The Journal of Chemical Physics

Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry

Cite as: J. Chem. Phys. **148**, 241401 (2018); https://doi.org/10.1063/1.5043213
Submitted: 06 June 2018 . Accepted: 11 June 2018 . Published Online: 27 June 2018

Matthias Rupp, O. Anatole von Lilienfeld, and Kieron Burke

Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry

Matthias Rupp,[1,a)] O. Anatole von Lilienfeld,[2,b)] and Kieron Burke[3,c)]
[1]*Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, 14195 Berlin, Germany*
[2]*Department of Chemistry, Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, University of Basel, 4056 Basel, Switzerland*
[3]*Departments of Chemistry and Physics, University of California, Irvine, California 92697, USA*

(Received 6 June 2018; accepted 11 June 2018; published online 27 June 2018)

A survey of the contributions to the Special Topic on Data-enabled Theoretical Chemistry is given, including a glossary of relevant machine learning terms. *Published by AIP Publishing.*
https://doi.org/10.1063/1.5043213

NOMENCLATURE

AI	Artificial intelligence, see Sec. II A
B3LYP	Becke, three-parameter, Lee-Yang-Parr, a hybrid DFT functional
CCSD(T)	Coupled cluster with single, double and perturbative triple excitations, an electronic structure method
DFT	Density functional theory, an electronic structure method
DFTB	Density functional theory tight binding, an electronic structure method
DNN	Deep neural network, see Sec. II C
EAM	Embedded atom model/method, an interatomic potential
GAP	Gaussian approximation potential, a machine learning potential
HOMO	Highest occupied molecular orbital
KRR	Kernel ridge regression, see Sec. II C
LUMO	Lowest unoccupied molecular orbital
MAE	Mean absolute error, see Sec. II D
MD	Molecular dynamics, a simulation technique
ML	Machine learning, see Sec. II A
MP2	Møller-Plesset perturbation theory to second order, an electronic structure method
QM/MM	Quantum mechanics/molecular mechanics, a molecular simulation method
(A)NN	(Artificial) neural network, see Sec. II C
QSPR	Quantitative structure-property relationship, see Sec. II A
RMSE	Root mean squared error, see Sec. II D
SINDy	Sparse identification of nonlinear dynamics, a machine learning method
SNAP	Spectral neighbor analysis potential, a machine learning potential
SVM	Support vector machine, see Sec. II C
tICA	Time structure independent component analysis, see Sec. II C

I. INTRODUCTION

Welcome to the Journal of Chemical Physics Special Topic on data-enabled theoretical chemistry. We expect that this will be a timely addition to this new and rapidly evolving field, with a variety of articles from the front lines.

Unless you have disconnected from all social media, you will have noticed that artificial intelligence, machine learning, big data, and other vague but computer-driven terms have invaded many realms of public life. Facial recognition software has been revolutionized by machine learning, cars now drive themselves, the world's best chess and go players are algorithms, and perhaps someday soon they will even be able to recommend a good movie.

The same revolution has also been occurring in many branches of theoretical and computational chemistry, driven by the same force: the never-ending increase in data being generated by computers. Our Special Topic is devoted to data-enabled chemistry, which we interpret broadly. We cover essentially all algorithmic developments that fit under the broad rubric of machine learning, using varying amounts of data, and driven by applications from small molecule chemistry to materials science to protein behavior.

In Fig. 1, we show papers being published involving machine learning and chemistry or materials over the last three

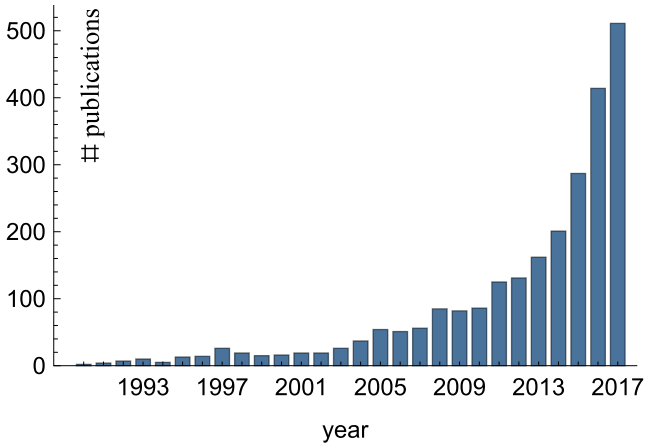

FIG. 1. *Number of publications per year* from a web of science search for articles with topics of machine learning and either chemistry or materials, taken June 5, 2018. The average number of citations per article is 12.

