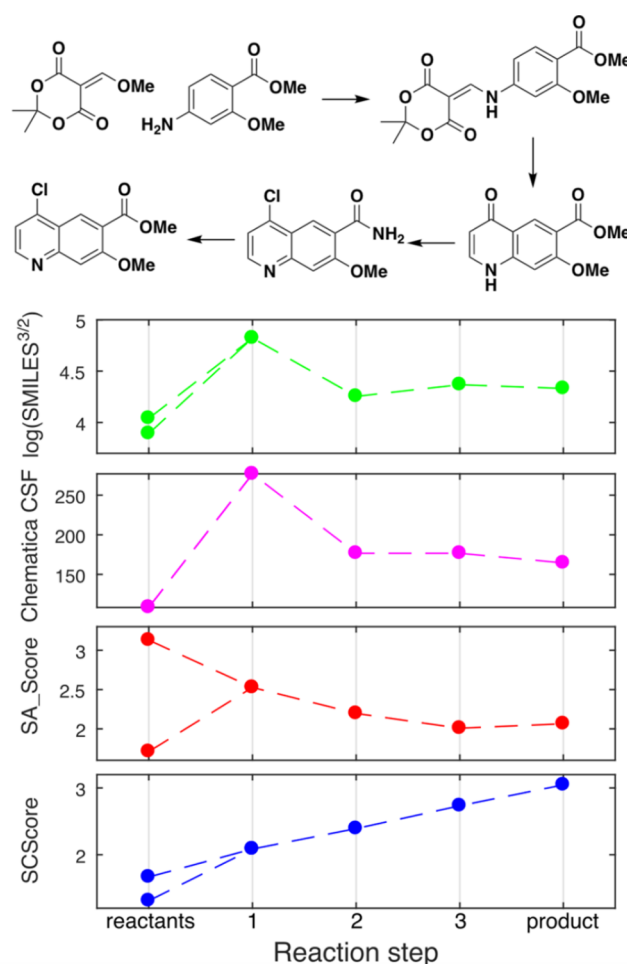
Figure 12. Use of different metrics to analyze the synthesis of a precursor to lenvatinib. Only the SCScore, a data-driven metric, correctly perceives a monotonic increase in complexity. ML models can give insights into which compounds are either reactants or products. Reprinted with permission from ref 647. Copyright 2018 American Chemical Society.

homogeneous (i.e., within a solution phase), heterogeneous (occurring at a solid/liquid interface), and biological (occurring within enzymes and riboenzymes), but it is best not to use these terms too strictly because actual reaction
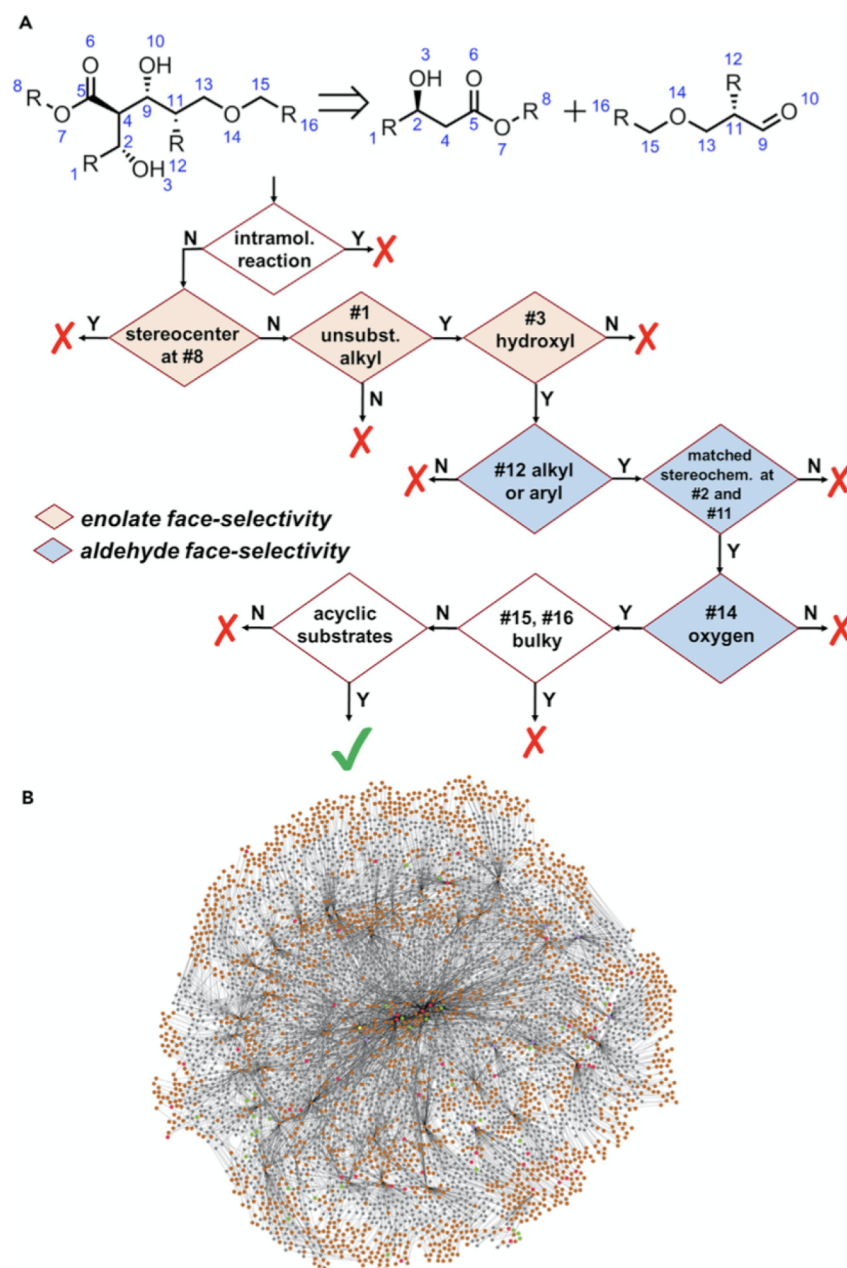
Figure 13. (A) Decision tree of one of the reaction rules within Chematica (double stereodifferentiating condensation of esters with aldehydes). The different conditions in the tree specify the range of admissible and possible substituents or atom types. (B) Reaction rules are used to explore the graph of synthetic possibilities (similar to the one shown here). Each node corresponds to a set of substrates. The combination of expert chemical knowledge, CompChem calculations and ML enables finding synthesizable paths. See ref 651. Reprinted from **Chem**, 4(3), Klucznik, T., et al., Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory, 522–532, Copyright (2018), with permission from Elsevier.

mechanisms can be quite complex and overall processes may sometimes exhibit characteristics of two or more of these classical processes.[655−657] Modern research in catalysis has been interested in studying chemical reactivity and reaction selectivity arising from stimuli from solar−thermal energy,[658,659] electrochemical potentials,[660] photons,[661−664] plasmas,[665,666] or other external resonances.[667] Catalysis makes up roughly 35% of the world's gross domestic product,[668] and it is important to guide toward the end goal of achieving greater sustainability with catalytic processes.[669−671]

These reasons help make catalysis a fertile training ground for applying and developing theoretical models (e.g., refs 672−674) that can be used along with CompChem or CompChem+ML. The research field is also burgeoning with many reports and review articles[544,675−679] that discuss perspectives and progress using ML methods for catalysis science. Here, we will mention notable examples. For example, CompChem+ML methods are enabling more data generation by allowing costly CompChem calculations to be run more efficiently, and more information means more comprehensive predictions of chemical and materials phase diagrams for catalysis[680,681] as well as stability and reactivity descriptors
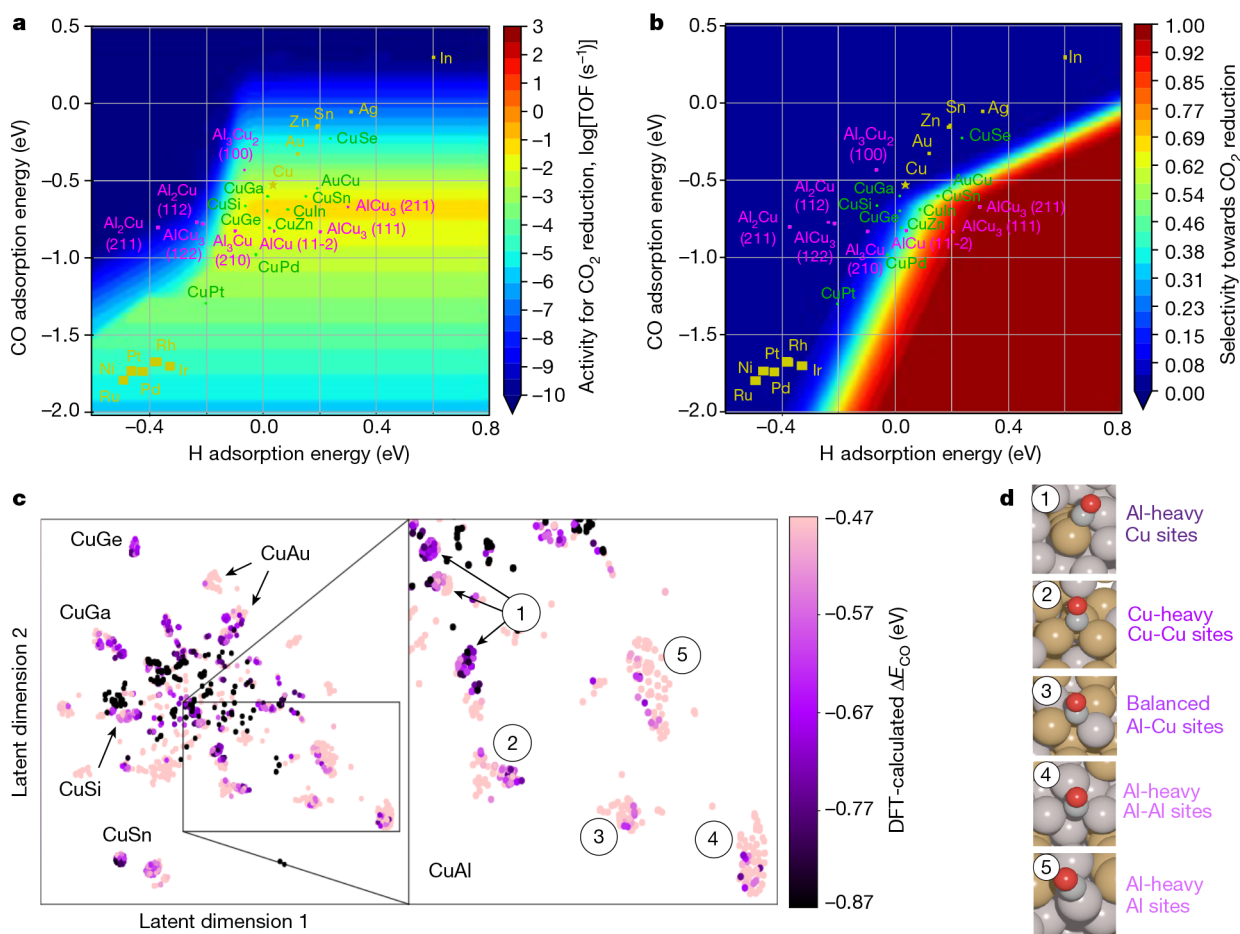
Figure 14. CompChem+ML screening of hypothetical Cu and Cu-based catalyst sites. (a) Two-dimensional activity volcano plot for CO_2 reduction. TOF, turnover frequency. (b) Two-dimensional selectivity volcano plot for CO_2 reduction. CO and H adsorption energies in panels a and b were calculated using DFT. Yellow data points are average adsorption energies of monometallics; green data points are average adsorption energies of copper alloys; and magenta data points are average, low-coverage adsorption energies of Cu–Al surfaces. (c) t-SNE[687] representation of approximately 4000 adsorption sites on which DFT calculations were performed with Cu-containing alloys. The Cu−Al clusters are labeled numerically. (d) Representative coordination sites for each of the clusters labeled in the t-SNE diagram. Each site archetype is labeled by the stoichiometric balance of the surface, that is, Al-heavy, Cu-heavy or balanced, and the binding site of the surface. See ref 688. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Nature*. Accelerated discovery of CO_2 electrocatalysts using active machine learning, Zhong, M., et al., Copyright 2020.
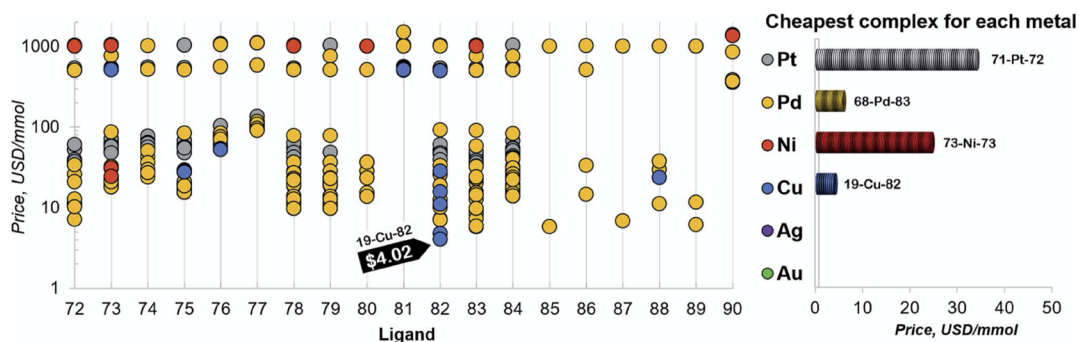
Figure 15. Estimated price (for one mmol in US dollars) of the catalysts in the selected range of $-32.1/-23.0$ kcal mol^{-1} (for ligand no. 72-90). The price is calculated as a summation of the commercial price of transition metal precursors (one mmol) and one mmol of each ligand. The cheapest complex for each metal is shown on the right. The estimated price of all the 557 catalysts is detailed in ref 694. Published by The Royal Society of Chemistry.

identified on the fly.[682−686] Figure 14 shows examples of the palettes of insight available using state-of-the-art CompChem +ML modeling for identifying activity and selectivity maps, as well as visualizations of data using t-SNE.[687]

Regarding modeling of deeply complex chemical environments, Artrith and Kolpak developed MLPs for investigating the relationships between solvent, surface composition and morphology, surface electronic structure, and catalytic activity
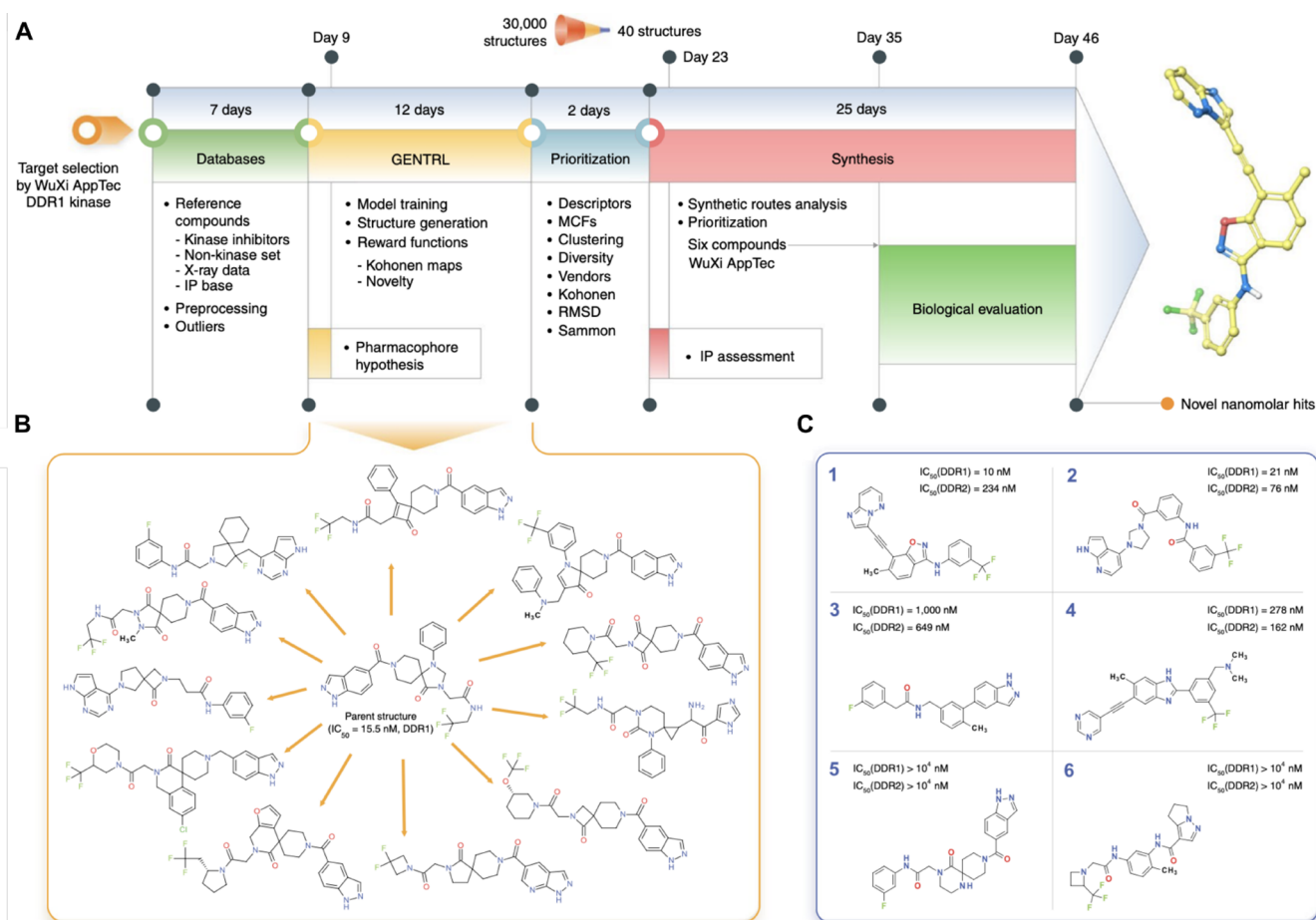
Figure 16. (A) Workflow and timeline for the design of candidates employing GENTRL. (B) Representative examples of the initial 30,000 structures compared to the parent DDR1 kinase inhibitor. (C) Compounds found to have the highest inhibition activity against human DDR1 kinase. CompChem+ML methods can considerably accelerate the discovery of drugs that are effective against a desired target. See ref 617. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Nature Biotechnology*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors, Zhavoronkov, A., et al., Copyright 2019.

in systems composed of thousands of atoms interfaces.[689] We expect such simulations for electro- and photocatalysis elucidation will continue to improve in size, scale, and accuracy. For other physical insights, new approaches by Kulik, Getman, and co-workers have also focused on developing ML models appropriate for elucidating complex *d*-orbital participation in homogeneous catalysis.[690] Rappe and co-workers have used regularized random forests to analyze how local chemical pressure effects adsorbate states on surface sites for the hydrogen evolution reaction.[691] Almost trivially simple ML approaches can be used in catalysis studies to deduce insights into interaction trends between single metal atoms and oxide supports,[692] to identify the significance of features (e.g., adsorbate type or coverage), where CompChem theories break down,[693] or they can be used to identify trends that result in optimal catalysis across multiple objectives, such as activity and cost (Figure 15).[694]

ML is also opening opportunities for CompChem+ML studies on highly detailed and complex networks of reactions.[695−700] Such models in principle can then significantly extend the range of utility of microkinetics modeling for predictions of products from catalysis.[701,702] ML also enables studies of complicated reaction networks that can allow predictions of regioselective products based on CompChem data,[703] asymmetric catalysis important for natural product

synthesis,[704,705] and biochemical reactions.[706] Efforts to better understand "above-the-arrow" optimizations of reaction conditions relate back to the challenge of retrosynthetic challenges.[707,708] Ideally, these efforts will continue while making use of rapid advances in CompChem+ML that enable predictive atomistic simulations to be run faster and more accurately. We see reason for excitement for different approaches, but we again stress the importance of ensuring that models will provide unique and physical results (see section 3 where we discuss the risk of "clever Hans" predictors[360]).