[a)]matthias.rupp@fhi-berlin.mpg.de. URL: www.mrupp.info.
[b)]anatole.vonlilienfeld@unibas.ch
[c)]kieron@uci.edu

TABLE I. *Overview* of contributions to the Special Topic.

Reference	Section	ML method	QM method	Systems	Keywords
30	III A	NN	DFT	Hydrocarbon molecules	Size-independence
31	III A	Multilinear regression	DFT	Small organic molecules	Representation, wavelets
32	III A	KRR	DFT	Organic molecules, water, solids	Representation, many-body terms
33	III A	NN	DFT	Small organic molecules	NN architecture
34	III A	NN	DFT	Small organic molecules	Representation, symmetry functions
35	III A	Regression	DFT	Small organic molecules	Polynomial fit, active learning
36	III A	KRR	DFT	Small organic molecules	Graph-based representation
76	III A	NN	DFT	Organic molecules	Covariant compositional networks
37	III B	KRR	DFT, CCSD(T)	Dimers, hydrogen-bonded complexes, and others	Non-covalent interactions
38	III B	GPR, NN	DFT	Liquid water, Al-Si-Mg alloy, organic molecules	Feature selection
39	III B	GPR	DFT	Li-C guest-host systems	Combination of potentials
40	III B	NN	DFT	Small organic molecules	Active learning
41	III B	NN	DFT	Small organic molecules	Molecular properties
42	III B	DNN	DFT	Organic molecules, bulk crystals, C_{20}-fullerene	DNN architecture
43	III B	GPR	DFT, force field	Na^+, Cl^- ion-water clusters	Ion-water interactions
44	III B	Regularized linear regression	DFT	Tantalum	Bispectrum quadratic terms
45	III B	GPR	DFT	Ni nanoclusters	Interatomic forces, k-body kernels
46	III B	NN	DFT	Nicotine, water cluster	Sampling, meta-dynamics
47	III B	NN	DFT	Cu surface grain boundaries	Hybrid QM/ML models
48	III C	NN	DFT	Water/ZnO($10\bar{1}0$) interface	Anharmonic vibrational spectra
49	III C	Linear regression	CCSD(T)	Formic acid dimer	Dipole moment surface, infrared spectrum
50	III C	NN, GPR	CCSD(T)	Water (ice, liquid, clusters)	Representation, invariant polynomials
51	III C	NN, GPR	Force field	Formaldehyde	Comparison, vibrational spectra
52	III D	NN, genetic algorithm	DFT	Li_xSi alloys	Phase diagrams of amorphous materials
53	III D	Regression trees	DFT	AB_2C_2 ternary intermetallics	Stable compound search
54	III D	Clustering	Harris approximation	Rigid-molecule crystals	Crystal structure prediction
55	III D	Monte Carlo tree search	EAM	Ag, Co grain boundaries	Segregation
56	III D	Binary classification trees	DFT	Inorganic crystals	Recommender system
57	III D	Monte Carlo tree search, GPR	DFT	Boron-doped graphene	Stable structure search
58	III E	Subset selection, outlier detection	DFT	Main group chemistry	Doubly hybrid functional
60	III E	NN	DFT	Model systems	Hartree-exchange-correlation potential
62	III E	KRR	DFT	Organic molecules	Representation
63	III E	KRR	DFT	Model systems	Exact conditions
64	III E	NN	DFT	Atoms and molecules	Kinetic energy density functional
65	III F	Sparse regression	Analytic potential	Model systems	Stochastic dynamical equations
66	III F	Time-lagged autoencoder	Force field	Model systems, villin peptide	Slow dynamics, dimensionality reduction
67	III F	Markov state model, tICA	Force field	Dye-labeled polyproline-20	Dynamics, transition probabilities
68	III G	None	DFT	Various (G3/99 test set)	Error statistics
69	III G	Autoencoder, NN	DFT	Donor-acceptor polymers	Screening, solar cells
70	III G	SVM	DFT	Organic polymers	Refraction index
71	III G	KRR	DFT	Perovskite oxides, elpasolite halides	Lanthanide-doped scintillators
72	III G	GPR	CCSD(T)	Small organic molecules	Geometry optimization
73	III G	Clustering	DFTB	Anatase TiO_2(001)	Global structure optimization
74	III G	SVM, graph analysis	Force field	Tyrosine phosphatase 1E	Proteins, dynamic allostery
75	III G	Data analysis	Force field	Antimicrobial peptides	Visualization

decades. The absolute rate is rather arbitrary, depending on the precise search terms, but the rapid growth is robust, as is the average citation rate of each article. There is no doubt that data-enabled chemistry is rapidly making a large impact in the field.

This editorial is designed for non-experts who are outside this field, and trying to figure out what is going on, and how they might want to get in on the action. We provide a brief glossary of machine-learning terms for non-experts in Sec. II, focusing on the concepts and algorithms used most often in physical chemistry and materials science. In Sec. III, using the introduced terminology, we briefly survey the contributions in this Special Topic, grouped by the physical and chemical processes and systems to which they are applied.

A nomenclature and a table are provided to aid the reader: the Nomenclature summarizes the used abbreviations and Table I presents an overview of all articles in the Special Topic, acting as a quick guide to the methods (both quantum chemical and computer science) and the systems included. Not only is it a quick way to find something in the issue, but it also represents a snapshot of the state of the field today.

II. SOME DATA-ENABLED TERMINOLOGY

This section is an introduction to common terminology in machine learning, with an emphasis on those concepts currently in use in the applications in this Special Topic. Terms used both in this editorial and throughout the Special Topic are set in small capitals, followed by their explanation. This is by no means a comprehensive explanation, and interested readers should consult further sources for more detailed explanations.

A. Machine learning and related scientific fields

MACHINE LEARNING (ML)[1,2] is an umbrella term referring to algorithms that improve with data ("learn from experience"),[3] mostly for analysis or prediction. Instead of being explicitly programmed to solve a specific problem, these algorithms rely on given data to make statements about new data. An example for a ML algorithm is regression (Fig. 2): Based on a finite number of points (EXAMPLES, SAMPLES), a function is inferred which enables predictions for new examples; the fit gets better the more examples there are. While ML encompasses many different tasks besides regression, such as classification, dimensionality reduction, clustering, anomaly detection, optimization, and offers a wide variety of specific algorithms, such as Gaussian process regression, support vector machines, principal component analysis, (deep) neural networks, the underlying principle of data-driven improvement remains the same.

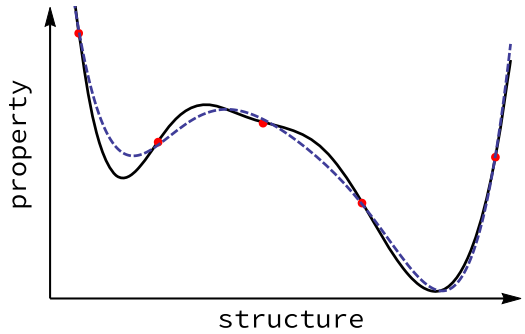

FIG. 2. *Sketch illustrating the idea of machine learning*,[4] using prediction of molecular energies as an example. The horizontal axis represents molecular space (molecules are points on the axis); the vertical axis represents energy. Instead of calculating all energies (solid line), only a few reference calculations are done (dots), and machine learning is used to learn the mapping from molecule to energy (dashed line).

ML is related to, but distinct from, artificial intelligence and data mining. ARTIFICIAL INTELLIGENCE (AI)[5] is the study of machines that exhibit intelligent behavior. The scope of this field is less clear-cut, evidenced by the lack of a formal definition of intelligence. AI traditionally involves (symbolic) knowledge representation and logical reasoning. DATA MINING is similar to ML but more concerned with extraction of new patterns in large datasets. PATTERN RECOGNITION is essentially a synonym for ML. For the more recent term DATA SCIENCE, no consensus has emerged yet, but it is often used to mean applied ML and statistics.

Two major application areas of ML closely related to this Special Topic are cheminformatics and materials informatics. CHEMINFORMATICS[6] (also chemoinformatics) is at the intersection of chemistry and computer science. In particular, QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIPS (QSPRs)[7] relate molecular features or descriptors to usually experimental and molecular properties, and VIRTUAL SCREENING[8] is the computational screening of large databases for compounds with desired properties. MATERIALS INFORMATICS[9] is a newer field at the intersection of materials science and computer science.