5.4. Drug Design

The central objective for drug discovery is to find structurally novel molecules with precise selectivity for a medicinal function. This involves identifying new chemical entities and obtaining structures with different physicochemical and polypharmacological properties (i.e., combinations of beneficial pharmacological effects or adverse side-effects).[709,710] Drug discovery involves the identification of targets (a property optimization task, as in material design) and the determination of compounds with good on-target effects and minimal off-target effects.[711] Traditionally, a drug discovery program may take around six years before a drug candidate can be used in clinical trials, and six or seven more years are required for three
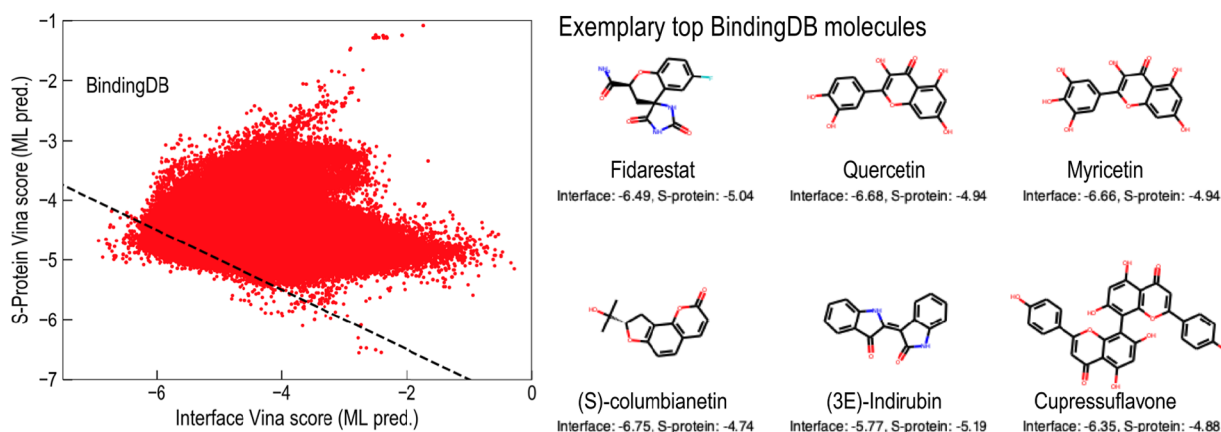
Figure 17. Vina scores predictions for the isolated protein (S-protein) and the protein-receptor complex (interface) for all the molecules in the BindingDB data sets and some exemplary top cases that satisfy the screening criteria. ML models trained on accurate CompChem databases are of upmost importance to efficiently gain insights into possible treatments, even for newly discovered diseases. Figure taken from ref 718. Copyright 2020 American Chemical Society.

clinical phases. Thus, it is important to identify adverse effects as soon as possible to minimize time and monetary costs.[712] Accelerating drug discovery relies on predicting how and where a certain drug binds to more than one protein, a phenomenon that sometimes results in polypharmacology. Researchers are developing ready-to-use tools aimed to facilitate research for drug discovery,[713] but CompChem+ML is expected to continue providing even more benefits to the drug development pipeline.[714]

In a recent study, Zhavoronkov et al.[617] developed a deep generative model for de novo small-molecule design: the generative tensorial reinforcement learning (GENTRL) model that was used to discover potent inhibitors of discoidin domain receptor 1 (DDR1), a kinase target implicated in fibrosis and other diseases. The drug discovery process was carried out in only 46 days, beginning with the recollection of appropriate data for training and finishing with the synthesis and experimental test of some compounds (Figure 16A). GENTRL was used to screen a total of 30 000 structures (some examples compared to the parent DDR1 kinase inhibitor are shown in Figure 16B) down to only 40 structures that were randomly selected ensuring a coverage of the resulting chemical space and distribution of root-mean squared deviation values. Six of these molecules were then selected for experimental validation (see Figure 16C), with one of them demonstrating favorable pharmacokinetics in mice. The predicted conformation of the successful compound according to pharmacophore modeling was very similar to the one predicted to be preferred and stable by CompChem methods. This work illustrates the utility of CompChem+ML approaches to give insights into drug design by rapidly giving compound candidates that are synthetically feasible and active against a desired target.

Besides generating new chemical structures with favorable pharmacokinetics, ML methods are also used in pharmaceutical research and development for peptide design, compound activity prediction and for assisting scoring protein−ligand interaction (docking).[709,715−717] An example of the latter was proposed by Batra et al.[718] for efficiently identifying ligands that can potentially limit the host−virus interactions of SARS-CoV-2. Those authors designed a high-throughput strategy based on CompChem+ML that involved high-fidelity docking studies to find candidates displaying high-binding affinities. The ML model was used to search through thousands of approved ligands by the Food and Drug Administration (FDA) and a million biomolecules in the BindingDB database.[514] From these, insights were obtained for more than 19 000 molecules satisfying the Vina score (i.e., an important physicochemical measure of the therapeutic process of a molecule that is used to rank molecular conformations and predict free energy of binding). Figure 17 shows the Vina score predictions that led to the selection of the best candidates, some of which are also illustrated in the figure. The Vina scores for the top ligands were further confirmed using expensive docking approaches, resulting in the identification of 75 FDA-approved and 100 other ligands potentially useful to treat SARS-CoV-2. This study highlights a reasonable CompChem+ML strategy for making useful suggestions to aid expert biologists and medical professionals to focus in fewer candidates when performing either robust CompChem efforts or synthesis and trial experiments.

6. CONCLUSIONS AND OUTLOOK

Recent CompChem methods, algorithms, and codes have empowered new studies for a wealth of physical and chemical insights into molecules and materials. Today, the combination of CompChem+ML can be equipped to address new and more challenging questions in different domains of physics, materials science, chemistry, biology, and medicine. Productive research efforts in this direction necessitate interdisciplinary teams and increasing availability of high-quality data across appropriate regions of chemical compound space. Discovering new chemicals and materials requires thorough investigations. One needs to predict reaction pathways and interactions between molecules, optimize environmental conditions for catalytic reactions, enhance selectivities that eliminate undesired side reactions or side effects, and navigate other system-specific degrees of freedom. Addressing this complexity calls for a statistical view on chemical design and discovery, and CompChem+ML provides a natural synergy for obtaining predictive insights to lead to wisdom and impact.

This Review provided a bird's-eye view of CompChem and ML and how they can be used together to make transformative impacts in the chemical sciences. The successes of CompChem+ML are particularly visible in physical chemistry and include drastic acceleration of molecular and materials modeling, discovery and prediction of chemicals with desired properties,

prediction of reaction pathways, and design of new catalysts and drug candidates. Nevertheless, we have only begun to scratch the surface of how successful applications of ML in chemistry can bring impact. There are many conceptual, theoretical, and practical challenges waiting to be solved to enable further synergies within the *troika* of CompChem, ML, and CPI. Here we enumerate some of the challenges that we consider to be the most pressing and interesting at this moment:

1. *Reliance on ML in CompChem algorithms must be increased*: ML algorithms can be integrated into CompChem algorithms at almost any simulation level (Figure 3). ML algorithms are already available to accelerate calculations of CompChem energies, navigations along reaction pathways, and sampling of larger regions of the PES, but the reluctance of their use impedes progress. In general, these algorithms must be made more effective, efficient, accessible, user-friendly, and reproducible to benefit fundamental and applied research (see for example, ref 719.).

2. *More general ML approaches are needed*: ML methods must continue to evolve beyond now-common applications of learning a narrow region of a PES or identifying straightforward structure/property relationships. New ML methods should have the capacity to predict energetic and electronic properties and their more convoluted relationships across chemical space. Such approaches should grow toward uniformly describing compositional (chemical arrangement of atoms in a molecule) and configurational (physical arrangement of atoms in space) degrees of freedom on equal footing. Further progress in this field requires developing new universal ML models suitable for insights across diverse systems and physicochemical properties.

3. *ML representations must include the right physics*: ML methods that are claimed to be accurate but incorrectly describe the true physics of a system will eventually fail to achieve meaningful insights while lowering the reputation of other work in the field. Current ML representations (descriptors) can successfully describe local chemical bonding, but few if any are treating long-range electrostatics, polarization, and van der Waals dispersion interactions that are critical for rationalizing physical systems, both large and small. Combining intermolecular interaction theory (a key focus of advanced CompChem methods) with ML is an important direction for future progress toward studying complex molecular systems.

4. *CompChem + ML applications need to strive toward achieving realistic complexity*: Investigations using highly accurate CompChem methods normally require overly simplified model systems while more realistic model systems necessitate less accurate but computationally efficient CompChem methods. This compromise should no longer be necessary. We are due for a paradigm shift in how thermodynamics, kinetics, and dynamics of systems in complex chemical environments (e.g., for multiscale biological processes like drug design and/or catalytic processes at solid−liquid interfaces under photochemical excitations, etc.) can be treated more faithfully with less corner-cutting. An emerging idea is to dispatch ML approaches into computationally efficient

model Hamiltonians for electronic interactions based on correlated wavefunction, KS-DFT, tight-binding, molecular orbital techniques, and/or the many-body dispersion method. ML can predict Hamiltonian parameters and the quantum-mechanical observables would be calculated via diagonalization of the corresponding Hamiltonian. The challenge is to find an appropriate balance between prediction accuracy and computational efficiency to dramatically enhance larger scale simulations.

5. *Much more experimental data is needed*: Validations of ML predictions require extensive comparisons with experimental observables such as reaction rates, spectroscopic observations, solvation energies, and melting temperatures. Such experiments may have previously been considered too routine, too mundane, or not insightful enough alone, but all high quality brings great value for future CompChem+ML efforts that tightly integrate quantum mechanics, statistical simulations, and fast ML predictions, all within a comprehensive molecular simulation framework.[720]

6. *Much more comprehensive data sets need to be assembled and curated*: Current CompChem+ML efforts have profited heavily by the availability of benchmark data sets for relatively small molecules that allow a comparison of existing models.[413,527] While efforts fixated on boosting prediction accuracies and shrinking down requisite training set sizes for ML models have had their merits, it is time to move on as further improvements are meaningless if the ML models are not making useful and insightful predictions themselves. More useful predictions will require knowledge from larger data sets, and these will inevitably contain heterogeneous combinations of different levels of theory or experiments that must be analyzed, "cleaned", and uncertainties adequately quantified for models to productively learn. Such hybrid data sets may be the key to arrive at novel hypotheses in chemistry that could then be experimentally tested.

7. *Bolder and deeper explorations of chemical space are needed*: So far most efforts to generate chemical data have focused on exploring parts of chemical space for new compounds for a targeted purpose. This should change. Combining ML model uncertainty estimates across broader swaths of chemical space could open pathways for fruitful statistical explorations, say, in an active learning framework. This could lead to discovering new synergies between data that otherwise would not have been possible to enable advances in scientific understanding and improve ML models. Generative models can bridge the gap between sampling and targeted structure generation imposing optimal compound properties, for example, for inverse chemical design.[125,621,622]

This and other reviews[20,554,634,720−724] have stated how ML has become instrumental for recent progress in CompChem. We would like to also mention inspirations that ML has drawn from being applied to physical and chemical problems.

ML methods generally assume that data is subject to measurement noise while CompChem data is generally approximate but also noise-free from a statistical perspective. ML modeling still requires regularization, but regularizers should reflect the underlying physics of molecular and

materials systems. ML models used in applications of vision contain discrete convolution filters that are suboptimal for chemical modeling, but recognition of this shortcoming has led to novel continuous convolution filters that are well suited for chemistry and have also become a popular novel architecture for core ML methods.[434]

Furthermore, invariances, symmetries, and conservation laws are key ingredients to physical and chemical systems. Incorporating them into ML has led to novel and useful models for chemistry since they can learn from significantly less data, which then makes it possible to build force fields at unprecedentedly high levels of theory.[206,207,372] Using these powerful ML techniques for computer vision, natural language processing, and other applications is currently being explored. Structural information from molecular graphs provide the basis for novel tensor NNs or message passing architectures,[352,421] as well as graph explanation methods.[725]

Many further challenges exist that have led or will lead to mutual *bidirectional* cross-fertilization between ML and chemistry. These interdisciplinary efforts also initiate progress in respective application domains. The power of this path is that solving a burning problem in chemistry with a novel crafted ML model may also result in unforeseen insights in how to better design core ML methods. Interestingly, the exploratory usage of ML for knowledge discovery in chemistry typically requires novel ML models and unforeseen scientific innovations, and this can lead to interesting insight that is not necessary limited to chemistry alone, rather it is likely to go beyond.

To conclude, the past decade has shown that it has not been enough to *just* apply existing ML algorithms, but breakthroughs are happening by a handshaking of innovations resulting in novel ML algorithms and architectures driven by the pursuit of novel insights in chemistry *while* retaining a deep understanding about the underlying physical and chemical principles. Research programs that foster interdisciplinary exchange, such as IPAM (www.ipam.ucla.edu), have seeded this progress, and these should be continued. Mixed teams with members educated in different aspects of physics, chemistry and ML have been instrumental. This also brings the need to solve the new educational challenge of developing new generations of researchers with an academic curriculum that interweaves chemistry, physics and computer science to enable a meaningful (multilingual) research contribution to this exciting emerging field.

AUTHOR INFORMATION

Corresponding Authors

John A. Keith − *Department of Chemical and Petroleum Engineering Swanson School of Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, United States;* ⊙ orcid.org/0000-0002-6583-6322; Email: jakeith@pitt.edu

Klaus-Robert Müller − *Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany; Department of Artificial Intelligence, Korea University, Seoul 02841, Korea; Max-Planck-Institut für Informatik, 66123 Saarbrücken, Germany; Google Research, Brain Team, Berlin, Germany;* ⊙ orcid.org/0000-0002-3861-7685; Email: klaus-robert.mueller@tu-berlin.de

Alexandre Tkatchenko − *Department of Physics and Materials Science, University of Luxembourg, L-1511*
Luxembourg City, Luxembourg; ⊙ orcid.org/0000-0002-1012-4854; Email: alexandre.tkatchenko@uni.lu

Authors

Valentin Vassilev-Galindo − *Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg;* ⊙ orcid.org/0000-0001-7532-3590

Bingqing Cheng − *Accelerate Programme for Scientific Discovery, Department of Computer Science and Technology, Cambridge CB3 0FD, United Kingdom;* ⊙ orcid.org/0000-0002-3584-9632

Stefan Chmiela − *Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, 10587 Berlin, Germany*

Michael Gastegger − *Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, 10587 Berlin, Germany*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.chemrev.1c00107

Notes

Biographies

John A. Keith is an associate professor and R.K. Mellon Faculty Fellow in Energy at the University of Pittsburgh in the department of chemical and petroleum engineering. He obtained his bachelors' in chemistry at Wesleyan University and a Ph.D. degree in computational chemistry at Caltech in 2007. After an Alexander von Humboldt postdoctoral fellowship at the Universität Ulm, he was an Associate Research Scholar at Princeton University. He was a recipient of an NSF-CAREER award in 2017. His research interests lie in the applications and development of computational chemistry for engineering chemical reactions and materials for electrocatalysis, anticorrosion coatings, and the development of chemicals having less of an environmental footprint. He was a recipient of a Luxembourg Science Foundation INTER Mobility award in 2019−2020 to do a research sabbatical in Prof. Alexandre Tkatchenko's group at the University of Luxembourg. This Review is a primary product of that visit.

Valentin Vassilev-Galindo graduated with honors from University of Veracruz (Mexico) with a Bachelor's degree in Chemical Engineering in 2014. Then, he enrolled to the Master program in Physical Chemistry at Cinvestav-Mérida (Mexico), where he worked under the supervision of Professor Gabriel Merino until receiving the MSc. degree in 2017. He is currently pursuing a PhD degree at the University of Luxembourg in the research group of Professor Alexandre Tkatchenko. His research is mainly related to machine learning potentials.

Bingqing Cheng is a Departmental Early Career Fellow at the Computer Laboratory, University of Cambridge, and a Junior Research Fellow at Trinity College. She received her Ph.D. from the École Polytechnique Fédérale de Lausanne (EPFL) in 2019. Her work focuses on theoretical predictions of material properties.

Stefan Chmiela is a senior researcher at the Berlin Institute for the Foundations of Learning and Data (BIFOLD). He received his Ph.D. from Technische Universität Berlin in 2019. His research interests include Hilbert space learning methods for applications in quantum chemistry, with particular focus on data efficiency and robustness.

Michael Gastegger is a postdoctoral researcher in the BASLEARN project of the Machine Learning Group at Technische Universität Berlin. He received his Ph.D. in Chemistry from the University of Vienna in Austria in 2017. His research interests include the development of machine learning methods for quantum chemistry and their application in simulations.

Klaus-Robert Müller has been a professor of computer science at Technische Universität Berlin since 2006; at the same time he is directing and codirecting the Berlin Machine Learning Center and the Berlin Big Data Center, respectively. He studied physics in Karlsruhe from 1984 to 1989 and obtained his Ph.D. degree in computer science at Technische Universität Karlsruhe in 1992. After completing a postdoctoral position at GMD FIRST in Berlin, he was a research fellow at the University of Tokyo from 1994 to 1995. In 1995, he founded the Intelligent Data Analysis group at GMD-FIRST (later Fraunhofer FIRST) and directed it until 2008. From 1999 to 2006, he was a professor at the University of Potsdam. He was awarded the Olympus Prize for Pattern Recognition (1999), the SEL Alcatel Communication Award (2006), the Science Prize of Berlin by the Governing Mayor of Berlin (2014), and the Vodafone Innovations Award (2017). In 2012, he was elected member of the German National Academy of Sciences-Leopoldina; in 2017, a member of the Berlin Brandenburg Academy of Sciences; and also in 2017, an external scientific member of the Max Planck Society. In 2019 and 2020, he became a Highly Cited researcher in the cross-disciplinary area. His research interests are intelligent data analysis and Machine Learning in the sciences (Neuroscience (specifically Brain-Computer Interfaces), Physics, Chemistry) and in industry.

Alexandre Tkatchenko is a Professor of Theoretical Chemical Physics at the University of Luxembourg and Visiting Professor at Technische Universität Berlin. He obtained his bachelor degree in Computer Science and a Ph.D. in Physical Chemistry at the Universidad Autonoma Metropolitana in Mexico City. Between 2008 and 2010, he was an Alexander von Humboldt Fellow at the Fritz Haber Institute of the Max Planck Society in Berlin. Between 2011 and 2016, he led an independent research group at the same institute. Tkatchenko serves on editorial boards of two society journals: *Physical Review Letters* (APS) and *Science Advances* (AAAS). He received a number of awards, including elected Fellow of the American Physical Society, the 2020 Dirac Medal from WATOC, the Gerhard Ertl Young Investigator Award of the German Physical Society, and two flagship grants from the European Research Council: a Starting Grant in 2011 and a Consolidator Grant in 2017. His group pushes the boundaries of quantum mechanics, statistical mechanics, and machine learning to develop efficient methods to enable accurate modeling and obtain new insights into complex materials.

ACKNOWLEDGMENTS

ACRONYMS

ACE	atomic cluster expansion
ACS	American Chemical Society
ACSF	atom-centered symmetry function
AE	autoencoders
AI	artificial intelligence
API	application programming interfaces
BoB	bag of bonds
BOP	bond order potential
BP	back-propagation
CGCNN	crystal graph convolution neural network
CNN	convolutional neural network
COSMO	conductor-like screening model
C-PCM	conductor polarizable continuum solvent model
CASPT2	complete active space perturbation theory
CASSCF	complete active space self-consistent field
CBS	complete basis set
CI	configuration interaction
CMD	centroid molecular dynamics
CompChem	computational chemistry
CPI	chemical and physical intuition
CV	collective variable
DDR1	discoidin domain receptor 1
DeepPot-SE	smooth edition version of the DeepMD potential
D-PCM	dielectric polarizable continuum solvent model
DFT	density-functional theory
DFTB	density functional tight binding
DLPNO	domain-based local pair natural orbital
DMRG	density matrix renormalization group theory
DTNN	deep tensor neural network
EAM	embedded atom method
EANN	embedded atom neural network
ECP	effective core potential
FCHL	Faber–Christensen–Huang–Lilienfeld
FCI	full configuration interaction
FDA	Food and Drug Administration
FES	free energy surface
FF	force field
FPS	farthest point sampling
GAN	generative adversarial network
GENTRL	generative tensorial reinforcement learning
GGA	generalized gradient approximation
GP	Gaussian processes
GPU	graphical processing units
GVB	generalized valence bond
HEAT	high accuracy extrapolated ab initio thermochemistry
HF	Hartree–Fock
HIP-NN	hierarchical interacting particle neural network
ICA	independent component analysis

IEFPCM	integral equation formulation of polarizable continuum solvent model
KE	kinetic energy
KPCA	kernel principal component analysis
KRR	kernel ridge regression
KS	Kohn–Sham
LDA	local density approximation
LJ	Lennard-Jones
MBTR	many-body tensor representation
MD	molecular dynamics
MEAM	modified embedded atom method
ML	machine learning
MLP	machine learning potential
MPNN	message-passing neural network
MRCC	multireference coupled cluster
MRCI	multireference configuration interaction
MS/MS	tandem mass spectroscopy
NDDO	neglect of diatomic differential overlap
NEB	nudged elastic band
NMR	nuclear magnetic resonance
NN	neural network
NQE	nuclear quantum effect
OF	orbital-free
OLED	organic light-emitting diode
PCA	principal component analysis
PCM	polarizable continuum solvent model
PES	potential energy surface
PIMD	path integral molecular dynamics
QM	quantum mechanics
QSAR/QSPR	quantitative structure activity/property relationship
RE-Match	regularized entropy match
RI	resolution of the identity
RISM	reference interaction site model
RL	reinforcement learning
RMSD	root mean squared displacement
RNN	recurrent neural network
SCRF	self-consistent reaction field
SOAP	smooth overlap of atomic positions
STM	scanning tunneling microscopy
SVM	support vector machine
t-SNE	t-distributed stochastic neighbor embedding
TD	time-dependent
UMAP	uniform manifold approximation and projection
XAI	explainable artificial intelligence
XANES	X-ray absorption near edge structure

REFERENCES

(1) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.

(2) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85–117.

(3) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, 2016; http://www.deeplearningbook.org.

(4) Capper, D.; Jones, D. T.; Sill, M.; Hovestadt, V.; Schrimpf, D.; Sturm, D.; Koelsche, C.; Sahm, F.; Chavez, L.; Reuss, D. E.; et al. DNA Methylation-Based Classification of Central Nervous System Tumours. *Nature* **2018**, *555*, 469–474.

(5) Klauschen, F.; Müller, K.-R.; Binder, A.; Bockmayr, M.; Hägele, M.; Seegerer, P.; Wienert, S.; Pruneri, G.; de Maria, S.; Badve, S.; et al. Scoring of Tumor-Infiltrating Lymphocytes: From Visual Estimation to Machine Learning. *Semin. Cancer Biol.* **2018**, *52*, 151–157.

(6) Jurmeister, P.; Bockmayr, M.; Seegerer, P.; Bockmayr, T.; Treue, D.; Montavon, G.; Vollbrecht, C.; Arnold, A.; Teichmann, D.; Bressem, K.; et al. Machine Learning Analysis of DNA Methylation Profiles Distinguishes Primary Lung Squamous Cell Carcinomas From Head and Neck Metastases. *Sci. Transl. Med.* **2019**, *11*, No. eaaw8513.

(7) Ardila, D.; Kiraly, A. P.; Bharadwaj, S.; Choi, B.; Reicher, J. J.; Peng, L.; Tse, D.; Etemadi, M.; Ye, W.; Corrado, G.; et al. End-to-End Lung Cancer Screening With Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography. *Nat. Med.* **2019**, *25*, 954–961.

(8) Binder, A.; Bockmayr, M.; Hagele, M.; Wienert, S.; Heim, D.; Hellweg, K.; Ishii, M.; Stenzinger, A.; Hocke, A.; Denkert, C.; et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nat. Mach. Intel.* **2021**, *3*, 355–366.

(9) Baldi, P.; Sadowski, P.; Whiteson, D. Searching for Exotic Particles in High-Energy Physics With Deep Learning. *Nat. Commun.* **2014**, *5*, 4308.

(10) Leinen, P.; Esders, M.; Schütt, K. T.; Wagner, C.; Müller, K.-R.; Tautz, F. S. Autonomous Robotic Nanofabrication With Reinforcement Learning. *Sci. Adv.* **2020**, *6*, No. eabb6987.

(11) Lengauer, T.; Sander, O.; Sierra, S.; Thielen, A.; Kaiser, R. Bioinformatics Prediction of HIV Coreceptor Usage. *Nat. Biotechnol.* **2007**, *25*, 1407–1410.

(12) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W.; Bridgland, A.; et al. Improved Protein Structure Prediction Using Potentials From Deep Learning. *Nature* **2020**, *577*, 706–710.

(13) Blankertz, B.; Tomioka, R.; Lemm, S.; Kawanabe, M.; Muller, K.-R. Optimizing Spatial Filters for Robust EEG Single-Trial Analysis. *IEEE Signal Process. Mag.* **2008**, *25*, 41–56.

(14) Perozzi, B.; Al-Rfou, R.; Skiena, S. DeepWalk: Online Learning of Social Representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; New York, NY, USA, 2014; pp 701–710.

(15) Thrun, S.; Burgard, W.; Fox, D. *Probabilistic Robotics*; MIT Press: Cambridge, MA, 2005.

(16) Won, D.-O.; Müller, K.-R.; Lee, S.-W. An Adaptive Deep Reinforcement Learning Framework Enables Curling Robots With Human-Like Performance in Real World Conditions. *Sci. Robot.* **2020**, *5*, No. eabb9764.

(17) Lewis, M. M. *Moneyball: The Art of Winning an Unfair Game*; W. W. Norton: New York, N.Y., 2003.

(18) Ferrucci, D.; Levas, A.; Bagchi, S.; Gondek, D.; Mueller, E. T. Watson: Beyond Jeopardy! *Artif. Intell.* **2013**, *199*, 93–105.

(19) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the Game of Go With Deep Neural Networks and Tree Search. *Nature* **2016**, *529*, 484–489.

(20) Tkatchenko, A. Machine Learning for Chemical Discovery. *Nat. Commun.* **2020**, *11*, 4125.

(21) Rowley, J. The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *J. Inf. Sci.* **2007**, *33*, 163–180.

(22) Box, G. E. Science and Statistics. *J. Am. Stat. Assoc.* **1976**, *71*, 791–799.

(23) McQuarrie, D.; Simon, J. *Physical Chemistry: A Molecular Approach*; University Science Books: Sausalito, CA, 1997.

(24) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2004.

(25) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press: New York, NY, 2002.

(26) Foresman, J.; Frisch, A.; Gaussian, I. *Exploring Chemistry With Electronic Structure Methods*; Gaussian, Inc.: Pittsburgh, PA, 1996.

(27) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; et al. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13*, No. e1005659.

(28) Oyeyemi, V. B.; Keith, J. A.; Carter, E. A. Accurate Bond Energies of Biodiesel Methyl Esters From Multireference Averaged Coupled-Pair Functional Calculations. *J. Phys. Chem. A* **2014**, *118*, 7392−7403.

(29) Anslyn, E.; Dougherty, D. *Modern Physical Organic Chemistry*; University Science Books: Sausalito, CA, 2006.

(30) Glazer, A. The Classification of Tilted Octahedra in Perovskites. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1972**, *28*, 3384−3392.

(31) Giessibl, F. J. Atomic Resolution of the Silicon (111)-(7 × 7) Surface by Atomic Force Microscopy. *Science* **1995**, *267*, 68−71.

(32) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-3 and Density Functional Theories for a Larger Experimental Test Set. *J. Chem. Phys.* **2000**, *112*, 7374.

(33) Haunschild, R.; Klopper, W. New Accurate Reference Energies for the G2/97 Test Set. *J. Chem. Phys.* **2012**, *136*, 164102.

(34) Taylor, P. R. *European Summer School in Quantum Chemistry*; Springer, Berlin, 1994; Vol. 125; pp 125−202.

(35) Bartlett, R. J. Many-Body Perturbation Theory and Coupled Cluster Theory for Electron Correlation in Molecules. *Annu. Rev. Phys. Chem.* **1981**, *32*, 359−401.

(36) Stöhr, M.; Van Voorhis, T.; Tkatchenko, A. Theory and Practice of Modeling Van Der Waals Interactions in Electronic-Structure Calculations. *Chem. Soc. Rev.* **2019**, *48*, 4118−4154.

(37) Lundberg, M.; Siegbahn, P. E. Quantifying the Effects of the Self-Interaction Error in DFT: When Do the Delocalized States Appear? *J. Chem. Phys.* **2005**, *122*, 224103.

(38) Morgante, P.; Peverati, R. The Devil in the Details: A Tutorial Review on Some Undervalued Aspects of Density Functional Theory Calculations. *Int. J. Quantum Chem.* **2020**, *120*, No. 26332.

(39) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648.

(40) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(41) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A Look at the Density Functional Theory Zoo With the Advanced GMTKN55 Database for General Main Group Thermochemistry, Kinetics and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184−32215.

(42) Zhao, Y.; González-García, N.; Truhlar, D. G. Benchmark Database of Barrier Heights for Heavy Atom Transfer, Nucleophilic Substitution, Association, and Unimolecular Reactions and Its Use to Test Theoretical Methods. *J. Phys. Chem. A* **2005**, *109*, 2012−2018.

(43) Sim, E.; Song, S.; Burke, K. Quantifying Density Errors in DFT. *J. Phys. Chem. Lett.* **2018**, *9*, 6385−6392.

(44) Riley, K. E.; Op't Holt, B. T.; Merz, K. M. Critical Assessment of the Performance of Density Functional Methods for Several Atomic and Molecular Properties. *J. Chem. Theory Comput.* **2007**, *3*, 407−433.

(45) Maldonado, A. M.; Hagiwara, S.; Choi, T. H.; Eckert, F.; Schwarz, K.; Sundararaman, R.; Otani, M.; Keith, J. A. Quantifying Uncertainties in Solvation Procedures for Modeling Aqueous Phase Reaction Mechanisms. *J. Phys. Chem. A* **2021**, *125*, 154−164.

(46) Abraham, M. J.; Apostolov, R. P.; Barnoud, J.; Bauer, P.; Blau, C.; Bonvin, A. M.; Chavent, M.; Chodera, J. D.; Condić-Jurkić, K.; Delemotte, L.; et al. Sharing Data From Molecular Simulations. *J. Chem. Inf. Model.* **2019**, *59*, 4093−4099.

(47) Wheeler, S. E.; Houk, K. N. Integration Grid Errors for Meta-Gga-Predicted Reaction Energies: Origin of Grid Errors for the M06 Suite of Functionals. *J. Chem. Theory Comput.* **2010**, *6*, 395−404.

(48) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; et al. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16*, 670−673.

(49) Lejaeghere, K.; Bihlmayer, G.; Björkman, T.; Blaha, P.; Blügel, S.; Blum, V.; Caliste, D.; Castelli, I. E.; Clark, S. J.; Dal Corso, A.; et al. Reproducibility in Density Functional Theory Calculations of Solids. *Science* **2016**, *351*, No. aad3000.

(50) Sonnenburg, S.; Braun, M. L.; Ong, C. S.; Bengio, S.; Bottou, L.; Holmes, G.; LeCun, Y.; Müller, K.- R.; Pereira, F.; Rasmussen, C. E.; et al. The Need for Open Source Software in Machine Learning. *J. Mach. Learn. Res.* **2007**, *8*, 2443−2466.

(51) Durrani, J. Computational Chemistry Faces a Coding Crisis. *Chemistry World*, 2020. https://www.chemistryworld.com/news/chemistrys--reproducibility--crisis--that--youve--probably--never--heard--of/4011693.article#/.

(52) Perkel, J. M. Challenge to Scientists: Does Your Ten-Year-Old Code Still Run? *Nature* **2020**, *584*, 656−658.

(53) Govoni, M.; Munakami, M.; Tanikanti, A.; Skone, J. H.; Runesha, H. B.; Giberti, F.; de Pablo, J.; Galli, G. Qresp, a Tool for Curating, Discovering and Exploring Reproducible Scientific Papers. *Sci. Data* **2019**, *6*, 190002.

(54) Kitchin, J. R. Examples of Effective Data Sharing in Scientific Publishing. *ACS Catal.* **2015**, *5*, 3894−3899.

(55) Álvarez-Moreno, M.; De Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the Computational Chemistry Big Data Problem: The ioChem-BD Platform. *J. Chem. Inf. Model.* **2015**, *55*, 95−103.

(56) Huber, S. P.; Zoupanos, S.; Uhrin, M.; Talirz, L.; Kahle, L.; Häuselmann, R.; Gresch, D.; Müller, T.; Yakutovich, A. V.; Andersen, C. W.; et al. AiiDA 1.0, a Scalable Computational Infrastructure for Automated Reproducible Workflows and Data Provenance. *Sci. Data* **2020**, *7*, 300.

(57) Heidrich, D.; Quapp, W. Saddle Points of Index 2 on Potential Energy Surfaces and Their Role in Theoretical Reactivity Investigations. *Theor. Chim. Acta* **1986**, *70*, 89−98.

(58) Ess, D. H.; Wheeler, S. E.; Iafe, R. G.; Xu, L.; Çelebi-Ölçüm, N.; Houk, K. N. Bifurcations on Potential Energy Surfaces of Organic Reactions. *Angew. Chem., Int. Ed.* **2008**, *47*, 7592−7601.

(59) Tarczay, G.; Császár, A. G.; Klopper, W.; Quiney, H. M. Anatomy of Relativistic Energy Corrections in Light Molecular Systems. *Mol. Phys.* **2001**, *99*, 1769−1794.

(60) Perdew, J. P.; Schmidt, K. Jacob's Ladder of Density Functional Approximations for the Exchange-Correlation Energy. *AIP Conference Proceedings.* **2000**, 1−20.

(61) Schrödinger, E. Quantisierung Als Eigenwertproblem (Erste Mitteilung). *Ann. Phys.* **1926**, *384*, 361−376.

(62) Schrödinger, E. Quantisierung Als Eigenwertproblem (Zweite Mitteilung). *Ann. Phys.* **1926**, *384*, 489−527.

(63) Schrödinger, E. Quantisierung Als Eigenwertproblem (Vierte Mitteilung). *Ann. Phys.* **1926**, *386*, 109−139.

(64) Born, M.; Oppenheimer, R. Zur Quantentheorie Der Molekeln. *Ann. Phys.* **1927**, *389*, 457−484.

(65) Curchod, B. F.; Martínez, T. J. Ab Initio Nonadiabatic Quantum Molecular Dynamics. *Chem. Rev.* **2018**, *118*, 3305−3336.

(66) Pavošević, F.; Culpitt, T.; Hammes-Schiffer, S. Multicomponent Quantum Chemistry: Integrating Electronic and Nuclear Quantum Effects via the Nuclear-Electronic Orbital Method. *Chem. Rev.* **2020**, *120*, 4222−4253.

(67) Peterson, K. A.; Dunning, T. H. Accurate Correlation Consistent Basis Sets for Molecular Core-Valence Correlation Effects: The Second Row Atoms Al-Ar, and the First Row Atoms B-Ne Revisited. *J. Chem. Phys.* **2002**, *117*, 10548.

(68) Hehre, W. J.; Stewart, R. F.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *J. Chem. Phys.* **1969**, *51*, 2657.

(69) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets for Atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571.

(70) Van Lenthe, E.; Baerends, E. J. Optimized Slater-Type Basis Sets for the Elements 1−118. *J. Comput. Chem.* **2003**, *24*, 1142−1156.

(71) Slater, J. C. Energy Band Calculations by the Augmented Plane Wave Method. *Adv. Quantum Chem.* **1964**, *1*, 35−58.

(72) MacDonald, A. H.; Picket, W. E.; Koelling, D. D. A Linearised Relativistic Augmented-Plane-Wave Method Utilising Approximate Pure Spin Basis Functions. *J. Phys. C: Solid State Phys.* **1980**, *13*, 2675.

(73) Louie, S. G.; Ho, K. M.; Cohen, M. L. Self-Consistent Mixed-Basis Approach to the Electronic Structure of Solids. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1979**, *19*, 1774.

(74) Goedecker, S.; Teter, M.; Hutter, J. Separable Dual-Space Gaussian Pseudopotentials. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 1703.

(75) Melius, C. F.; Goddard, W. A. Ab Initio Effective Potentials for Use in Molecular Quantum Mechanics. *Phys. Rev. A: At., Mol., Opt. Phys.* **1974**, *10*, 1528.

(76) Wadt, W. R.; Hay, P. J. Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for Main Group Elements Na to Bi. *J. Chem. Phys.* **1985**, *82*, 284.

(77) Hay, P. J.; Wadt, W. R. Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for the Transition Metal Atoms Sc to Hg. *J. Chem. Phys.* **1985**, *82*, 270.

(78) Cao, X.; Dolg, M. Segmented Contraction Scheme for Small-Core Lanthanide Pseudopotential Basis Sets. *J. Mol. Struct.: THEOCHEM* **2002**, *581*, 139−147.

(79) Metz, B.; Stoll, H.; Dolg, M. Small-Core Multiconfiguration-Dirac-Hartree-Fock-Adjusted Pseudopotentials for Post-D Main Group Elements: Application to PbH and PbO. *J. Chem. Phys.* **2000**, *113*, 2563.

(80) Dolg, M. In *Handbook of Relativistic Quantum Chemistry*; Liu, W., Ed.; Springer: Berlin, 2016; pp 449−478.

(81) Shaw, R. W.; Harrison, W. A. Reformulation of the Screened Heine-Abarenkov Model Potential. *Phys. Rev.* **1967**, *163*, 604.

(82) Kahn, L. R.; Baybutt, P.; Truhlar, D. G. Ab Initio Effective Core Potentials: Reduction of All-Electron Molecular Structure Calculations to Calculations Involving Only Valence Electrons. *J. Chem. Phys.* **1976**, *65*, 3826.

(83) Christiansen, P. A.; Lee, Y. S.; Pitzer, K. S. Improved Ab Initio Effective Core Potentials for Molecular Calculations. *J. Chem. Phys.* **1979**, *71*, 4445.

(84) Hamann, D. R.; Schlüter, M.; Chiang, C. Norm-Conserving Pseudopotentials. *Phys. Rev. Lett.* **1979**, *43*, 1494.

(85) Vanderbilt, D. Optimally Smooth Norm-Conserving Pseudopotentials. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1985**, *32*, 8412.

(86) Garrity, K. F.; Bennett, J. W.; Rabe, K. M.; Vanderbilt, D. Pseudopotentials for High-Throughput DFT Calculations. *Comput. Mater. Sci.* **2014**, *81*, 446−452.

(87) Kresse, G.; Hafner, J. Norm-Conserving and Ultrasoft Pseudopotentials for First-Row and Transition Elements. *J. Phys.: Condens. Matter* **1994**, *6*, 8245.

(88) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 1758.

(89) Troullier, N.; Martins, J. A Straightforward Method for Generating Soft Transferable Pseudopotentials. *Solid State Commun.* **1990**, *74*, 613−616.

(90) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. Systematically Convergent Basis Sets With Relativistic Pseudopotentials. II. Small-Core Pseudopotentials and Correlation Consistent Basis Sets for the Post-D Group 16−18 Elements. *J. Chem. Phys.* **2003**, *119*, 11113.

(91) Roy, L. E.; Hay, P. J.; Martin, R. L. Revised Basis Sets for the LANL Effective Core Potentials. *J. Chem. Theory Comput.* **2008**, *4*, 1029−1031.

(92) Pyykkö, P. Relativistic Effects in Chemistry: More Common Than You Thought. *Annu. Rev. Phys. Chem.* **2012**, *63*, 45−64.

(93) Dirac, P. A. M. The Quantum Theory of the Electron. *Proc. R. Soc. London A* **1928**, *117*, 610−624.

(94) Tecmer, P.; Boguslawski, K.; Kędziera, D. In *Handbook of Computational Chemistry*; Leszczynski, J., Ed.; Springer: Dordrecht, 2016; pp 1−43.

(95) Feynman, R. *Quantum Electrodynamics*; CRC Press: Boca Raton, FL, 2018.