B. Types of problems machine learning address

One way to categorize problem types in ML is according to the types of examples involved. In SUPERVISED LEARNING, examples are pairs of input x and label y, for example molecules and their energy and the task is to predict the label of new examples, that is, to learn the function $f: x \to y$. In UNSUPERVISED LEARNING, only inputs x are given and the task is to find structure in the data. An example would be identifying a reaction coordinate from molecular dynamics (MD) data. Mixed forms are possible as well: In SEMI-SUPERVISED LEARNING, only some examples are labeled, with the idea that large amounts of unlabeled data can still help with predictions by characterizing the manifold on which the data lie. An example would be a large combinatorial chemistry database of molecules where only some have been measured or calculated.

Frequent types of problems within supervised learning are classification and regression. In CLASSIFICATION, labels belong to a finite set of outcomes, where one distinguishes between two possible labels in BINARY CLASSIFICATION, for example active and inactive, and, multiple possible labels in MULTI-CLASS CLASSIFICATION, for example different phases. The special case with only one possible label is ONE-CLASS LEARNING (also NOVELTY DETECTION, OUTLIER DETECTION, OR ANOMALY DETECTION), where examples from a single class are given and the task is to detect whether new examples fall outside of this class or not. In REGRESSION, labels are continuous. Usually, these are scalar values, but vectors, distributions, or other structured objects

like graphs can also be predicted using STRUCTURED-OUTPUT LEARNING.[10]

Frequent problem types within unsupervised learning are dimensionality reduction and clustering. In DIMENSIONALITY REDUCTION,[11] the goal is to find a subspace or manifold of low dimension on which the data live. CLUSTERING attempts to group samples into clusters such that samples within a cluster are more similar to each other than to samples in other clusters.

There are many other concepts that have found their way into data-enabled theoretical chemistry and materials science: In ACTIVE LEARNING,[12] the training data are not sampled randomly but "actively" chosen by the ML algorithm; this often enables achieving the same prediction error with much smaller training sets. In REINFORCEMENT LEARNING, the ML algorithm chooses an action from a set of possible actions based on the state of its environment. It is then rewarded accordingly and the process repeats. The goal of the algorithm is to maximize reward.

C. Specific algorithms

Many ML algorithms exist, but the ones used most often in cheminformatics and materials informatics belong to two large families, kernel-based ML and (deep) artificial neural networks.

In KERNEL-BASED ML,[13,14] inputs x are non-linearly transformed into a higher dimensional space, where problems can become linear with the right transformation. As working directly in these high-dimensional feature spaces is impractical, kernel functions k are used. These are computed in the original input space, but yield inner product values, and thus geometric information, in the high-dimensional space. Since their invention in the 1990s,[15,16] many linear ML algorithms have been "kernelized." Popular algorithms include SUPPORT VECTOR MACHINES (SVMs), KERNEL PRINCIPAL COMPONENT ANALYSIS,[15,16] KERNEL RIDGE REGRESSION (KRR),[17] and GAUSSIAN PROCESS REGRESSION (GPR)[18] (also called KRIGING due to its origins in geostatistics). While KRR is a frequentist algorithm and GPR is a Bayesian one, their predictions are formally identical, which is why the terms KRR and GPR are occasionally used interchangeably in practice.

Artificial NEURAL NETWORKS (NNs)[19,20] are repeated compositions of simple functions, where the inputs of one function are the weighted outputs of other functions. These functions are typically arranged in consecutive layers. In graph representations of NNs, vertices correspond to functions and edges correspond to weighted connections between them. Determining the weights is a non-convex optimization problem. DEEP NNs (DNNs)[21] are characterized by having many functional layers. This depth enables them to learn internal representations of the data of increasing complexity and abstraction.

Kernel learning and NN are simply two different ways of fitting a flexible function to data. Many other learning algorithms exist, including tree-based algorithms such as DECISION TREES, REGRESSION TREES, and RANDOM FORESTS.

A classic algorithm for dimensionality reduction is PRINCIPAL COMPONENT ANALYSIS (PCA),[22,23] which finds orthogonal directions of maximal variance in the data. Many variants of this idea exist, such as Independent Component Analysis (ICA), which finds independent latent variables and explains data as mixtures of these variables. For TIME-STRUCTURE INDEPENDENT COMPONENT ANALYSIS (tICA), these variables are chosen to maximize autocorrelation. A NN approach to dimensionality reduction are AUTOENCODER networks, where the size of functional layers first decreases, then increases again and the task is to reproduce the inputs. Having the data go through a "bottleneck" forces the autoencoder NN to find a low-dimensional representation of the data.