(96) Nakajima, T.; Hirao, K. The Douglas-Kroll-Hess Approach. *Chem. Rev.* **2012**, *112*, 385−402.

(97) Van Lenthe, J. H.; Faas, S.; Snijders, J. G. Gradients in the Ab Initio Scalar Zeroth-Order Regular Approximation (ZORA) Approach. *Chem. Phys. Lett.* **2000**, *328*, 107−112.

(98) Visscher, L. The Dirac Equation in Quantum Chemistry: Strategies to Overcome the Current Computational Problems. *J. Comput. Chem.* **2002**, *23*, 759−766.

(99) Hartree, D. R.; Hartree, W. Self-Consistent Field, With Exchange, for Beryllium. *Proc. R. Soc. London A* **1935**, *150*, 9−33.

(100) Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81*, 385.

(101) Fock, V. Näherungsmethode Zur Lösung Des Quantenmechanischen Mehrkörperproblems. *Eur. Phys. J. A* **1930**, *61*, 126−148.

(102) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2007.

(103) Roothaan, C. C. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23*, 69−89.

(104) Hall, G. G. The Molecular Orbital Theory of Chemical Valency VIII. A Method of Calculating Ionization Potentials. *Proc. R. Soc. London A* **1951**, *205*, 541−552.

(105) Hermann, J.; Schätzle, Z.; Noé, F. Deep-Neural-Network Solution of the Electronic Schrödinger Equation. *Nat. Chem.* **2020**, *12*, 891−897.

(106) Pfau, D.; Spencer, J. S.; Matthews, A. G.; Foulkes, W. M. C. Ab Initio Solution of the Many-Electron Schrödinger Equation With Deep Neural Networks. *Phys. Rev. Res.* **2020**, *2*, 033429.

(107) Eriksen, J. J.; Anderson, T. A.; Deustua, J. E.; Ghanem, K.; Hait, D.; Hoffmann, M. R.; Lee, S.; Levine, D. S.; Magoulas, I.; Shen, J.; et al. The Ground State Electronic Energy of Benzene. *J. Phys. Chem. Lett.* **2020**, *11*, 8922−8929.

(108) Helgaker, T.; Jorgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; John Wiley & Sons, Inc.: Hoboken, NJ, 2013.

(109) Bartlett, R. J.; Musiał, M. Coupled-Cluster Theory in Quantum Chemistry. *Rev. Mod. Phys.* **2007**, *79*, 291−352.

(110) Řezáč, J.; Hobza, P. Describing Noncovalent Interactions Beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, *9*, 2151−2155.

(111) Moran, D.; Simmonett, A. C.; Leach, F. E.; Allen, W. D.; Schleyer, P. V.; Schaefer, H. F. Popular Theoretical Methods Predict Benzene and Arenes to Be Nonplanar. *J. Am. Chem. Soc.* **2006**, *128*, 9342−9343.

(112) Samala, N. R.; Jordan, K. D. Comment on a Spurious Prediction of a Non-Planar Geometry for Benzene at the MP2 Level of Theory. *Chem. Phys. Lett.* **2017**, *669*, 230−232.

(113) Titov, A. V.; Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. Generating Efficient Quantum Chemistry Codes for Novel Architectures. *J. Chem. Theory Comput.* **2013**, *9*, 213−221.

(114) Seritan, S.; Bannwarth, C.; Fales, B. S.; Hohenstein, E. G.; Isborn, C. M.; Kokkila-Schumacher, S. I.; Li, X.; Liu, F.; Luehr, N.; Snyder, J. W.; et al. TeraChem: A Graphical Processing Unit-Accelerated Electronic Structure Package for Large-Scale Ab Initio Molecular Dynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020**, No. e1494.

(115) Anderson, A. G.; Goddard, W. A.; Schröder, P. Quantum Monte Carlo on Graphical Processing Units. *Comput. Phys. Commun.* **2007**, *177*, 298−306.

(116) Andrade, X.; Aspuru-Guzik, A. Real-Space Density Functional Theory on Graphical Processing Units: Computational Approach and Comparison to Gaussian Basis Set Methods. *J. Chem. Theory Comput.* **2013**, *9*, 4360−4373.

(117) Friesner, R. A. Solution of the Hartree-Fock Equations by a Pseudospectral Method: Application to Diatomic Molecules. *J. Chem. Phys.* **1986**, *85*, 1462.

(118) Martinez, T. J.; Mehta, A.; Carter, E. A. Pseudospectral Full Configuration Interaction. *J. Chem. Phys.* **1992**, *97*, 1876.

(119) Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; 2002.

(120) Sierka, M.; Hogekamp, A.; Ahlrichs, R. Fast Evaluation of the Coulomb Potential for Electron Densities Using Multipole Accelerated Resolution of Identity Approximation. *J. Chem. Phys.* **2003**, *118*, 9136−9148.

(121) Riplinger, C.; Neese, F. An Efficient and Near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138*, 034106.

(122) Kong, L.; Bischoff, F. A.; Valeev, E. F. Explicitly Correlated R12/F12 Methods for Electronic Structure. *Chem. Rev.* **2012**, *112*, 75−107.

(123) Austin, B. M.; Zubarev, D. Y.; Lester Jr, W. A. Quantum Monte Carlo and Related Approaches. *Chem. Rev.* **2012**, *112*, 263−288.

(124) Marti, K. H.; Reiher, M. The Density Matrix Renormalization Group Algorithm in Quantum Chemistry. *Z. Phys. Chem.* **2010**, *224*, 583−599.

(125) Schütt, K.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. Unifying Machine Learning and Quantum Chemistry With a Deep Neural Network for Molecular Wavefunctions. *Nat. Commun.* **2019**, *10*, 5024.

(126) McGibbon, R. T.; Taube, A. G.; Donchev, A. G.; Siva, K.; Hernández, F.; Hargus, C.; Law, K. H.; Klepeis, J. L.; Shaw, D. E. Improving the accuracy of Møller-Plesset perturbation theory with neural networks. *J. Chem. Phys.* **2017**, *147*, 161725.

(127) Townsend, J.; Vogiatzis, K. D. Transferable MP2-Based Machine Learning for Accurate Coupled-Cluster Energies. *J. Chem. Theory Comput.* **2020**, *16*, 7453−7461.

(128) Coe, J. P. Machine Learning Configuration Interaction. *J. Chem. Theory Comput.* **2018**, *14*, 5739−5749.

(129) Jeong, W. S.; Stoneburner, S. J.; King, D.; Li, R.; Walker, A.; Lindh, R.; Gagliardi, L. Automation of Active Space Selection for Multireference Methods via Machine Learning on Chemical Bond Dissociation. *J. Chem. Theory Comput.* **2020**, *16*, 2389−2399.

(130) Montgomery, J. A.; Frisch, M. J.; Ochterski, J. W.; Petersson, G. A. A Complete Basis Set Model Chemistry. VII. Use of the Minimum Population Localization Method. *J. Chem. Phys.* **2000**, *112*, 6532−6542.

(131) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) Theory for Molecules Containing First and Second-Row Atoms. *J. Chem. Phys.* **1998**, *109*, 7764−7776.

(132) Karton, A.; Rabinovich, E.; Martin, J. M.; Ruscic, B. W4 Theory for Computational Thermochemistry: In Pursuit of Confident Sub-kJ/Mol Predictions. *J. Chem. Phys.* **2006**, *125*, 144108.

(133) Tajti, A.; Szalay, P. G.; Császár, A. G.; Kállay, M.; Gauss, J.; Valeev, E. F.; Flowers, B. A.; Vázquez, J.; Stanton, J. F. HEAT: High Accuracy Extrapolated Ab Initio Thermochemistry. *J. Chem. Phys.* **2004**, *121*, 11599.

(134) Karton, A. A Computational Chemist's Guide to Accurate Thermochemistry for Organic Molecules. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6*, 292−310.

(135) Zaspel, P.; Huang, B.; Harbrecht, H.; von Lilienfeld, O. A. Boosting Quantum Machine Learning Models With a Multilevel Combination Technique: Pople Diagrams Revisited. *J. Chem. Theory Comput.* **2019**, *15*, 1546−1559.

(136) Park, J. W.; Al-Saadon, R.; Macleod, M. K.; Shiozaki, T.; Vlaisavljevich, B. Multireference Electron Correlation Methods: Journeys Along Potential Energy Surfaces. *Chem. Rev.* **2020**, *120*, 5878−5909.

(137) Hirao, K. Multireference Møller-Plesset Method. *Chem. Phys. Lett.* **1992**, *190*, 374−380.

(138) Buenker, R. J.; Peyerimhoff, S. D.; Butscher, W. Applicability of the Multi-Reference Double-Excitation CI (MRD-CI) Method to the Calculation of Electronic Wavefunctions and Comparison With Related Techniques. *Mol. Phys.* **1978**, *35*, 771−791.

(139) Levine, B. G.; Coe, J. D.; Martínez, T. J. Optimizing Conical Intersections Without Derivative Coupling Vectors: Application to Multistate Multireference Second-Order Perturbation Theory (MS-CASPT2). *J. Phys. Chem. B* **2008**, *112*, 405−413.

(140) Jiang, W.; Deyonker, N. J.; Wilson, A. K. Multireference Character for 3d Transition-Metal-Containing Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 460−468.

(141) Lee, T. J. Comparison of the T1 and D1 Diagnostics for Electronic Structure Theory: A New Definition for the Open-Shell D11 Diagnostic. *Chem. Phys. Lett.* **2003**, *372*, 362−367.

(142) Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J. Data-Driven Approaches Can Overcome the Cost-Accuracy Trade-Off in Multireference Diagnostics. *J. Chem. Theory Comput.* **2020**, *16*, 4373−4387.

(143) Bobrowicz, F. W.; Goddard, W. A. In *Methods of Electronic Structure Theory*; Schaefer, H. F., Ed.; Springer: Boston, MA, 1977; pp 79−127.

(144) Roos, B. O.; Taylor, P. R.; Sigbahn, P. E. A Complete Active Space SCF Method (CASSCF) Using a Density Matrix Formulated Super-Ci Approach. *Chem. Phys.* **1980**, *48*, 157−173.

(145) Szalay, P. G.; Müller, T.; Gidofalvi, G.; Lischka, H.; Shepard, R. Multiconfiguration Self-Consistent Field and Multireference Configuration Interaction Methods and Applications. *Chem. Rev.* **2012**, *112*, 108−181.

(146) Pulay, P. A Perspective on the CASPT2Method. *Int. J. Quantum Chem.* **2011**, *111*, 3273−3279.

(147) Lyakh, D. I.; Musiał, M.; Lotrich, V. F.; Bartlett, R. J. Multireference Nature of Chemistry: The Coupled-Cluster View. *Chem. Rev.* **2012**, *112*, 182−243.

(148) Evangelista, F. A. Perspective: Multireference Coupled Cluster Theories of Dynamical Electron Correlation. *J. Chem. Phys.* **2018**, *149*, 030901.

(149) Jensen, K. P.; Roos, B. O.; Ryde, U. Erratum: O2-Binding to Heme: Electronic Structure and Spectrum of Oxyheme, Studied by Multiconfigurational Methods. *J. Inorg. Biochem.* **2005**, *99* (1), 45−54; *J. Inorg. Biochem.* **2005**, *99*, 978.

(150) Pople, J. A. Two-Dimensional Chart of Quantum Chemistry. *J. Chem. Phys.* **1965**, *43*, S229.

(151) Parr, R.; Weitao, Y. *Density-Functional Theory of Atoms and Molecules*; International Series of Monographs on Chemistry; Oxford University Press: New York, NY, 1994.

(152) Witt, W. C.; Del Rio, B. G.; Dieterich, J. M.; Carter, E. A. Orbital-Free Density Functional Theory for Materials Research. *J. Mater. Res.* **2018**, *33*, 777−795.

(153) Hung, L.; Huang, C.; Shin, I.; Ho, G. S.; Lignères, V. L.; Carter, E. A. Introducing PROFESS 2.0: A Parallelized, Fully Linear Scaling Program for Orbital-Free Density Functional Theory Calculations. *Comput. Phys. Commun.* **2010**, *181*, 2208−2209.

(154) Mi, W.; Shao, X.; Su, C.; Zhou, Y.; Zhang, S.; Li, Q.; Wang, H.; Zhang, L.; Miao, M.; Wang, Y.; et al. ATLAS: A Real-Space Finite-Difference Implementation of Orbital-Free Density Functional Theory. *Comput. Phys. Commun.* **2016**, *200*, 87−95.

(155) Mi, W.; Genova, A.; Pavanello, M. Nonlocal Kinetic Energy Functionals by Functional Integration. *J. Chem. Phys.* **2018**, *148*, 184107.

(156) Ayers, P. W. Generalized Density Functional Theories Using the K -Electron Densities: Development of Kinetic Energy Functionals. *J. Math. Phys.* **2005**, *46*, 062107.

(157) Huang, C.; Carter, E. A. Nonlocal Orbital-Free Kinetic Energy Density Functional for Semiconductors. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2010**, *81*, 045206.

(158) Burakovsky, L.; Ticknor, C.; Kress, J. D.; Collins, L. A.; Lambert, F. Transport Properties of Lithium Hydride at Extreme Conditions From Orbital-Free Molecular Dynamics. *Phys. Rev. E* **2013**, *87*, 023104.

(159) Sjostrom, T.; Daligault, J. Ionic and Electronic Transport Properties in Dense Plasmas by Orbital-Free Density Functional Theory. *Phys. Rev. E* **2015**, *92*, 063304.

(160) Kang, D.; Luo, K.; Runge, K.; Trickey, S. B. Two-Temperature Warm Dense Hydrogen as a Test of Quantum Protons Driven by Orbital-Free Density Functional Theory Electronic Forces. *Matter Radiat. Extremes* **2020**, *5*, 064403.

(161) Snyder, J. C.; Rupp, M.; Hansen, K.; Blooston, L.; Müller, K.-R.; Burke, K. Orbital-Free Bond Breaking via Machine Learning. *J. Chem. Phys.* **2013**, *139*, 224104.

(162) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K. R. Bypassing the Kohn-Sham Equations With Machine Learning. *Nat. Commun.* **2017**, *8*, 872.

(163) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133−A1138.

(164) Maurer, R. J.; Freysoldt, C.; Reilly, A. M.; Brandenburg, J. G.; Hofmann, O. T.; Björkman, T.; Lebègue, S.; Tkatchenko, A. Advances in Density-Functional Calculations for Materials Modeling. *Annu. Rev. Mater. Res.* **2019**, *49*, 1−30.

(165) Jacobsen, H.; Cavallo, L. In *Handbook of Computational Chemistry*; Leszczynski, J., Ed.; Springer: Dordrecht, 2012; pp 95−133.

(166) *Learn Density Functional Theory*. https://dft.uci.edu/learnDFT.php (accessed 2020-11-30).

(167) Tran, F.; Stelzl, J.; Blaha, P. Rungs 1 to 4 of DFT Jacob's Ladder: Extensive Test on the Lattice Constant, Bulk Modulus, and Cohesive Energy of Solids. *J. Chem. Phys.* **2016**, *144*, 204120.

(168) Kozuch, S.; Martin, J. M. Spin-Component-Scaled Double Hybrids: An Extensive Search for the Best Fifth-Rung Functionals Blending DFT and Perturbation Theory. *J. Comput. Chem.* **2013**, *34*, 2327−2344.

(169) Janesko, B. G. Reducing Density-Driven Error Without Exact Exchange. *Phys. Chem. Chem. Phys.* **2017**, *19*, 4793−4801.

(170) Gerber, I. C.; Ángyán, J. G.; Marsman, M.; Kresse, G. Range Separated Hybrid Density Functional With Long-Range Hartree-Fock Exchange Applied to Solids. *J. Chem. Phys.* **2007**, *127*, 054101.

(171) Pisani, C.; Dovesi, R.; Roetti, C. *Hartree-Fock Ab Initio Treatment of Crystalline Systems*; Springer: Berlin Heidelberg, 2012; Vol. *48*.

(172) Shishkin, M.; Sato, H. DFT+ U in Dudarev's Formulation With Corrected Interactions Between the Electrons With Opposite Spins: The Form of Hamiltonian, Calculation of Forces, and Bandgap Adjustments. *J. Chem. Phys.* **2019**, *151*, 024102.

(173) Petukhov, A. G.; Mazin, I. I.; Chioncel, L.; Lichtenstein, A. I. Correlated Metals and the LDA + U Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2003**, *67*, 153106.

(174) Sun, Q.; Chan, G. K. L. Quantum Embedding Theories. *Acc. Chem. Res.* **2016**, *49*, 2705−2712.

(175) Cortona, P. Self-Consistently Determined Properties of Solids Without Band-Structure Calculations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *44*, 8454.

(176) Huang, P.; Carter, E. A. Advances in Correlated Electronic Structure Methods for Solids, Surfaces, and Nanostructures. *Annu. Rev. Phys. Chem.* **2008**, *59*, 261−290.

(177) Manby, F. R.; Stella, M.; Goodpaster, J. D.; Miller, T. F. A Simple, Exact Density-Functional-Theory Embedding Scheme. *J. Chem. Theory Comput.* **2012**, *8*, 2564−2568.

(178) Libisch, F.; Huang, C.; Carter, E. A. Embedded Correlated Wavefunction Schemes: Theory and Applications. *Acc. Chem. Res.* **2014**, *47*, 2768−2775.

(179) Casida, M.; Huix-Rotllant, M. Progress in Time-Dependent Density-Functional Theory. *Annu. Rev. Phys. Chem.* **2012**, *63*, 287−323.

(180) Casida, M. E.; Casida, K. C.; Salahub, D. R. Excited-State Potential Energy Curves From Time-Dependent Density-Functional Theory: A Cross Section of Formaldehyde's 1A1Manifold. *Int. J. Quantum Chem.* **1998**, *70*, 933−941.

(181) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K. R.; Burke, K. Finding Density Functionals With Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.

(182) Schmidt, J.; Benavides-Riveros, C. L.; Marques, M. A. Machine Learning the Physical Nonlocal Exchange-Correlation Functional of Density-Functional Theory. *J. Phys. Chem. Lett.* **2019**, *10*, 6425−6431.

(183) Meyer, R.; Weichselbaum, M.; Hauser, A. W. Machine Learning Approaches Toward Orbital-Free Density Functional Theory: Simultaneous Training on the Kinetic Energy Density

Functional and Its Functional Derivative. *J. Chem. Theory Comput.* **2020**, *16*, 5685−5694.

(184) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum Chemical Accuracy From Density Functional Approximations via Machine Learning. *Nat. Commun.* **2020**, *11*, 5223.

(185) Nagai, R.; Akashi, R.; Sugino, O. Completing Density Functional Theory by Machine Learning Hidden Messages From Molecules. *Npj Comput. Mater.* **2020**, *6*, 43.

(186) Thiel, W. Semiempirical Quantum-Chemical Methods. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 145−157.

(187) Pople, J.; Beveridge, D. *Approximate Molecular Orbital Theory*; McGraw-Hill: United Kingdom, 1970.

(188) Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. 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.

(189) Stewart, J. J. 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.

(190) Dewar, M. J.; Thiel, W. Ground States of Molecules. 38. The MNDO Method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899−4907.

(191) Dral, P. O.; Wu, X.; Spörkel, L.; Koslowski, A.; Thiel, W. Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Benchmarks for Ground-State Properties. *J. Chem. Theory Comput.* **2016**, *12*, 1097−1120.

(192) Koskinen, P.; Mäkinen, V. Density-Functional Tight-Binding for Beginners. *Comput. Mater. Sci.* **2009**, *47*, 237−253.

(193) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; et al. Self-Consistent-Charge Density-Functional Tight-Binding Method for Simulations of Complex Materials Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *58*, 7260.

(194) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method With Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.

(195) Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 2120−2125.

(196) Hegde, G.; Bowen, R. C. Machine-Learned Approximations to Density Functional Theory Hamiltonians. *Sci. Rep.* **2017**, *7*, 42669.

(197) Stöhr, M.; Medrano Sandonas, L.; Tkatchenko, A. Accurate Many-Body Repulsive Potentials for Density-Functional Tight Binding From Deep Tensor Neural Networks. *J. Phys. Chem. Lett.* **2020**, *11*, 6835−6843.

(198) Li, H.; Collins, C.; Tanha, M.; Gordon, G. J.; Yaron, D. J. A Density Functional Tight Binding Layer for Deep Learning of Chemical Hamiltonians. *J. Chem. Theory Comput.* **2018**, *14*, 5764−5776.

(199) Poltavsky, I.; Zheng, L.; Mortazavi, M.; Tkatchenko, A. Quantum Tunneling of Thermal Protons Through Pristine Graphene. *J. Chem. Phys.* **2018**, *148*, 204707.

(200) Ceriotti, M.; Fang, W.; Kusalik, P. G.; McKenzie, R. H.; Michaelides, A.; Morales, M. A.; Markland, T. E. Nuclear Quantum Effects in Water and Aqueous Systems: Experiment, Theory, and Current Challenges. *Chem. Rev.* **2016**, *116*, 7529−7550.

(201) Marx, D.; Parrinello, M. Ab Initio Path Integral Molecular Dynamics: Basic Ideas. *J. Chem. Phys.* **1996**, *104*, 4077−4082.

(202) Chandler, D.; Wolynes, P. G. Exploiting the Isomorphism Between Quantum Theory and Classical Statistical Mechanics of Polyatomic Fluids. *J. Chem. Phys.* **1981**, *74*, 4078−4095.

(203) Cao, J.; Voth, G. A. The Formulation of Quantum Statistical Mechanics Based on the Feynman Path Centroid Density. IV. Algorithms for Centroid Molecular Dynamics. *J. Chem. Phys.* **1994**, *101*, 6168−6183.

(204) Hele, T. J. H.; Willatt, M. J.; Muolo, A.; Althorpe, S. C. Communication: Relation of Centroid Molecular Dynamics and Ring-

Polymer Molecular Dynamics to Exact Quantum Dynamics. *J. Chem. Phys.* **2015**, *142*, 191101.

(205) Wang, L.; Ceriotti, M.; Markland, T. E. Quantum Fluctuations and Isotope Effects in Ab Initio Descriptions of Water. *J. Chem. Phys.* **2014**, *141*, 104502.

(206) Sauceda, H. E.; Vassilev-Galindo, V.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. Dynamical Strengthening of Covalent and Non-Covalent Molecular Interactions by Nuclear Quantum Effects at Finite Temperature. *Nat. Commun.* **2021**, *12*, 442.

(207) Chmiela, S.; Sauceda, H. E.; Müller, K. R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations With Machine-Learned Force Fields. *Nat. Commun.* **2018**, *9*, 3887.

(208) Wang, X.; Ramírez-Hinestrosa, S.; Dobnikar, J.; Frenkel, D. The Lennard-Jones Potential: When (Not) to Use It. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10624−10633.

(209) Li, P.; Song, L. F.; Merz, K. M. Parameterization of Highly Charged Metal Ions Using the 12−6−4 LJ-type Nonbonded Model in Explicit Water. *J. Phys. Chem. B* **2015**, *119*, 883−895.

(210) Girifalco, L. A.; Weizer, V. G. Application of the Morse Potential Function to Cubic Metals. *Phys. Rev.* **1959**, *114*, 687.

(211) Buckingham, R. A. The Classical Equation of State of Gaseous Helium, Neon and Argon. *Proc. R. Soc. London A* **1938**, *168*, 264−283.

(212) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926.

(213) Stillinger, F. H.; Weber, T. A. Computer Simulation of Local Order in Condensed Phases of Silicon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1985**, *31*, 5262.

(214) Weiner, P. K.; Kollman, P. A. AMBER: Assisted Model Building With Energy Refinement. A General Program for Modeling Molecules and Their Interactions. *J. Comput. Chem.* **1981**, *2*, 287−303.

(215) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 198−210.

(216) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(217) Brooks, B. R.; Brooks III, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(218) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656−1676.

(219) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; Van Gunsteren, W. F. Definition and Testing of the GROMOS Force-Field Versions 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40*, 843.

(220) Daura, X.; Mark, A. E.; Van Gunsteren, W. F. Parametrization of Aliphatic CHn United Atoms of GROMOS96 Force Field. *J. Comput. Chem.* **1998**, *19*, 535−547.

(221) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(222) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. Optimized Intermolecular Potential Functions for Liquid Hydrocarbons. *J. Am. Chem. Soc.* **1984**, *106*, 6638−6646.

(223) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897−8909.

(224) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(225) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular

Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(226) Sun, H. Compass: An Ab Initio Force-Field Optimized for Condensed-Phase Applications - Overview With Details on Alkane and Benzene Compounds. *J. Phys. Chem. B* **1998**, *102*, 7338−7364.

(227) Heinz, H.; Lin, T. J.; Kishore Mishra, R.; Emami, F. S. Thermodynamically Consistent Force Fields for the Assembly of Inorganic, Organic, and Biological Nanostructures: The INTERFACE Force Field. *Langmuir* **2013**, *29*, 1754−1765.

(228) Gale, J. D. Empirical Potential Derivation for Ionic Materials. *Philos. Mag. B* **1996**, *73*, 3−19.

(229) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A.; et al. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549−2564.

(230) Lemkul, J. A.; Huang, J.; Roux, B.; Mackerell, A. D. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, *116*, 4983−5013.

(231) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. Parametrizing a Polarizable Force Field From Ab Initio Data. I. The Fluctuating Point Charge Model. *J. Chem. Phys.* **1999**, *110*, 741.

(232) Babin, V.; Leforestier, C.; Paesani, F. Development of a "First Principles" Water Potential With Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient. *J. Chem. Theory Comput.* **2013**, *9*, 5395−5403.

(233) Kumar, R.; Wang, F. F.; Jenness, G. R.; Jordan, K. D. A Second Generation Distributed Point Polarizable Water Model. *J. Chem. Phys.* **2010**, *132*, 014309.

(234) Xu, P.; Guidez, E. B.; Bertoni, C.; Gordon, M. S. Perspective: Ab Initio Force Field Methods Derived From Quantum Mechanics. *J. Chem. Phys.* **2018**, *148*, 090901.

(235) 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.

(236) Baskes, M. I. Modified Embedded-Atom Potentials for Cubic Materials and Impurities. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 2727.

(237) Finnis, M. W.; Sinclair, J. E. A Simple Empirical N-Body Potential for Transition Metals. *Philos. Mag. A* **1984**, *50*, 45−55.

(238) Sutton, A. P.; Chen, J. Long-Range Finnis-Sinclair Potentials. *Philos. Mag. Lett.* **1990**, *61*, 139−146.

(239) Brenner, D. W. Empirical Potential for Hydrocarbons for Use in Simulating the Chemical Vapor Deposition of Diamond Films. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1990**, *42*, 9458.

(240) Tersoff, J. New Empirical Approach for the Structure and Energy of Covalent Systems. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 6991.

(241) Tersoff, J. Modeling Solid-State Chemistry: Interatomic Potentials for Multicomponent Systems. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1989**, *39*, 5566−5568.

(242) Brenner, D. W.; Shenderova, O. A.; Harrison, J. A.; Stuart, S. J.; Ni, B.; Sinnott, S. B. A Second-Generation Reactive Empirical Bond Order (REBO) Potential Energy Expression for Hydrocarbons. *J. Phys.: Condens. Matter* **2002**, *14*, 783.

(243) Liang, T.; Shan, T. R.; Cheng, Y. T.; Devine, B. D.; Noordhoek, M.; Li, Y.; Lu, Z.; Phillpot, S. R.; Sinnott, S. B. Classical Atomistic Simulations of Surfaces and Heterogeneous Interfaces With the Charge-Optimized Many Body (COMB) Potentials. *Mater. Sci. Eng., R* **2013**, *74*, 255−279.

(244) Yu, J.; Sinnott, S. B.; Phillpot, S. R. Charge Optimized Many-Body Potential for the Si/SiO_2 System. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *75*, 085311.

(245) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; et al. The ReaxFF Reactive Force-Field: Development, Applications and Future Directions. *Npj Comput. Mater.* **2016**, *2*, 15011.

(246) Van Duin, A. C.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396−9409.

(247) Rappé, A. K.; Bormann-Rochotte, L. M.; Wiser, D. C.; Hart, J. R.; Pietsch, M. A.; Casewit, C. J.; Skiff, W. M. APT a Next Generation QM-based Reactive Force Field Model. *Mol. Phys.* **2007**, *105*, 301−324.

(248) Warshel, A.; Weiss, R. M. An Empirical Valence Bond Approach for Comparing Reactions in Solutions and in Enzymes. *J. Am. Chem. Soc.* **1980**, *102*, 6218−6226.

(249) Wu, Y.; Chen, H.; Wang, F.; Paesani, F.; Voth, G. A. An Improved Multistate Empirical Valence Bond Model for Aqueous Proton Solvation and Transport. *J. Phys. Chem. B* **2008**, *112*, 467−482.

(250) Hartke, B.; Grimme, S. Reactive Force Fields Made Simple. *Phys. Chem. Chem. Phys.* **2015**, *17*, 16715−16718.

(251) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5*, 129−145.

(252) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. Class IV Charge Models: A New Semiempirical Approach in Quantum Chemistry. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87−110.

(253) Mehler, E. L.; Solmajer, T. Electrostatic Effects in Proteins: Comparison of Dielectric and Charge Models. *Protein Eng., Des. Sel.* **1991**, *4*, 903−910.

(254) Chen, J.; Martínez, T. J. QTPIE: Charge Transfer With Polarization Current Equalization. A Fluctuating Charge Model With Correct Asymptotics. *Chem. Phys. Lett.* **2007**, *438*, 315−320.

(255) Poier, P. P.; Jensen, F. Describing Molecular Polarizability by a Bond Capacity Model. *J. Chem. Theory Comput.* **2019**, *15*, 3093−3107.

(256) Rappé, A. K.; Goddard, W. A. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.* **1991**, *95*, 3358−3363.

(257) Akimov, A. V.; Prezhdo, O. V. Large-Scale Computations in Chemistry: A Bird's Eye View of a Vibrant Field. *Chem. Rev.* **2015**, *115*, 5797−5890.

(258) Harrison, J. A.; Schall, J. D.; Maskey, S.; Mikulski, P. T.; Knippenberg, M. T.; Morrow, B. H. Review of Force Fields and Intermolecular Potentials Used in Atomistic Computational Materials Research. *Appl. Phys. Rev.* **2018**, *5*, 031104.

(259) Piquemal, J. P.; Jordan, K. D. Preface: Special Topic: From Quantum Mechanics to Force Fields. *J. Chem. Phys.* **2017**, *147*, 161401.

(260) Lennard-Jones, J. E. Cohesion. *Proc. Phys. Soc.* **1931**, *43*, 461−482.

(261) Tersoff, J. Empirical Interatomic Potential for Silicon With Improved Elastic Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *38*, 9902.

(262) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible With the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671−690.

(263) Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P. AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.* **2018**, *14*, 2084−2108.

(264) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations With AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542−1555.

(265) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations With AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878−3888.

(266) Stone, J. E.; Hardy, D. J.; Ufimtsev, I. S.; Schulten, K. GPU-accelerated Molecular Modeling Coming of Age. *J. Mol. Graphics Modell.* **2010**, *29*, 116−125.

(267) Glaser, J.; Nguyen, T. D.; Anderson, J. A.; Lui, P.; Spiga, F.; Millan, J. A.; Morse, D. C.; Glotzer, S. C. Strong Scaling of General-Purpose Molecular Dynamics Simulations on GPUs. *Comput. Phys. Commun.* **2015**, *192*, 97−107.

(268) Lagardère, L.; Jolly, L. H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z. F.; Harger, M.; Torabifard, H.; Cisneros, G. A.; Schnieders, M. J.; et al. Tinker-Hp: A Massively Parallel Molecular Dynamics Package for Multiscale Simulations of Large Complex Systems With Advanced Point Dipole Polarizable Force Fields. *Chem. Sci.* **2018**, *9*, 956−972.

(269) Zhang, Y.; Hu, C.; Jiang, B. Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning With a Physically Inspired Representation. *J. Phys. Chem. Lett.* **2019**, *10*, 4962−4967.

(270) Agrawal, A.; Choudhary, A. Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Mater.* **2016**, *4*, 053208.

(271) Pun, G. P.; Batra, R.; Ramprasad, R.; Mishin, Y. Physically Informed Artificial Neural Networks for Atomistic Modeling of Materials. *Nat. Commun.* **2019**, *10*, 2339.

(272) Guo, F.; Wen, Y.-S.; Feng, S.-Q.; Li, X.-D.; Li, H.-S.; Cui, S.-X.; Zhang, Z.-R.; Hu, H.-Q.; Zhang, G.-Q.; Cheng, X.-L. Intelligent-ReaxFF: Evaluating the Reactive Force Field Parameters With Machine Learning. *Comput. Mater. Sci.* **2020**, *172*, 109393.

(273) Narayanan, B.; Chan, H.; Kinaci, A.; Sen, F. G.; Gray, S. K.; Chan, M. K.; Sankaranarayanan, S. K. Machine Learnt Bond Order Potential to Model Metal-Organic (Co-C) Heterostructures. *Nanoscale* **2017**, *9*, 18229−18239.

(274) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(275) Wood, M. A.; Thompson, A. P. Extending the Accuracy of the SNAP Interatomic Potential Form. *J. Chem. Phys.* **2018**, *148*, 241721.

(276) Bartók, 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.

(277) Boes, J. R.; Groenenboom, M. C.; Keith, J. A.; Kitchin, J. R. Neural Network and ReaxFF Comparison for Au Properties. *Int. J. Quantum Chem.* **2016**, *116*, 979−987.

(278) Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The Power of Coarse Graining in Biomolecular Simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 225−248.

(279) Mennucci, B. Polarizable Continuum Model. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 386−404.

(280) Jäger, M.; Schäfer, R.; Johnston, R. L. First Principles Global Optimization of Metal Clusters and Nanoalloys. *Adv. Phys. X* **2018**, *3*, S100009.

(281) Dieterich, J. M.; Hartke, B. OGOLEM: Global Cluster Structure Optimisation for Arbitrary Mixtures of Flexible Molecules. A Multiscaling, Object-Oriented Approach. *Mol. Phys.* **2010**, *108*, 279−291.

(282) Wales, D. J.; Doye, J. P. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing Up to 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111−5116.

(283) Zhang, J.; Dolg, M. ABCluster: The Artificial Bee Colony Algorithm for Cluster Global Optimization. *Phys. Chem. Chem. Phys.* **2015**, *17*, 24173−24181.

(284) Goedecker, S. Minima Hopping: An Efficient Search Method for the Global Minimum of the Potential Energy Surface of Complex Molecular Systems. *J. Chem. Phys.* **2004**, *120*, 9911.

(285) Schlegel, H. B. Geometry Optimization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 790−809.

(286) Sheppard, D.; Terrell, R.; Henkelman, G. Optimization Methods for Finding Minimum Energy Paths. *J. Chem. Phys.* **2008**, *128*, 134106.

(287) Schlegel, B. H. Estimating the Hessian for Gradient-Type Geometry Optimizations. *Theor. Chim. Acta* **1984**, *66*, 333−340.

(288) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths. *J. Chem. Phys.* **2000**, *113*, 9901.

(289) Sheppard, D.; Xiao, P.; Chemelewski, W.; Johnson, D. D.; Henkelman, G. A Generalized Solid-State Nudged Elastic Band Method. *J. Chem. Phys.* **2012**, *136*, 074103.

(290) Zimmerman, P. M. Growing String Method With Interpolation and Optimization in Internal Coordinates: Method and Examples. *J. Chem. Phys.* **2013**, *138*, 184102.

(291) Samanta, A.; Weinan, E. Optimization-Based String Method for Finding Minimum Energy Path. *Commun. Comput. Phys.* **2013**, *14*, 265−275.

(292) Burger, S. K.; Yang, W. Quadratic String Method for Determining the Minimum-Energy Path Based on Multiobjective Optimization. *J. Chem. Phys.* **2006**, *124*, 054109.

(293) Peterson, A. A. Acceleration of Saddle-Point Searches With Machine Learning. *J. Chem. Phys.* **2016**, *145*, 074106.

(294) Garijo del Río, E.; Mortensen, J. J.; Jacobsen, K. W. Local Bayesian Optimizer for Atomic Structures. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *100*, 104103.

(295) Garrido Torres, J. A.; Jennings, P. C.; Hansen, M. H.; Boes, J. R.; Bligaard, T. Low-Scaling Algorithm for Nudged Elastic Band Calculations Using a Surrogate Machine Learning Model. *Phys. Rev. Lett.* **2019**, *122*, 156001.

(296) Meyer, R.; Schmuck, K. S.; Hauser, A. W. Machine Learning in Computational Chemistry: An Evaluation of Method Performance for Nudged Elastic Band Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 6513−6523.

(297) Koistinen, O. P.; Ásgeirsson, V.; Vehtari, A.; Jónsson, H. Nudged Elastic Band Calculations Accelerated With Gaussian Process Regression Based on Inverse Interatomic Distances. *J. Chem. Theory Comput.* **2019**, *15*, 6738−6751.

(298) Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann Generators: Sampling Equilibrium States of Many-Body Systems With Deep Learning. *Science* **2019**, *365*, No. eaaw1147.

(299) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. Operators in Quantum Machine Learning: Response Properties in Chemical Space. *J. Chem. Phys.* **2019**, *150*, 064105.

(300) Gastegger, M.; Schütt, K. T.; Müller, K.-R. Machine Learning of Solvent Effects on Molecular Spectra and Reactions. *arXiv*, 2020, 2010.14942. https://arxiv.org/abs/2010.14942.

(301) Varghese, J. J.; Mushrif, S. H. Origins of Complex Solvent Effects on Chemical Reactivity and Computational Tools to Investigate Them: A Review. *React. Chem. Eng.* **2019**, *4*, 165−206.

(302) Basdogan, Y.; Maldonado, A. M.; Keith, J. A. Advances and Challenges in Modeling Solvated Reaction Mechanisms for Renewable Fuels and Chemicals. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, No. 1446.

(303) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999−3094.

(304) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760−768.

(305) Klamt, A. The COSMO and COSMO-RS Solvation Models. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 699−709.

(306) Hirata, F., Ed. *Molecular Theory of Solvation*; Kluwer Academic Publishers: Norwell, MA, 2003; Vol. 24.

(307) Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic Interaction of a Solute With a Continuum. A Direct Utilizaion of Ab Initio Molecular Potentials for the Prevision of Solvent Effects. *Chem. Phys.* **1981**, *55*, 117−129.

(308) Cances, E.; Mennucci, B.; Tomasi, J. A New Integral Equation Formalism for the Polarizable Continuum Model: Theoretical Background and Applications to Isotropic and Anisotropic Dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032−3041.

(309) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.

(310) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102*, 1995−2001.

(311) Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents With Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799−805.

(312) Klamt, A. Conductor-Like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224−2235.

(313) Bernales, V. S.; Marenich, A. V.; Contreras, R.; Cramer, C. J.; Truhlar, D. G. Quantum Mechanical Continuum Solvation Models for Ionic Liquids. *J. Phys. Chem. B* **2012**, *116*, 9122−9129.

(314) Truchon, J. F.; Pettitt, B. M.; Labute, P. A Cavity Corrected 3d-Rism Functional for Accurate Solvation Free Energies. *J. Chem. Theory Comput.* **2014**, *10*, 934−941.

(315) Nishihara, S.; Otani, M. Hybrid Solvation Models for Bulk, Interface, and Membrane: Reference Interaction Site Methods Coupled With Density Functional Theory. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *96*, 115429.

(316) Kamerlin, S. C.; Haranczyk, M.; Warshel, A. Progress in Ab Initio QM/MM Free-Energy Simulations of Electrostatic Energies in Proteins: Accelerated QM/MM Studies of pKa, Redox Reactions and Solvation Free Energies. *J. Phys. Chem. B* **2009**, *113*, 1253−1272.

(317) Boereboom, J. M.; Fleurat-Lessard, P.; Bulo, R. E. Explicit Solvation Matters: Performance of QM/MM Solvation Models in Nucleophilic Addition. *J. Chem. Theory Comput.* **2018**, *14*, 1841−1852.