D. Model building

Unlike classical potentials, which are parametrized once for a class of molecules or materials and then deployed, ML models, being more flexible mathematical functions, should be applied only to molecules or materials sampled from the same distribution as the ones used to train the model—otherwise, the ML model will operate outside of its DOMAIN OF APPLICABILITY, resulting in uncontrolled and essentially arbitrary errors. For this reason, ML models are often retrained, for example dynamically by adding training data "on-the-fly" during the course of a simulation. Deciding when to make a prediction and when to do a reference calculation to update the model requires UNCERTAINTY ESTIMATES, that is, assessments of the reliability of individual predictions.

The ROOT MEAN SQUARED ERROR (RMSE) is the canonical measure of how wrong a set of predictions is. It is the RMSE that is minimized by many algorithms by default. This typically leads to "full" solutions, such as all coefficients in an expansion being non-zero. By contrast, SPARSITY of solutions, that is, solutions with most coefficients zero, can be achieved by minimizing the MEAN ABSOLUTE ERROR (MAE) or L^1-norm instead.

For VALIDATION of a ML model, the errors reported must always be on OUT-OF-SAMPLE data, that is, data not used for training the model, including any pre-processing steps. An easy way to achieve this is to set aside a HOLD-OUT SET in the beginning, to be used only for validation, and only after the ML model's training is complete. For small datasets, where this might not be feasible, statistical validation techniques such as cross-validation can be used. These essentially reuse the data by splitting it multiple times into the training and hold-out set and then average over the results.

The training or model-building process can include steps such as optimizing free parameters, often called HYPERPARAMETERS, or, FEATURE SELECTION, where only some of the descriptors or variables used to represent the inputs are retained. HYPERPARAMETER OPTIMIZATION usually is a non-convex optimization problem, but well-behaved in practice. For few parameters, it can be addressed via grid search, minimizing the hold-out RMSE over a logarithmic grid; alternatives are maximizing the likelihood of the model given the data, or choosing good values via heuristics.

The out-of-sample error of ML models must decay with training set size (otherwise it would not be machine *learning*). For many models, the leading error term varies as a/N^b, where N is number of training data.[24–26] LEARNING CURVES are plots of the out-of-sample prediction error as a function of N, usually on a log-log scale.

III. SURVEY OF AREAS COVERED

We next survey the areas covered by the articles in our Special Topic. We have organized them according to the type of chemical problem being addressed, as far as is possible. This makes it easier to see both the breadth of the problems and which topics have the most interest, as well as to compare different ML approaches to the same problem.

A. Prediction of energies and other properties throughout chemical compound space

Chemical space is astronomically vast.[27,28] Given some molecule, defined by its number of electrons and the set of nuclei at their equilibrium geometries, we can typically predict its observables with satisfying accuracy using *ab initio* quantum chemical methods such as CCSD(T) in a sufficiently large basis. This is feasible for smaller molecules, and DFT can be used (less reliably) for larger ones. But even DFT (or computationally less demanding semi-empirical quantum chemistry methods) is not fast enough to search all of chemical compound space, whose size grows combinatorially with the number of atoms and distinct elements. Thus, an important problem is to search chemical compound space to find new drugs (and materials space to find new materials) with desired functionalities.

A basic property is the ground-state energy of a molecule. But there are also many other interesting properties at the ground-state configuration, such as dipole moments, ionization potentials, and vibrational frequencies. Some of these can be extracted from the same electronic structure calculation from which the molecule's energy was obtained, while others require additional computation. Given the impossibility of calculating all properties of all possible molecules, it is interesting to ask if a ML algorithm, trained on known examples, can be used to predict the properties of new molecules at much reduced computational cost.[29] If so, chemical compound space can be searched orders of magnitude more quickly. Many groups are therefore formulating ways to do this.

Note that often researchers use DFT (or even DFTB) results for both training and testing their algorithms. In those cases, the ML algorithm is tested against the DFT calculations, not experiments or more accurate quantum chemical methods. The idea is that, once an algorithm is sufficiently robust and useful, it can then be trained on more accurate data and, presumably, work just as well. These days, many ML approaches already produce MAE below those typical of density functionals.

Yang *et al.*[30] introduce a size-independent NN model of heats of formation trained on small organic molecules that can be applied to large molecules. For these, the MAE from reference B3LYP numbers is reduced to 1.7 kcal/mol.

On the other hand, Eickenberg *et al.*[31] introduce a ML model based on a solid harmonic wavelet scattering representation of organic molecules and demonstrated competitive performance for predicted atomization energies. Meanwhile, Hy *et al.*[76] use a new kind of NN, called a covariant compositional network, to deduce properties from molecular graphs alone, yielding promising results on databases of small molecules.