(318) Gregersen, B. A.; Lopez, X.; York, D. M. Hybrid QM/MM Study of Thio Effects in Transphosphorylation Reactions: The Role of Solvation. *J. Am. Chem. Soc.* **2004**, *126*, 7504−7513.

(319) Maldonado, A. M.; Basdogan, Y.; Berryman, J. T.; Rempe, S. B.; Keith, J. A. First-Principles Modeling of Chemistry in Mixed Solvents: Where to Go From Here? *J. Chem. Phys.* **2020**, *152*, 130902.

(320) Skyner, R.; McDonagh, J.; Groom, C.; Van Mourik, T.; Mitchell, J. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174−6191.

(321) Basdogan, Y.; Groenenboom, M. C.; Henderson, E.; De, S.; Rempe, S. B.; Keith, J. A. Machine Learning-Guided Approach for Studying Solvation Environments. *J. Chem. Theory Comput.* **2020**, *16*, 633−642.

(322) Pratt, L. R.; Laviolette, R. A. Quasi-Chemical Theories of Associated Liquids. *Mol. Phys.* **1998**, *94*, 909−915.

(323) Rempe, S. B.; Pratt, L. R.; Hummer, G.; Kress, J. D.; Martin, R. L.; Redondo, A. The Hydration Number of Li+ in Liquid Water. *J. Am. Chem. Soc.* **2000**, *122*, 966−967.

(324) Chew, A. K.; Jiang, S.; Zhang, W.; Zavala, V. M.; Van Lehn, R. C. Fast Predictions of Liquid-Phase Acid-Catalyzed Reaction Rates Using Molecular Dynamics Simulations and Convolutional Neural Networks. *Chem. Sci.* **2020**, *11*, 12464−12476.

(325) Zhang, P.; Shen, L.; Yang, W. Solvation Free Energy Calculations With Quantum Mechanics/Molecular Mechanics and Machine Learning Models. *J. Phys. Chem. B* **2019**, *123*, 901−908.

(326) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative Correlation of Physical and Chemical Properties With Chemical Structure: Utility for Prediction. *Chem. Rev.* **2010**, *110*, 5714−5789.

(327) Muratov, E. N.; et al. QSAR Without Borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.

(328) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.

(329) Geerlings, P.; Chamorro, E.; Chattaraj, P. K.; De Proft, F.; Gázquez, J. L.; Liu, S.; Morell, C.; Toro-Labbé, A.; Vela, A.; Ayers, P. Conceptual Density Functional Theory: Status, Prospects, Issues. *Theor. Chem. Acc.* **2020**, *139*, 36.

(330) von Rudorff, G. F.; von Lilienfeld, O. A. Alchemical Perturbation Density Functional Theory. *Phys. Rev. Res.* **2020**, *2*, 023220.

(331) Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. R. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *arXiv*, 2012, 1207.0580. https://arxiv.org/abs/1207.0580.

(332) Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* **2015**, 1−9.

(333) Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211−252.

(334) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet Classification With Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84−90.

(335) Blei, D. M.; Ng, A. Y.; Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993−1022.

(336) Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137−1155.

(337) Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K., et al. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. *arXiv*, 2016, 1609.08144. https://arxiv.org/abs/1609.08144.

(338) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv*, 2013, 1301.3781. https://arxiv.org/abs/1301.3781.

(339) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998−6008.

(340) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, 2009.

(341) Rasmussen, C. E. Gaussian Processes in Machine Learning. *Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science*: Berlin, 2004; pp 63−71.

(342) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, 2006.

(343) Hornik, K.; Stinchcombe, M.; White, H. Multilayer Feedforward Networks Are Universal Approximators. *Neural Netw.* **1989**, *2*, 359−366.

(344) Tran, D.; Ranganath, R.; Blei, D. M. The Variational Gaussian Process. *arXiv preprint*, 2015, 1511.06499. https://arxiv.org/abs/1511.06499.

(345) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, 1995.

(346) Poggio, T.; Girosi, F. Networks for Approximation and Learning. *Proc. IEEE* **1990**, *78*, 1481−1497.

(347) Smola, A. J.; Schölkopf, B.; Müller, K.-R. The Connection Between Regularization Operators and Support Vector Kernels. *Neural Netw.* **1998**, *11*, 637−649.

(348) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*; MIT Press: Cambridge, MA, 2005.

(349) Caruana, R.; Lawrence, S.; Giles, C. L. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. *Advances in Neural Information Processing Systems* **2001**, 402−408.

(350) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.

(351) Geman, S.; Bienenstock, E.; Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Comput.* **1992**, *4*, 1−58.

(352) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights From Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

(353) Bietti, A.; Mairal, J. On the Inductive Bias of Neural Tangent Kernels. *arXiv*, 2019, 1905.12173. https://arxiv.org/abs/1905.12173

(354) Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.-R. Explaining Nonlinear Classification Decisions With Deep Taylor Decomposition. *Pattern Recognit.* **2017**, *65*, 211−222.

(355) Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., Müller, K.-R., Eds. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Lecture Notes in Computer Science; Springer: New York, NY, 2019; Vol. *11700*.

(356) Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; Müller, K.-R. How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.* **2010**, *11*, 1803−1831.

(357) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One* **2015**, *10*, No. e0130140.

(358) Montavon, G.; Samek, W.; Müller, K.-R. Methods for Interpreting and Understanding Deep Neural Networks. *Digit. Signal Process.* **2018**, *73*, 1−15.

(359) Holzinger, A. From Machine Learning to Explainable AI. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*; 2018; pp 55−66.

(360) Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nat. Commun.* **2019**, *10*, 1096.

(361) Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; Muller, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* **2021**, *109*, 247−278.

(362) Bongard, J.; Lipson, H. Automated Reverse Engineering of Nonlinear Dynamical Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 9943−9948.

(363) Schmidt, M.; Lipson, H. Distilling Free-Form Natural Laws From Experimental Data. *Science* **2009**, *324*, 81−85.

(364) 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.

(365) Boninsegna, L.; Nüske, F.; Clementi, C. Sparse Learning of Stochastic Dynamical Equations. *J. Chem. Phys.* **2018**, *148*, 241723.

(366) Hoffmann, M.; Fröhner, C.; Noé, F. Reactive SINDy: Discovering Governing Reactions From Concentration Data. *J. Chem. Phys.* **2019**, *150*, 025101.

(367) Watters, N.; Zoran, D.; Weber, T.; Battaglia, P.; Pascanu, R.; Tacchetti, A. Visual Interaction Networks: Learning a Physics Simulator From Video. *Adv. Neural Inf. Process. Syst.* **2017**, 4539−4547.

(368) Raissi, M. Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations. *J. Mach. Learn. Res.* **2018**, *19*, 932−955.

(369) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science* **2018**, *360*, 186−190.

(370) Chuang, K. V.; Keiser, M. J. Comment on "Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science* **2018**, *362*, No. eaat8603.

(371) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on "Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science* **2018**, *362*, No. eaat8763.

(372) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3*, No. e1603015.

(373) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.

(374) Cheng, B.; Griffiths, R.-R.; Wengert, S.; Kunkel, C.; Stenczel, T.; Zhu, B.; Deringer, V. L.; Bernstein, N.; Margraf, J. T.; Reuter, K.; et al. Mapping Materials and Molecules. *Acc. Chem. Res.* **2020**, *53*, 1981−1991.

(375) Reinhardt, A.; Pickard, C. J.; Cheng, B. Predicting the Phase Diagram of Titanium Dioxide With Random Search and Pattern Recognition. *Phys. Chem. Chem. Phys.* **2020**, *22*, 12697−12705.

(376) Meila, M.; Koelle, S.; Zhang, H. A Regression Approach for Explaining Manifold Embedding Coordinates. *arXiv*, 2018, 1811.11891. https://arxiv.org/abs/1811.11891.

(377) Cox, M. A.; Cox, T. F. *Handbook of Data Visualization*; Springer, 2008; pp 315−347.

(378) Schölkopf, B.; Smola, A.; Müller, K.-R. Kernel Principal Component Analysis. *International Conference on Artificial Neural Networks*; 1997; pp 583−588.

(379) Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **1998**, *10*, 1299−1319.

(380) Maaten, L. v. d.; Hinton, G. Visualizing Data Using T-Sne. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(381) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the Representation of Complex Free-Energy Landscapes Using Sketch-Map. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023−13028.

(382) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, 2018, 1802.03426. https://arxiv.org/abs/1802.03426.

(383) Ruff, L.; Kauffmann, J. R.; Vandermeulen, R. A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T. G.; Müller, K.-R. A Unifying Review of Deep and Shallow Anomaly Detection. *Proc. IEEE* **2021**, *109*, 756−795.

(384) Kaelbling, L. P.; Littman, M. L.; Moore, A. W. Reinforcement Learning: A Survey. *J. Artif. Intell. Res.* **1996**, *4*, 237−285.

(385) Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* **1958**, *65*, 386−408.

(386) Rosenblatt, F. Principles of Neurodynamics. *Perceptrons and the Theory of Brain Mechanisms*; Cornell Aeronautical Lab, Inc.: Buffalo, NY, 1961.

(387) Minsky, M.; Papert, S. A. *Perceptrons: An Introduction to Computational Geometry*; MIT Press: Cambridge, MA, 2017.

(388) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533−536.

(389) Lecun, Y. Une procédure d'apprentissage pour réseau à seuil asymétrique (A learning scheme for asymmetric threshold networks). *Proceedings of Cognitiva 85*; Paris, France, 1985; pp 599−604.

(390) Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, NY, 1995.

(391) Funahashi, K.-I. On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Netw.* **1989**, *2*, 183−192.

(392) Cybenko, G. Approximation by Superpositions of a Sigmoidal Function. *Math. Control. Signals Syst.* **1989**, *2*, 303−314.

(393) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.

(394) Müller, K.-R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans. Neural Netw.* **2001**, *12*, 181−201.

(395) Schölkopf, B.; Smola, A. J. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, 2002.

(396) Schölkopf, B.; Mika, S.; Burges, C. J.; Knirsch, P.; Müller, K.-R.; Rätsch, G.; Smola, A. J. Input Space Versus Feature Space in Kernel-Based Methods. *IEEE Trans. Neural Netw.* **1999**, *10*, 1000−1017.

(397) Müller, K.-R.; Smola, A. J.; Rätsch, G.; Schölkopf, B.; Kohlmorgen, J.; Vapnik, V. Predicting Time Series With Support Vector Machines. *International Conference on Artificial Neural Networks*; 1997; pp 999−1004.

(398) Harmeling, S.; Ziehe, A.; Kawanabe, M.; Müller, K.-R. Kernel-Based Nonlinear Blind Source Separation. *Neural Comput.* **2003**, *15*, 1089−1124.

(399) Braun, M. L.; Buhmann, J. M.; Müller, K.-R. On Relevant Dimensions in Kernel Feature Spaces. *J. Mach. Learn. Res.* **2008**, *9*, 1875−1908.

(400) Montavon, G.; Braun, M. L.; Müller, K.-R. Kernel Analysis of Deep Networks. *J. Mach. Learn. Res.* **2011**, *12*, 2563−2581.

(401) Montavon, G.; Braun, M. L.; Krueger, T.; Müller, K.-R. Analyzing Local Structure in Kernel-Based Learning: Explanation, Complexity, and Reliability Assessment. *IEEE Signal Process. Mag.* **2013**, *30*, 62−74.

(402) Mitchell, J. B. O. Machine Learning Methods in Chemo-informatics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 468−481.

(403) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity With Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(404) Mizoguchi, T.; Kiyohara, S. Machine Learning Approaches for ELNES/XANES. *Microscopy* **2020**, *69*, 92−109.

(405) Carr, D. A.; Lach-Hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Machine Learning Approach for Structure-Based Zeolite Classification. *Microporous Mesoporous Mater.* **2009**, *117*, 339−349.

(406) Legrain, F.; Carrete, J.; Van Roekeghem, A.; Madsen, G. K.; Mingo, N. Materials Screening for the Discovery of New Half-Heuslers: Machine Learning Versus Ab Initio Methods. *J. Phys. Chem. B* **2018**, *122*, 625−632.

(407) Lam Pham, T.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; Chi Dam, H. Machine Learning Reveals Orbital Interaction in Materials. *Sci. Technol. Adv. Mater.* **2017**, *18*, 756−765.

(408) Sugiyama, M.; Krauledat, M.; Müller, K.-R. Covariate Shift Adaptation by Importance Weighted Cross Validation. *J. Mach. Learn. Res.* **2007**, *8*, 985−1005.

(409) Sugiyama, M.; Kawanabe, M. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*; MIT Press: Cambridge, MA, 2012.

(410) Zien, A.; Rätsch, G.; Mika, S.; Schölkopf, B.; Lengauer, T.; Müller, K.-R. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *Bioinformatics* **2000**, *16*, 799−807.

(411) Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(412) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(413) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies With Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(414) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094−1101.

(415) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K. R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.

(416) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL Revisited: Faster and More Accurate Quantum Machine Learning. *J. Chem. Phys.* **2020**, *152*, 044107.

(417) 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.

(418) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. *arXiv*, 2017, 1704.06439. https://arxiv.org/abs/1704.06439.

(419) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K.-R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.

(420) Drautz, R. Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *99*, 014104.

(421) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *34th International Conference on Machine Learning ICML 2017*; 2017; pp 2053−2070.

(422) Schütt, K. T.; Kindermans, P. J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K. R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 992−1002.

(423) Kocer, E.; Mason, J. K.; Erturk, H. A Novel Approach to Describe Chemical Environments in High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2019**, *150*, 154102.

(424) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Finger-prints. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2224−2232.

(425) Li, Z.; Wang, S.; Chin, W. S.; Achenie, L. E.; Xin, H. High-Throughput Screening of Bimetallic Catalysts Enabled by Machine Learning. *J. Mater. Chem. A* **2017**, *5*, 24131−24138.

(426) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for De Novo Molecular Design. *J. Cheminf.* **2018**, *10*, 31.

(427) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(428) Ehmki, E. S. R.; Schmidt, R.; Ohm, F.; Rarey, M. Comparing Molecular Patterns Using the Example of SMARTS: Applications and Filter Collection Analysis. *J. Chem. Inf. Model.* **2019**, *59*, 2572−2586.

(429) Schmidt, R.; Ehmki, E. S.; Ohm, F.; Ehrlich, H. C.; Mashychev, A.; Rarey, M. Comparing Molecular Patterns Using the Example of SMARTS: Theory and Algorithms. *J. Chem. Inf. Model.* **2019**, *59*, 2560−2571.

(430) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(431) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(432) Behler, J. Neural Network Potential-Energy Surfaces in Chemistry: A Tool for Large-Scale Simulations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930−17955.

(433) Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(434) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet-A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(435) Unke, O. T.; Meuwly, M. A Reactive, Scalable, and Transferable Model for Molecular Energies From a Neural Network Approach Based on Local Information. *J. Chem. Phys.* **2018**, *148*, 241708.

(436) Murray, I. Gaussian Processes and Fast Matrix-Vector Multiplies. *NUMML 2009 Numerical Mathematics in Machine Learning ICML 2009 Workshop.* **2009**.

(437) Wilson, A. G.; Hu, Z.; Salakhutdinov, R.; Xing, E. P. Deep Kernel Learning. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* **2016**, *51*, 370−378.

(438) Gardner, J. R.; Pleiss, G.; Wu, R.; Weinberger, K. Q.; Wilson, A. G. Product Kernel Interpolation for Scalable Gaussian Processes. *arXiv*, 2018, 1802.08903. https://arxiv.org/abs/1802.08903.

(439) Gardner, J.; Pleiss, G.; Weinberger, K. Q.; Bindel, D.; Wilson, A. G. Gpytorch: Blackbox Matrix-Matrix Gaussian Process Inference With Gpu Acceleration. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7576−7586.

(440) Wang, K.; Pleiss, G.; Gardner, J.; Tyree, S.; Weinberger, K. Q.; Wilson, A. G. Exact Gaussian Processes on a Million Data Points. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14648−14659.

(441) LeCun, Y. A.; Bottou, L.; Orr, G. B.; Müller, K.-R. *In Neural Networks: Tricks of the Trade*; Lecture Notes in Computer Science; Montavon, G., Orr, G. B., Müller, K.-R., Eds.; Springer-Verlag: Berlin, 2012; Vol. 7700; pp 9−48.

(442) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.

(443) MacKay, D. J. The Evidence Framework Applied to Classification Networks. *Neural Comput.* **1992**, *4*, 720−736.

(444) MacKay, D. J. A Practical Bayesian Framework for Backpropagation Networks. *Neural Comput.* **1992**, *4*, 448−472.

(445) Kwok, J. T.-Y. The Evidence Framework Applied to Support Vector Machines. *IEEE Trans. Neural Netw.* **2000**, *11*, 1162−1173.

(446) Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716−723.

(447) Schwarz, G. Estimating the Dimension of a Model. *Ann. Statis.* **1978**, *6*, 461−464.

(448) Murata, N.; Yoshizawa, S.; Amari, S. Network Information Criterion-Determining the Number of Hidden Units for an Artificial Neural Network Model. *IEEE Trans. Neural Netw.* **1994**, *5*, 865−872.

(449) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, 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.

(450) Wang, J.; Chmiela, S.; Müller, K.-R.; Noé, F.; Clementi, C. Ensemble Learning of Coarse-Grained Molecular Dynamics Force Fields With a Kernel Approach. *J. Chem. Phys.* **2020**, *152*, 194106.

(451) Bonati, L.; Rizzi, V.; Parrinello, M. Data-Driven Collective Variables for Enhanced Sampling. *J. Phys. Chem. Lett.* **2020**, *11*, 2998−3004.

(452) Willatt, M. J.; Musil, F.; Ceriotti, M. Atom-Density Representations for Machine Learning. *J. Chem. Phys.* **2019**, *150*, 154110.

(453) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-Inspired Structural Representations for Molecules and Materials. *arXiv*, 2021, 2101.04673. https://arxiv.org/abs/2101.04673.

(454) Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for Measuring Distances in Configuration Spaces. *J. Chem. Phys.* **2013**, *139*, 184118.

(455) Barker, J.; Bulin, J.; Hamaekers, J.; Mathias, S. In *Scientific Computing and Algorithms in Industrial Simulations*; Griebel, M., Schüller, A., Schweitzer, M. A., Eds.; Springer: Berlin, 2017; pp 25−42.

(456) Tsuzuki, H.; Branicio, P. S.; Rino, J. P. Structural Characterization of Deformed Crystals by Analysis of Common Atomic Neighborhood. *Comput. Phys. Commun.* **2007**, *177*, 518−523.

(457) Çaylak, O.; Anatole von Lilienfeld, O.; Baumeier, B. Wasserstein Metric for Improved Quantum Machine Learning With Adjacency Matrix Representations. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 03LT01.

(458) Gastegger, M.; Schwiedrzik, L.; Bittermann, M.; Berzsenyi, F.; Marquetand, P. WACSF - Weighted Atom-Centered Symmetry Functions as Descriptors in Machine Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241709.

(459) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids Across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.

(460) Pronobis, W.; Tkatchenko, A.; Müller, K.-R. Many-Body Descriptors for Predicting Molecular Properties With Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules. *J. Chem. Theory Comput.* **2018**, *14*, 2991−3003.

(461) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep Potential Molecular Dynamics: A Scalable Model With the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.

(462) Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; E, W. End-to-End Symmetry Preserving Inter-Atomic Potential Energy Model for Finite and Extended Systems. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 4436−4446.

(463) Anderson, B.; Hy, T. S.; Kondor, R. Cormorant: Covariant Molecular Neural Networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14537−14546.

(464) Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor Field Networks: Rotation-and Translation-Equivariant Neural Networks for 3d Point Clouds. *arXiv*, 2018, 1802.08219. https://arxiv.org/abs/1802.08219.

(465) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241730.

(466) Schlömer, T.; Heck, D.; Deussen, O. Farthest-Point Optimized Point Sets with Maximized Minimum Distance. *Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics*; New York, NY, USA, 2011; p 135−142.

(467) Mahoney, M. W.; Drineas, P. CUR Matrix Decompositions for Improved Data Analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 697−702.

(468) Pozdnyakov, S. N.; Willatt, M. J.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Incompleteness of Atomic Structure Representations. *Phys. Rev. Lett.* **2020**, *125*, 166001.

(469) von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier Series of Atomic Radial Distribution Functions: A Molecular Fingerprint for Machine Learning Models of Quantum Chemical Properties. *Int. J. Quantum Chem.* **2015**, *115*, 1084−1093.

(470) Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J.; Csanyi, G.; Ceriotti, M. Machine Learning Unifies the Modelling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, No. e1701816.

(471) Monserrat, B.; Brandenburg, J. G.; Engel, E. A.; Cheng, B. Extracting Ice Phases From Liquid Water: Why a Machine-Learning Water Model Generalizes So Well. *arXiv*, 2020, 2006.13316, https://arxiv.org/abs/2006.13316.

(472) Deringer, V. L.; Csányi, G. Machine Learning Based Interatomic Potential for Amorphous Carbon. *Phys. Rev. B: Condens. Matter Mater.* **2017**, *95*, 094203.

(473) Rowe, P.; Deringer, V. L.; Gasparotto, P.; Csányi, G.; Michaelides, A. An Accurate and Transferable Machine Learning Potential for Carbon. *J. Chem. Phys.* **2020**, *153*, 034702.

(474) Yue, S.; Muniz, M. C.; Calegari Andrade, M. F.; Zhang, L.; Car, R.; Panagiotopoulos, A. Z. When Do Short-Range Atomistic Machine-Learning Models Fall Short? *J. Chem. Phys.* **2021**, *154*, 034111.

(475) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. General-Purpose Machine Learning Potentials Capturing Nonlocal Charge Transfer. *Acc. Chem. Res.* **2021**, *54*, 808−817.

(476) 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.

(477) Bereau, T.; Andrienko, D.; von Lilienfeld, O. A. Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules. *J. Chem. Theory Comput.* **2015**, *11*, 3225−3233.

(478) Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic 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.

(479) Grisafi, A.; Ceriotti, M. Incorporating Long-Range Physics in Atomic-Scale Machine Learning. *J. Chem. Phys.* **2019**, *151*, 204105.

(480) Grisafi, A.; Nigam, J.; Ceriotti, M. Multi-Scale Approach for the Prediction of Atomic Scale Properties. *Chem. Sci.* **2021**, *12*, 2078−2090.

(481) Glielmo, A.; Sollich, P.; De Vita, A. Accurate Interatomic Force Fields via Machine Learning With Covariant Kernels. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 214302.

(482) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, 036002.

(483) Willatt, M. J.; Musil, F.; Ceriotti, M. Feature Optimization for Atomistic Machine Learning Yields a Data-Driven Construction of the Periodic Table of the Elements. *Phys. Chem. Chem. Phys.* **2018**, *20*, 29661−29668.

(484) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies Using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148*, 241715.

(485) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.

(486) Jørgensen, P. B.; Jacobsen, K. W.; Schmidt, M. N. Neural Message Passing With Edge Updates for Predicting Properties of Molecules and Materials. *arXiv*, 2018, 1806.03146. https://arxiv.org/abs/1806.03146.

(487) Klicpera, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. *International Conference on Learning Representations*; 2020.

(488) Shao, Y.; Hellstrom, M.; Mitev, P. D.; Knijff, L.; Zhang, C. PiNN: A Python Library for Building Atomic Neural Networks of Molecules and Materials. *J. Chem. Inf. Model.* **2020**, *60*, 1184−1193.

(489) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.

(490) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.

(491) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **2019**, *5*, 57−64.

(492) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio Jr, R. A.; Ceriotti, M. Accurate Molecular Polarizabilities With Coupled Cluster Theory and Machine Learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3401−3406.

(493) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.

(494) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-Learning Approach for One- And Two-Body Corrections to Density Functional Theory: Applications to Molecular and Condensed Water. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *88*, 054104.

(495) Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F. A Universal Density Matrix Functional From Molecular Orbital-Based Machine Learning: Transferability Across Organic Molecules. *J. Chem. Phys.* **2019**, *150*, 131103.

(496) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower Than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.

(497) Hollingsworth, J.; Baker, T. E.; Burke, K. Can Exact Conditions Improve Machine-Learned Density Functionals? *J. Chem. Phys.* **2018**, *148*, 241743.

(498) Li, L.; Snyder, J. C.; Pelaschier, I. M.; Huang, J.; Niranjan, U. N.; Duncan, P.; Rupp, M.; Müller, K. R.; Burke, K. Understanding Machine-Learned Density Functionals. *Int. J. Quantum Chem.* **2016**, *116*, 819−833.

(499) Vu, K.; Snyder, J. C.; Li, L.; Rupp, M.; Chen, B. F.; Khelif, T.; Müller, K. R.; Burke, K. Understanding Kernel Ridge Regression: Common Behaviors From Simple Functions to Density Functionals. *Int. J. Quantum Chem.* **2015**, *115*, 1115−1128.

(500) Nagai, R.; Akashi, R.; Sasaki, S.; Tsuneyuki, S. Neural-Network Kohn-Sham Exchange-Correlation Potential and Its Out-of-Training Transferability. *J. Chem. Phys.* **2018**, *148*, 241737.

(501) Carleo, G.; Troyer, M. Solving the Quantum Many-Body Problem With Artificial Neural Networks. *Science* **2017**, *355*, 602−606.

(502) Manzhos, S.; Carrington, T. An Improved Neural Network Method for Solving the Schrödinger Equation. *Can. J. Chem.* **2009**, *87*, 864−871.

(503) Huang, B.; von Lilienfeld, O. A. Quantum Machine Learning Using Atom-in-Molecule-Based Fragments Selected on the Fly. *Nat. Chem.* **2020**, *12*, 945−951.

(504) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics With on-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.

(505) Podryabinkin, E. V.; Shapeev, A. V. Active Learning of Linearly Parametrized Interatomic Potentials. *Comput. Mater. Sci.* **2017**, *140*, 171−180.

(506) *Quantum-machine.org.* http://quantum-machine.org/datasets/ .

(507) de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L.-Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S.; et al. New Frontiers for the Materials Genome Initiative. *Npj Comput. Mater.* **2019**, *5*, 41.

(508) *The Materials Project.* https://materialsproject.org/.

(509) *The NOMAD Laboratory.* https://nomad-repository.eu/.

(510) Calderon, C. E.; Plata, J. J.; Toher, C.; Oses, C.; Levy, O.; Fornari, M.; Natan, A.; Mehl, M. J.; Hart, G.; Buongiorno Nardelli, M.; et al. The AFLOW Standard for High-Throughput Materials Science Calculations. *Comput. Mater. Sci.* **2015**, *108*, 233−238.

(511) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential With DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.

(512) Smith, J. S.; Isayev, O.; Roitberg, A. E. Data Descriptor: ANI-1, a Data Set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules. *Sci. Data* **2017**, *4*, 170193.

(513) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Sci. Data* **2020**, *7*, 134.

(514) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

(515) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.

(516) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal-Organic Frameworks: A Tool to Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26*, 6185−6192.

(517) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, With Input Files. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711−720.

(518) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe Up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.

(519) Earl, D. J.; Deem, M. W. Toward a Database of Hypothetical Zeolite Structures. *Ind. Eng. Chem. Res.* **2006**, *45*, 5449−5454.

(520) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.

(521) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513−530.

(522) Zitnick, C. L.; Chanussot, L.; Das, A.; Goyal, S.; Heras-Domingo, J.; Ho, C.; Hu, W.; Lavril, T.; Palizhati, A.; Riviere, M., et al. An Introduction to Electrocatalyst Design Using Machine Learning for Renewable Energy Storage. *arXiv*, 2020, 2010.09435. https://arxiv.org/abs/2010.09435.

(523) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery With High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501−1509.

(524) Nakata, M.; Shimazaki, T.; Hashimoto, M.; Maeda, T. PubChemQC PM6: Data Sets of 221 Million Molecules With Optimized Molecular Geometries and Electronic Properties. *J. Chem. Inf. Model.* **2020**, *60*, 5891−5899.

(525) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300−1308.

(526) Hoja, J.; Medrano Sandonas, L.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio Jr, R. A.; Tkatchenko, A. QM7-X: A Comprehensive Dataset of Quantum-Mechanical Properties Spanning the Chemical Space of Small Organic Molecules. *Sci. Data* **2021**, *8*, 43.

(527) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.

(528) Kim, E.; Huang, K.; Tomala, A.; Matthews, S.; Strubell, E.; Saunders, A.; McCallum, A.; Olivetti, E. Machine-Learned and Codified Synthesis Parameters of Oxide Materials. *Sci. Data* **2017**, *4*, 170127.

(529) Tetko, I. V.; Maran, U.; Tropsha, A. Public (Q)SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development. *Mol. Inf.* **2017**, *36*, 1600082.

(530) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole von Lilienfeld, O. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, 095003.

(531) 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.

(532) Ceriotti, M. Unsupervised Machine Learning in Atomistic Simulations, Between Predictions and Understanding. *J. Chem. Phys.* **2019**, *150*, 150901.

(533) Montavon, G.; Hansen, K.; Fazli, S.; Rupp, M.; Biegler, F.; Ziehe, A.; Tkatchenko, A.; Lillienfeld, A.; Muller, K.-R. Learning invariant representations of molecules for atomization energy prediction. *NeurIPS* **2012**, *25*, 440−448.

(534) Fraux, G.; Cersonsky, R.; Ceriotti, M. Chemiscope: Interactive Structure-Property Explorer for Materials and Molecules. *J. Open Source Softw.* **2020**, *5*, 2117.

(535) Qiang, Y.; Xindong, W. 10 Challenging Problems in Data Mining Research. *Int. J. Inf. Technol. Decis. Mak.* **2006**, *5*, 597−604.

(536) Wu, X.; Kumar, V.; Ross Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Yu, P. S.; et al. Top 10 Algorithms in Data Mining. *Knowl. Inf. Syst.* **2008**, *14*, 1−37.

(537) Li, H.; Zhang, Z.; Zhao, Z.-Z. Data-Mining for Processes in Chemistry, Materials, and Engineering. *Processes* **2019**, *7*, 151.

(538) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised Word Embeddings Capture Latent Knowledge From Materials Science Literature. *Nature* **2019**, *571*, 95−98.

(539) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.

(540) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights From Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436−9444.

(541) Kowalski, B. R. Chemometrics: Views and Propositions. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 201−203.

(542) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information From the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894−1904.

(543) Sparks, T. D.; Gaultois, M. W.; Oliynyk, A.; Brgoch, J.; Meredig, B. Data Mining Our Way to the Next Generation of Thermoelectrics. *Scr. Mater.* **2016**, *111*, 10−15.

(544) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge From Data Through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403−7429.

(545) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73−76.

(546) Xie, S.; Stewart, G.; Hamlin, J.; Hirschfeld, P.; Hennig, R. Functional Form of the Superconducting Critical Temperature From Machine Learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *100*, 174513.

(547) Staker, J.; Marshall, K.; Abel, R.; McQuaw, C. M. Molecular Structure Extraction From Documents Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1017−1029.

(548) Timoshenko, J.; Lu, D.; Lin, Y.; Frenkel, A. I. Supervised Machine-Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles. *J. Phys. Chem. Lett.* **2017**, *8*, 5091−5098.

(549) Ziatdinov, M.; Maksov, A.; Kalinin, S. V. Learning Surface Molecular Structures via Machine Vision. *Npj Comput. Mater.* **2017**, *3*, 31.

(550) Zhou, X. X.; Zeng, W. F.; Chi, H.; Luo, C.; Liu, C.; Zhan, J.; He, S. M.; Zhang, Z. PDeep: Predicting MS/MS Spectra of Peptides With Deep Learning. *Anal. Chem.* **2017**, *89*, 12690−12697.

(551) Car, R.; Parrinello, M. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55*, 2471−2474.

(552) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.

(553) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Adv. Mater.* **2019**, *31*, 1902765.

(554) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, DOI: 10.1021/acs.chemrev.0c01111.

(555) Dral, P. O.; Owens, A.; Yurchenko, S. N.; Thiel, W. Structure-Based Sampling and Self-Correcting Machine Learning for Accurate Calculations of Potential Energy Surfaces and Vibrational Levels. *J. Chem. Phys.* **2017**, *146*, 244108.

(556) Handley, C. M.; Popelier, P. L. Potential Energy Surfaces Fitted by Artificial Neural Networks. *J. Phys. Chem. A* **2010**, *114*, 3371−3383.

(557) Jiang, B.; Li, J.; Guo, H. Potential Energy Surfaces From High Fidelity Fitting of Ab Initio Points: The Permutation Invariant Polynomial - Neural Network Approach. *Int. Rev. Phys. Chem.* **2016**, *35*, 479−506.

(558) Kolb, B.; Zhao, B.; Li, J.; Jiang, B.; Guo, H. Permutation Invariant Potential Energy Surfaces for Polyatomic Reactions Using Atomistic Neural Networks. *J. Chem. Phys.* **2016**, *144*, 224103.

(559) Shao, K.; Chen, J.; Zhao, Z.; Zhang, D. H. Communication: Fitting Potential Energy Surfaces With Fundamental Invariant Neural Network. *J. Chem. Phys.* **2016**, *145*, 071101.

(560) Li, J.; Song, K.; Behler, J. A Critical Comparison of Neural Network Potentials for Molecular Reaction Dynamics With Exact Permutation Symmetry. *Phys. Chem. Chem. Phys.* **2019**, *21*, 9672−9682.

(561) Fu, B.; Zhang, D. H. Ab Initio Potential Energy Surfaces and Quantum Dynamics for Polyatomic Bimolecular Reactions. *J. Chem. Theory Comput.* **2018**, *14*, 2289−2303.

(562) Ballard, A. J.; Das, R.; Martiniani, S.; Mehta, D.; Sagun, L.; Stevenson, J. D.; Wales, D. J. Energy Landscapes for Machine Learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 12585−12603.

(563) Chen, J.; Xu, X.; Xu, X.; Zhang, D. H. A Global Potential Energy Surface for the $H_2 + OH \leftrightarrow H_2O + H$ Reaction Using Neural Networks. *J. Chem. Phys.* **2013**, *138*, 154301.

(564) Brown, D. F.; Gibbs, M. N.; Clary, D. C. Combining Ab Initio Computations, Neural Networks, and Diffusion Monte Carlo: An Efficient Method to Treat Weakly Bound Molecules. *J. Chem. Phys.* **1996**, *105*, 7597−7604.

(565) Bernstein, N.; Csányi, G.; Deringer, V. L. De Novo Exploration and Self-Guided Learning of Potential-Energy Surfaces. *Npj Comput. Mater.* **2019**, *5*, 99.

(566) Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. sGDML: Constructing Accurate and Data Efficient Molecular Force Fields Using Machine Learning. *Comput. Phys. Commun.* **2019**, *240*, 38−45.

(567) Pártay, L. B.; Bartók, A. P.; Csányi, G. Efficient Sampling of Atomic Configurational Spaces. *J. Phys. Chem. B* **2010**, *114*, 10502−10512.

(568) Thompson, A.; Swiler, L.; Trott, C.; Foiles, S.; Tucker, G. Spectral Neighbor Analysis Method for Automated Generation of Quantum-Accurate Interatomic Potentials. *J. Comput. Phys.* **2015**, *285*, 316−330.

(569) Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A.; et al. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *J. Phys. Chem. A* **2020**, *124*, 731−745.

(570) Behler, J. Representing Potential Energy Surfaces by High-Dimensional Neural Network Potentials. *J. Phys.: Condens. Matter* **2014**, *26*, 183001.

(571) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K. R. SchNetPack: A Deep Learning Toolbox for Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.

(572) Han, J.; Zhang, L.; Car, R.; E, W. Deep Potential: A General Representation of a Many-Body Potential Energy Surface. *Commun. Comput. Phys.* **2018**, *23*, 629−639.

(573) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(574) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.

(575) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.

(576) Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. How Van Der Waals Interactions Determine the Unique Properties of Water. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 8368−8373.

(577) Tuckerman, M. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford University Press: New York, NY, 2010.

(578) Cheng, B.; Engel, E. A.; Behler, J.; Dellago, C.; Ceriotti, M. Ab Initio Thermodynamics of Liquid and Solid Water. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 1110−1115.

(579) Reinhardt, A.; Cheng, B. Quantum-Mechanical Exploration of the Phase Diagram of Water. *Nat. Commun.* **2021**, *12*, 588.

(580) Niu, H.; Bonati, L.; Piaggi, P. M.; Parrinello, M. Ab Initio Phase Diagram And Nucleation of Gallium. *Nat. Commun.* **2020**, *11*, 2654.

(581) Cheng, B.; Mazzola, G.; Pickard, C. J.; Ceriotti, M. Evidence for Supercritical Behaviour of High-Pressure Liquid Hydrogen. *Nature* **2020**, *585*, 217−220.

(582) Cheng, B.; Behler, J.; Ceriotti, M. Nuclear Quantum Effects in Water at the Triple Point: Using Theory as a Link Between Experiments. *J. Phys. Chem. Lett.* **2016**, *7*, 2210−2215.

(583) Morawietz, T.; Marsalek, O.; Pattenaude, S. R.; Streacker, L. M.; Ben-Amotz, D.; Markland, T. E. The Interplay of Structure and Dynamics in the Raman Spectrum of Liquid Water Over the Full Frequency and Temperature Range. *J. Phys. Chem. Lett.* **2018**, *9*, 851−857.

(584) Ko, H.-Y.; Zhang, L.; Santra, B.; Wang, H.; E, W.; DiStasio, R. A., Jr.; Car, R. Isotope Effects in Liquid Water via Deep Potential Molecular Dynamics. *Mol. Phys.* **2019**, *117*, 3269−3281.

(585) Sauceda, H. E.; Chmiela, S.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. Molecular Force Fields With Gradient-Domain Machine Learning: Construction and Application to Dynamics of Small Molecules With Coupled Cluster Forces. *J. Chem. Phys.* **2019**, *150*, 114102.

(586) Bisbo, M. K.; Hammer, B. Efficient Global Structure Optimization With a Machine-Learned Surrogate Model. *Phys. Rev. Lett.* **2020**, *124*, 086102.

(587) Meldgaard, S. A.; Kolsbjerg, E. L.; Hammer, B. Machine Learning Enhanced Global Optimization by Clustering Local Environments to Enable Bundled Atomic Energies. *J. Chem. Phys.* **2018**, *149*, 134104.

(588) Deringer, V. L.; Pickard, C. J.; Proserpio, D. M. Hierarchically Structured Allotropes of Phosphorus From Data-Driven Exploration. *Angew. Chem., Int. Ed.* **2020**, *59*, 15880.

(589) Chiavazzo, E.; Covino, R.; Coifman, R. R.; Gear, C. W.; Georgiou, A. S.; Hummer, G.; Kevrekidis, I. G. Intrinsic Map Dynamics Exploration for Uncharted Effective Free-Energy Landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, No. E5494.

(590) Rogal, J.; Schneider, E.; Tuckerman, M. E. Neural-Network-Based Path Collective Variables for Enhanced Sampling of Phase Transformations. *Phys. Rev. Lett.* **2019**, *123*, 245701.

(591) Bonati, L.; Rizzi, V.; Parrinello, M. Data-Driven Collective Variables for Enhanced Sampling. *J. Phys. Chem. Lett.* **2020**, *11*, 2998−3004.

(592) Mones, L.; Bernstein, N.; Csányi, G. Exploration, Sampling, and Reconstruction of Free Energy Surfaces With Gaussian Process Regression. *J. Chem. Theory Comput.* **2016**, *12*, 5100−5110.

(593) Debnath, J.; Parrinello, M. Gaussian Mixture-Based Enhanced Sampling for Statics and Dynamics. *J. Phys. Chem. Lett.* **2020**, *11*, 5076−5080.

(594) Schneider, E.; Dai, L.; Topper, R. Q.; Drechsel-Grau, C.; Tuckerman, M. E. Stochastic Neural Network Approach for Learning High-Dimensional Free Energy Surfaces. *Phys. Rev. Lett.* **2017**, *119*, 150601.

(595) Bonati, L.; Zhang, Y.-Y.; Parrinello, M. Neural Networks-Based Variationally Enhanced Sampling. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 17641−17647.

(596) Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The Advent of Generative Chemistry. *ACS Med. Chem. Lett.* **2020**, *11*, 1496−1505.