Often, the efficiency of a ML algorithm depends crucially on the way the data are represented. Faber *et al.*[32] introduce a many-body representation of atoms in their environment and reported "chemical accuracy" (1 kcal/mol) for energies of organic molecules and solids with few thousand training points. Interpolation across the periodic table even enables prediction of energies of molecules with elements that were not included in the training set.

Lubbers *et al.*[33] introduce a hierarchical NN approach with competitive performance for predicting atomization energies of organic molecules, as well as energies and forces of thousands of snapshots of benzene, malonaldehyde, salicyclic acid, and toluene. Their method can also be applied to MD simulations and gives a measure of model uncertainty automatically. Gastegger *et al.*[34] develop element-specific weighting functions for atom-centered symmetry function-based representations in NN. Upon use of the weighting functions, they showed that less symmetry functions are necessary and the prediction error of atomization energies in organic molecules is systematically reduced.

Gubaev *et al.*[35] conceive a local tensor based ML approach which depends on the property being intensive or extensive, and they combined it with ACTIVE LEARNING in order to achieve state-of-the-art performance for atomization energies, polarizabilities, and HOMO/LUMO eigenvalues in organic molecules. Collins *et al.*[36] show that graph-based molecular representations profit from inclusion of interatomic distance information while remaining size-independent, as evinced for competitive prediction errors of atomization energies in organic molecules.

B. Interatomic potentials

Classical MD simulations with interatomic potentials can handle a million atoms or more and are used to study dynamic processes in biology and chemistry. Unfortunately, the necessary computational efficiency is sometimes obtained only at the expense of predictive power. Typically, relying on complex classical force fields, which ignore the underlying electronic structure and dynamics, can produce inconsistent answers to important questions. This limitation becomes especially acute when covalent bonds are formed or broken, when atoms vary their hybridization state, or during considerable changes in chemical environments, as in, for example, molten alloys. Then developing and testing force fields for all possible configurations become an unsurmountable task. Given this challenge, and the relevance of a dynamic description of atomistic processes throughout the exact sciences, a large number of articles in the Special Topic are devoted to the question if and how interatomic potentials can be constructed via ML, for example by training on (usually) DFT calculations.

Bereau *et al.*[37] predict parameters for intermolecular force fields throughout chemical space. These parameters include atomic charges, dipole moments, quadrupole moments, polarizabilities, atomic electron density screening factors, and normalization constants. Out-of-sample predictions on well-established van der Waals benchmark datasets indicate errors below or about 1 kcal/mol.

A crucial consideration for ML methods is the way that the inputs are represented, which can have a strong impact on performance. Imbalzano *et al.*[38] provide an automated protocol for FEATURE SELECTION, showing how this can simplify construction of ML potentials. They illustrated their procedure on NN potentials for water and aluminum ternary alloys, as well as a GPR potential for formation energies of molecules.

Gaussian approximation potentials (GAPs) are one of the success stories of ML in chemistry. They provide an automated approach to constructing accurate interatomic potentials that recreate the underlying electronic structure energetics at a fraction of the computational cost. Fujikake *et al.*[39] study the issue of guest atoms in host structures, with the specific case of Li in C, showing how to add the Li interactions to a pre-existing GAP potential for C.

An important question, usually left to human bias and intuition, is the selection of data upon which to train: When generating an interatomic potential, which sets of electronic structure calculations do you perform to create the database to train on and test against? Smith *et al.*[40] present a fully automatic way of generating datasets for the specific purpose of training ML potentials. Query-by-committee ACTIVE LEARNING uses disagreements between predictions of different models to improve sampling and reduce the amount of data needed over random sampling. Results are given on a new COMP6 database of small organic molecules containing CHNO.

Unke and Meuwly[41] are focused on creating methods that span both configurational space and chemical space. Their method decomposes energy into local atomic contributions, with prediction errors on atomization energies on the order of half a kcal/mol after training on 35 000 organic molecules. They demonstrate predictive capability on both reactive and non-reactive MD simulations.

Advanced deep learning methods are applied by Schütt *et al.*[42] They present SchNet, a DNN that learns chemically relevant information about atom types across the periodic table. It is general and flexible and uses deep learning to avoid the need for clever choices of descriptors. It can be applied to both molecules and materials and has been shown to reduce the computational cost of DFT-MD simulations of fullerenes by 3–4 orders of magnitude.

Another type of ML method is GPR or KRIGING, and Di Pasquale *et al.*[43] use it to predict energies of ions solvated in water. Energies are based on atomic energies obtained from the topological partitioning called interacting quantum atoms. This method provides accurate results and is part of an advanced force field development, FFLUX.

Spectral Neighbor Analysis Potentials (SNAPs) express the energy of an atom linearly in terms of bispectrum components of neighboring atoms. Wood and Thompson[44] show that accuracy can be improved by including quadratic contributions at a modest increase in cost, making it particularly suitable for large-scale MD simulations of materials.

Metallic nanoclusters are important in many areas of chemistry, but realistic simulations are limited by the computational cost of DFT-MD. Zeni *et al.*[45] study such systems via classical n-body potentials derived from ML ("M-FFs") by constructing n-body kernels that can be exactly mapped to non-parametric classical potential forms such as 3D splines. This circumvents summing over training set entries for predictions, accelerating simulations by orders of magnitude. They find that 2-body potentials are insufficiently accurate to capture the behavior of Ni clusters, but 3-body potentials are. Choice of training data also plays an essential role.

Another important question is which regions of configuration space to sample when constructing a ML force field. Herr *et al.*[46] explore application of metadynamics to training sets prior to selection for training. Metadynamics avoids the problem of being stuck in the vicinity of local minima. In comparison to data retrieved from MD or normal-mode analysis based sampling, the resulting NN exhibits improved or more efficient performance.

Finally, QM/MM schemes are popular in computational molecular biology but often suffer from limitations of the MM model and ambiguities at the interface. Zhang *et al.*[47] review this field for the specific case of a ML force field for the MM contribution. They point out both advantages and disadvantages of the ML approach.

C. Potential energy surfaces of specific molecules

This section could arguably be part of the previous one. But in this section, the molecule is fixed, and a highly accurate potential energy surface is desired, for a fixed number of atoms.

A difficult problem is the simulation of water on oxide surfaces, as measured by infrared spectroscopy of OH anharmonic stretches. MD simulations at the DFT level should be sufficiently accurate but are too expensive computationally. Quaranta *et al.*[48] use a NN potential trained on such calculations to perform MD and solve the nuclear Schrödinger equation for a large number of configurations to determine vibrational spectra. They found that many different species contribute in overlapping regions of the spectrum and that the stretching frequencies depend strongly on the hydrogen bonding.

For many purposes, DFT-level calculations suffice but not for the infrared spectrum of weakly bound dimers. The potential energy surface is a function of all 45 internuclear distances and must be calculated at CCSD(T) levels of accuracy in order to accurately reflect the anharmonic couplings. Qu and Bowman[49] present a novel fit to the dipole moment and solve the nuclear Schrödinger equation using various levels of anharmonic theory to generate the infrared spectrum.

Nguyen *et al.*[50] perform a careful study of the general methodology for constructing interatomic potentials, focusing on two- and three-body interactions in water using coupled-cluster energies. They compare different approaches: GAP, NN, and permutation-invariant polynomials, finding comparable levels of accuracy in the fit.

In a related way, Kamath *et al.*[51] study the potential energy surface of formaldehyde, in order to compare NN with GPR, using exactly the same data. In each case, they calculate vibrational spectra. They found GPR to perform better for a fixed number of data points, with a relatively accurate spectrum from as few as 300 data points.

D. Stability of solids

Another important field is the relative stability of different arrangements of atoms in solids, be they metallic alloys

or molecular crystals. Searching all possible arrangements is again a Herculean task, which could be tremendously accelerated if the patterns of the output could be machine-learned instead of having to be recalculated over and over.

Artrith *et al.*[52] address the problem of creating atomic potentials for alloys. There are a few cases where good potentials have been intuited in the past, but the essentially infinite number of possibilities and simulation conditions leads to a strong need for automation. Essentially, direct simulation with first-principles methods is hopelessly expensive for many problems and properties of interest. They use NN to speed up the sampling for amorphous and disordered materials and use the subsequent potential to calculate the phase diagram.

On the other hand, Schmidt *et al.*[53] scan many materials, looking specifically at ternary compounds to find the most stable structures. Here they find that ML reduces the calculational cost by about a factor of 4, but the high accuracy needed for such predictions limits the benefits of the ML approach to this problem.