(597) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360−365.

(598) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37*, 1700111.

(599) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery With Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(600) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37*, 1700153.

(601) Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A. Constrained Graph Variational Autoencoders for Molecule Design. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7795−7804.

(602) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *Proceedings of the 35th International Conference on Machine Learning* **2018**, 2323−2332.

(603) Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-Directed Variational Autoencoder for Structured Data. *arXiv*, **2018**, 1802.08786. https://arxiv.org/abs/1802.08786.

(604) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. *Proceedings of the 34th International Conference on Machine Learning* **2017**, 1945−1954.

(605) Jørgensen, P. B.; Mesta, M.; Shil, S.; García Lastra, J. M.; Jacobsen, K. W.; Thygesen, K. S.; Schmidt, M. N. Machine Learning-Based Screening of Complex Molecules for Polymer Solar Cells. *J. Chem. Phys.* **2018**, *148*, 241735.

(606) Jin, W.; Yang, K.; Barzilay, R.; Jaakkola, T. Learning Multimodal Graph-to-Graph Translation for Molecule Optimization. *arXiv*, **2019**, 1812.01070, ver. 3. https://arxiv.org/abs/1812.01070.

(607) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(608) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for De Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15*, 4398−4405.

(609) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* **2018**, *37*, 1700123.

(610) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for De Novo Generation of New Molecules With Desired Molecular Properties in Silico. *Mol. Pharmaceutics* **2017**, *14*, 3098−3104.

(611) Yu, L.; Zhang, W.; Wang, J.; Yu, Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.* **2017**, 2852−2858.

(612) Guimaraes, G. L.; Sanchez-Lengeling, B.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv*, **2017**, 1705.10843. https://arxiv.org/abs/1705.10843.

(613) De Cao, N.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. *arXiv*, **2018**, 1805.11973. https://arxiv.org/abs/1805.11973.

(614) Maziarka, Ł.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; Warchoł, M. Mol-CycleGAN: A Generative Model for Molecular Optimization. *J. Cheminf.* **2020**, *12*, 2.

(615) You, J.; Liu, B.; Ying, R.; Pande, V.; Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6410−6421.

(616) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for De Novo Drug Design. *Sci. Adv.* **2018**, *4*, No. eaap7885.

(617) Zhavoronkov, A.; et al. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.

(618) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective De Novo Drug Design With Conditional Graph Generative Model. *J. Cheminf.* **2018**, *10*, 33.

(619) Li, Y.; Vinyals, O.; Dyer, C.; Pascanu, R.; Battaglia, P. Learning Deep Generative Models of Graphs. *arXiv*, **2018**, 1803.03324. https://arxiv.org/abs/1803.03324.

(620) Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular Geometry Prediction Using a Deep Generative Graph Neural Network. *Sci. Rep.* **2019**, *9*, 20381.

(621) Gebauer, N. W. A.; Gastegger, M.; Schütt, K. T. Generating Equilibrium Molecules With Deep Neural Networks. *arXiv*, **2018**, 1810.11347. https://arxiv.org/abs/1810.11347.

(622) Gebauer, N.; Gastegger, M.; Schütt, K. Symmetry-Adapted Generation of 3d Point Sets for the Targeted Discovery of Molecules. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 7564−7576.

(623) Caccin, M.; Li, Z.; Kermode, J. R.; De Vita, A. A Framework for Machine-Learning-Augmented Multiscale Atomistic Simulations on Parallel Supercomputers. *Int. J. Quantum Chem.* **2015**, *115*, 1129−1139.

(624) Gkeka, P.; et al. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *J. Chem. Theory Comput.* **2020**, *16*, 4757−4775.

(625) Saunders, M. G.; Voth, G. A. Coarse-Graining Methods for Computational Biology. *Annu. Rev. Biophys.* **2013**, *42*, 73−93.

(626) John, S. T.; Csányi, G. Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials. *J. Phys. Chem. B* **2017**, *121*, 10934−10949.

(627) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. DeePCG: Constructing Coarse-Grained Models via Deep Neural Networks. *J. Chem. Phys.* **2018**, *149*, 034101.

(628) Cesari, A.; Gil-Ley, A.; Bussi, G. Combining Simulations and Solution Experiments as a Paradigm for RNA Force Field Refinement. *J. Chem. Theory Comput.* **2016**, *12*, 6192−6200.

(629) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38*, 1291−1307.

(630) Goh, G. B.; Vishnu, A.; Siegel, C.; Hodas, N. Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* **2018**, 302−310.

(631) Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191−201.

(632) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T.; et al. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120−1127.

(633) Kolb, B.; Lentz, L. C.; Kolpak, A. M. Discovering Charge Density Functionals and Structure-Property Relationships With PROPhet: A General Framework for Coupling Machine Learning and First-Principles Methods. *Sci. Rep.* **2017**, *7*, 1192.

(634) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space With Quantum-Based Machine Learning. *Nat. Rev. Chem.* **2020**, *4*, 347−358.

(635) Kuhn, C.; Beratan, D. N. Inverse Strategies for Molecular Design. *J. Phys. Chem.* **1996**, *100*, 10595−10599.

(636) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8*, 15679.

(637) Zhang, L.; Mao, H.; Liu, Q.; Gani, R. Chemical Product Design - Recent Advances and Perspectives. *Curr. Opin. Chem. Eng.* **2020**, *27*, 22−34.

(638) Park, C. W.; Wolverton, C. Developing an Improved Crystal Graph Convolutional Neural Network Framework for Accelerated Materials Discovery. *Phys. Rev. Mater.* **2020**, *4*, 063801.

(639) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(640) Dewar, M. J.; Storch, D. M. Comparative Tests of Theoretical Procedures for Studying Chemical Reactions. *J. Am. Chem. Soc.* **1985**, *107*, 3898−3902.

(641) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. *arXiv*, 2016, 1612.09529. https://arxiv.org/abs/1612.09529.

(642) Segler, M. H.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses With Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.

(643) Segler, M.; Preuss, M.; Waller, M. P. Towards "Alphachem": Chemical Synthesis Planning With Tree Search and Deep Neural Network Policies. *5th International Conference on Learning Representations, ICLR 2017—Workshop Track Proceedings*, 2019.

(644) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inf.* **2014**, *33*, 469−476.

(645) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.

(646) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178−192.

(647) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(648) Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-Aided Synthesis Design: 40 Years On. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 79−107.

(649) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103−1113.

(650) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned From a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58*, 252−261.

(651) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem.* **2018**, *4*, 522−532.

(652) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.* **2020**, *59*, 22858−22893.

(653) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem., Int. Ed.* **2020**, *59*, 23414−23436.

(654) Lindström, B.; Pettersson, L. J. A Brief History of Catalysis. *CATTECH* **2003**, *7*, 130−138.

(655) Cui, X.; Li, W.; Ryabchuk, P.; Junge, K.; Beller, M. Bridging Homogeneous and Heterogeneous Catalysis by Heterogeneous Single-Metal-Site Catalysts. *Nat. Catal.* **2018**, *1*, 385−397.

(656) Suen, N.-T.; Hung, S.-F.; Quan, Q.; Zhang, N.; Xu, Y.-J.; Chen, H. M. Electrocatalysis for the Oxygen Evolution Reaction: Recent Development and Future Perspectives. *Chem. Soc. Rev.* **2017**, *46*, 337−365.

(657) Benson, E. E.; Kubiak, C. P.; Sathrum, A. J.; Smieja, J. M. Electrocatalytic and Homogeneous Approaches to Conversion of CO_2 to Liquid Fuels. *Chem. Soc. Rev.* **2009**, *38*, 89−99.

(658) Steinfeld, A. Solar Thermochemical Production of Hydrogen—a Review. *Sol. Energy* **2005**, *78*, 603−615.

(659) Lee, J. H.; Seo, Y.; Park, Y. D.; Anthony, J. E.; Kwak, D. H.; Lim, J. A.; Ko, S.; Jang, H. W.; Cho, K.; Lee, W. H. Effect of Crystallization Modes in TIPS-pentacene/insulating Polymer Blends on the Gas Sensing Properties of Organic Field-Effect Transistors. *Sci. Rep.* **2019**, *9*, 21.

(660) Zeradjanin, A. R.; Grote, J.-P.; Polymeros, G.; Mayrhofer, K. J. A Critical Review on Hydrogen Evolution Electrocatalysis: Re-Exploring the Volcano-Relationship. *Electroanalysis* **2016**, *28*, 2256−2269.

(661) Wenderich, K.; Mul, G. Methods, Mechanism, and Applications of Photodeposition in Photocatalysis: A Review. *Chem. Rev.* **2016**, *116*, 14587−14619.

(662) Hou, W.; Cronin, S. B. A Review of Surface Plasmon Resonance-Enhanced Photocatalysis. *Adv. Funct. Mater.* **2013**, *23*, 1612−1619.

(663) Julliard, M.; Chanon, M. Photoelectron-Transfer Catalysis: Its Connections With Thermal and Electrochemical Analogs. *Chem. Rev.* **1983**, *83*, 425−506.

(664) Kavarnos, G. J.; Turro, N. J. Photosensitization by Reversible Electron Transfer: Theories, Experimental Evidence, and Examples. *Chem. Rev.* **1986**, *86*, 401−449.

(665) Bogaerts, A.; Tu, X.; Whitehead, J. C.; Centi, G.; Lefferts, L.; Guaitella, O.; Azzolina-Jury, F.; Kim, H.-H.; Murphy, A. B.; Schneider, W. F.; et al. The 2020 Plasma Catalysis Roadmap. *J. Phys. D: Appl. Phys.* **2020**, *53*, 443001.

(666) Van Durme, J.; Dewulf, J.; Leys, C.; Van Langenhove, H. Combining Non-Thermal Plasma With Heterogeneous Catalysis in Waste Gas Treatment: A Review. *Appl. Catal., B* **2008**, *78*, 324−333.

(667) Campos-Gonzalez-Angulo, J. A.; Ribeiro, R. F.; Yuen-Zhou, J. Resonant Catalysis of Thermally Activated Chemical Reactions With Vibrational Polaritons. *Nat. Commun.* **2019**, *10*, 4685.

(668) Ma, Z.; Zaera, F. *Encyclopedia of Inorganic and Bioinorganic Chemistry*; Wiley Online Library, 2014; pp 1−16.

(669) Anastas, P. T.; Crabtree, R. H. *Handbook of Green Chemistry*; Vol. 2: Green Catalysis, Heterogeneous Catalysis; Wiley-VCH: Weinheim, Germany, 2013.

(670) Trost, B. M. On Inventing Reactions for Atom Economy. *Acc. Chem. Res.* **2002**, *35*, 695−705.

(671) Sheldon, R. A.; Arends, I.; Hanefeld, U. *Green Chemistry and Catalysis*; Wiley-VCH: Weinheim, Germany, 2007.

(672) Hammer, B.; Nørskov, J. K. Theoretical Surface Science and Catalysis-Calculations and Concepts. *Adv. Catal.* **2000**, *45*, 71−129.

(673) Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. From the Sabatier Principle to a Predictive Theory of Transition-Metal Heterogeneous Catalysis. *J. Catal.* **2015**, *328*, 36−42.

(674) Calle-Vallejo, F.; Loffreda, D.; Koper, M. T.; Sautet, P. Introducing Structural Sensitivity Into Adsorption-Energy Scaling Relations by Means of Coordination Numbers. *Nat. Chem.* **2015**, *7*, 403−410.

(675) Kitchin, J. R. Machine Learning in Catalysis. *Nat. Catal.* **2018**, *1*, 230−232.

(676) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260−2297.

(677) Grajciar, L.; Heard, C. J.; Bondarenko, A. A.; Polynski, M. V.; Meeprasert, J.; Pidko, E. A.; Nachtigall, P. Towards Operando Computational Modeling in Heterogeneous Catalysis. *Chem. Soc. Rev.* **2018**, *47*, 8307−8348.

(678) Goldsmith, B. R.; Esterhuizen, J.; Liu, J. X.; Bartel, C. J.; Sutton, C. Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE J.* **2018**, *64*, 2311−2323.

(679) McCullough, K.; Williams, T.; Mingle, K.; Jamshidi, P.; Lauterbach, J. High-Throughput Experimentation Meets Artificial Intelligence: A New Pathway to Catalyst Discovery. *Phys. Chem. Chem. Phys.* **2020**, *22*, 11174−11196.

(680) Wexler, R. B.; Qiu, T.; Rappe, A. M. Automatic Prediction of Surface Phase Diagrams Using Ab Initio Grand Canonical Monte Carlo. *J. Phys. Chem. C* **2019**, *123*, 2321−2328.

(681) Ulissi, Z. W.; Singh, A. R.; Tsai, C.; Nørskov, J. K. Automated Discovery and Construction of Surface Phase Diagrams Using Machine Learning. *J. Phys. Chem. Lett.* **2016**, *7*, 3931−3935.

(682) Roling, L. T.; Li, L.; Abild-Pedersen, F. Configurational Energies of Nanoparticles Based on Metal-Metal Coordination. *J. Phys. Chem. C* **2017**, *121*, 23002−23010.

(683) Roling, L. T.; Choksi, T. S.; Abild-Pedersen, F. A Coordination-Based Model for Transition Metal Alloy Nanoparticles. *Nanoscale* **2019**, *11*, 4438−4452.

(684) Yan, Z.; Taylor, M. G.; Mascareno, A.; Mpourmpakis, G. Size-, Shape-, and Composition-Dependent Model for Metal Nanoparticle Stability Prediction. *Nano Lett.* **2018**, *18*, 2696−2704.

(685) Dean, J.; Taylor, M. G.; Mpourmpakis, G. Unfolding Adsorption on Metal Nanoparticles: Connecting Stability With Catalysis. *Sci. Adv.* **2019**, *5*, No. eaax5101.

(686) Núñez, M.; Lansford, J. L.; Vlachos, D. G. Optimization of the Facet Structure of Transition-Metal Catalysts Applied to the Oxygen Reduction Reaction. *Nat. Chem.* **2019**, *11*, 449−456.

(687) van der Maaten, L. Accelerating T-SNE Using Tree-Based Algorithms. *J. Mach. Learn. Res.* **2014**, *15*, 3221−3245.

(688) Zhong, M.; et al. Accelerated Discovery of CO_2 Electrocatalysts Using Active Machine Learning. *Nature* **2020**, *581*, 178−183.

(689) Artrith, N.; Kolpak, A. M. Understanding the Composition and Activity of Electrocatalytic Nanoalloys in Aqueous Solvents: A Combination of DFT and Accurate Neural Network Potentials. *Nano Lett.* **2014**, *14*, 2670−2676.

(690) Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J. Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal-Oxo Intermediate Formation. *ACS Catal.* **2019**, *9*, 8243−8255.

(691) Wexler, R. B.; Martirez, J. M. P.; Rappe, A. M. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni_2P From Nonmetal Surface Doping Interpreted via Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 4678−4683.

(692) O'Connor, N. J.; Jonayat, A. S.; Janik, M. J.; Senftle, T. P. Interaction Trends Between Single Metal Atoms and Oxide Supports Identified With Density Functional Theory and Statistical Learning. *Nat. Catal.* **2018**, *1*, 531−539.

(693) Griego, C. D.; Zhao, L.; Saravanan, K.; Keith, J. A. Machine Learning Corrected Alchemical Perturbation Density Functional Theory for Catalysis Applications. *AIChE J.* **2020**, *66*, No. 17041.

(694) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069−7077.

(695) Boes, J. R.; Mamun, O.; Winther, K.; Bligaard, T. Graph Theory Approach to High-Throughput Surface Adsorption Structure Generation. *J. Phys. Chem. A* **2019**, *123*, 2281−2285.

(696) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621.

(697) Bort, W.; Baskin, I. I.; Gimadiev, T.; Mukanov, A.; Nugmanov, R.; Sidorov, P.; Marcou, G.; Horvath, D.; Klimchuk, O.; Madzhidov, T.; Varnek, A. Discovery of Novel Chemical Reactions by Deep Generative Recurrent Neural Network. *Sci. Rep.* **2021**, *11*, 3178.

(698) Kayala, M. A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52*, 2526−2540.

(699) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.

(700) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks. *Nat. Mach. Intell.* **2021**, *3*, 144−152.

(701) Blondal, K.; Jelic, J.; Mazeau, E.; Studt, F.; West, R. H.; Goldsmith, C. F. Computer-Generated Kinetics for Coupled Heterogeneous/Homogeneous Systems: A Case Study in Catalytic Combustion of Methane on Platinum. *Ind. Eng. Chem. Res.* **2019**, *58*, 17682−17691.

(702) Rangarajan, S.; Maravelias, C. T.; Mavrikakis, M. Sequential-Optimization-Based Framework for Robust Modeling and Design of Heterogeneous Catalytic Systems. *J. Phys. Chem. C* **2017**, *121*, 25847−25863.

(703) Banerjee, S.; Sreenithya, A.; Sunoj, R. B. Machine Learning for Predicting Product Distributions in Catalytic Regioselective Reactions. *Phys. Chem. Chem. Phys.* **2018**, *20*, 18311−18318.

(704) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571*, 343−348.

(705) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, No. eaau5631.

(706) Ravasco, J. M.; Coelho, J. A. Predictive Multivariate Models for Bioorthogonal Inverse-Electron Demand Diels-Alder Reactions. *J. Am. Chem. Soc.* **2020**, *142*, 4235−4241.

(707) Davies, I. W. The Digitization of Organic Synthesis. *Nature* **2019**, *570*, 175−181.

(708) Li, J.; Eastgate, M. D. Making Better Decisions During Synthetic Route Design: Leveraging Prediction to Achieve Greenness-by-Design. *React. Chem. Eng.* **2019**, *4*, 1595−1607.

(709) Schneider, G.; Fechner, U. Computer-Based De Novo Design of Drug-Like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.

(710) Reker, D.; Schneider, G. Active-Learning Strategies in Computer-Assisted Drug Discovery. *Drug Discovery Today* **2015**, *20*, 458−465.

(711) Wainberg, M.; Merico, D.; Delong, A.; Frey, B. J. Deep Learning in Biomedicine. *Nat. Biotechnol.* **2018**, *36*, 829−838.

(712) Khaket, T. P.; Aggarwal, H.; Dhanda, S.; Singh, J. *Industrial Enzymes: Trends, Scope and Relevance*; Nova Science Publishers, Inc.: Hauppauge, NY, 2014; pp 110−143.

(713) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 5918−5922.

(714) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery With Explainable Artificial Intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573−584.

(715) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.

(716) Piroozmand, F.; Mohammadipanah, F.; Sajedi, H. Spectrum of Deep Learning Algorithms in Drug Discovery. *Chem. Biol. Drug Des.* **2020**, *96*, 886−901.

(717) Fjell, C. D.; Hiss, J. A.; Hancock, R. E.; Schneider, G. Designing Antimicrobial Peptides: Form Follows Function. *Nat. Rev. Drug Discovery* **2012**, *11*, 37−51.

(718) Batra, R.; Chan, H.; Kamath, G.; Ramprasad, R.; Cherukara, M. J.; Sankaranarayanan, S. K. Screening of Therapeutic Agents for COVID-19 Using Machine Learning and Ensemble Docking Studies. *J. Phys. Chem. Lett.* **2020**, *11*, 7058−7065.

(719) Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J. Chem ML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, No. 1458.

(720) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.

(721) von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angew. Chem., Int. Ed.* **2018**, *57*, 4164−4169.

(722) von Lilienfeld, O. A.; Burke, K. Retrospective on a Decade of Machine Learning for Chemical Discovery. *Nat. Commun.* **2020**, *11*, 4895.

(723) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49*, 6154−6168.

(724) Schütt, K. T.; Chmiela, S.; von Lilienfeld, O. A.; Tkatchenko, A.; Tsuda, K.; Müller, K.-R. *Machine Learning Meets Quantum Physics*; Springer Lecture Notes in Physics, Springer: Cham, Switzerland, 2020; Vol. *968*.

(725) Schnake, T.; Eberle, O.; Lederer, J.; Nakajima, S.; Schütt, K. T.; Müller, K.-R.; Montavon, G. XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks. *arXiv*, 2020, 2006.03589, ver. 1. https://arxiv.org/abs/2006.03589v1.