An important problem is that of finding stable polymorphs of molecular crystals. Li *et al.*[54] introduce Genarris, a Python package that does inexpensive approximate DFT calculations and analyzes results with a relative coordinate descriptor developed specifically for this task. It uses ML for CLUSTERING and can be targeted for various outcomes, ranging from random structure generation to finding a maximally diverse set of structures to seed a genetic algorithm.

A quite different problem is that of grain boundaries in materials, where all sorts of non-stoichiometric defects appear. Kiyohara and Mizoguchi[55] use a Monte Carlo tree search to model grain-boundary segregation and test it on silver impurities in copper. They find that the search algorithm reduces the number of evaluations by a factor of 100 and yields insight into the nature of the most relevant sites.

Returning to searching chemical compound space, Seko *et al.*[56] look at all possible inorganic crystals, which is a much vaster space than those that have been discovered so far. They propose descriptors to estimate the relevance of chemical composition to stability. They train and test on experimental databases and also estimate phase stability from first-principles calculations.

Graphene is a promising material for future electronic applications. Dieb *et al.*[57] consider doping graphene with boron atoms. High levels of doping have been recently made and measured. Their aim is to find the most stable structures, using first principles calculations and ML to perform the search. They find useful patterns and predict properties as a function of boron doping.

E. Finding new density functionals

Density functional theory (DFT) calculations are currently of limited accuracy and reliability, and often fail badly for materials that are of key technological interest. Several of the papers in this Special Topic address the idea of using ML to improve existing functionals or to create entirely new ones.

Mardirossian and Head-Gordon[58] develop ML technology to optimize exchange-correlation functionals at different levels on Jacob's ladder[59] of increasing sophistication. Their work is at the highest rung, in which a double-hybrid functional is optimized (but not overfitted) to a dataset of nearly 5000 molecular energies, screening trillions of possible functionals, but ending up with only 14 parameters. This might prove an invaluable combination of accuracy and computational efficiency.

Another place where ML methods can be fruitfully applied is to find the exact (or at least a much more accurate) exchange-correlation functional, without fitting a given form of approximation. Nagai *et al.*[60] take small model problems, in which the exact density and energy are known, and use inversion techniques to find the exact Hartree-exchange-correlation energy and potentials. In the framework of Levy and Zahariev,[61] they then train and test a NN for this object. This work can be classified as going beyond the existing approximations used currently in DFT.

On the other hand, Ji and Jung[62] use a grid-based local representation of various electronic properties to predict DFT energies, densities, and exchange-correlation potentials for 16 small main-group molecules, with errors below 1 kcal/mol when trained for each molecule separately. The errors rise only to 4 kcal/mol if a small subset of the molecules is used for training, holding out the promise of a transferable method sensitive to the chemical environment.

The work of Hollingsworth *et al.*[63] is focused on whether or not simple exact conditions, which have been highly useful in guiding human-based functional design, are useful for improving learning curves of ML functional approximations. While they examine the question for the Kohn-Sham kinetic energy of simple models, their results should provide a guide for applications to the exchange-correlation energy, such as in the work of Nagai *et al.*[60] They find that, while exact conditions do improve learning rates, the improvement is only significant when there is similarity in the densities within the training manifold.

Seino *et al.*[64] work with approximate forms for the energy density of the Kohn-Sham kinetic energy to improve over existing approximations to orbital-free DFT. They expanded in higher gradients than are typically included in human approximations, and use ML to find coefficients and density dependencies, and compare their accuracies to many existing orbital-free functionals.

F. Analyzing molecular dynamics simulations

Even with classical force fields, there is tremendous interest in speeding up specific aspects of MD simulations, such as rare-event sampling or slow, long-term motions of long molecules. A related interest is the extraction of information from the large amounts of data generated by MD simulations.

The work of Boninsegna *et al.*[65] is focused on finding collective variables to determine long-time and coarse-grained motions from MD data. There is substantial history of *ad hoc* intuitive approaches to these problems, but their Sparse Identification of Nonlinear Dynamics (SINDy) approach does this automatically, and they prove the correctness of their approach in the limit of infinite data. A similar problem is tackled by Wehmeyer and Noé[66] using a DNN AUTOENCODER, which finds low-dimensional features (that is, the slow dynamics of the underlying stochastic processes) embedded in a higher

dimensional feature space. They test their methodology on simple model systems and a 125 μs trajectory of the fast-folding peptide villin.

Finally, Matsunaga and Sugita[67] approach this topic from a different viewpoint. They construct a Markov state model from MD trajectories and then refine that model using ML methods applied to experimental data. Thus their methodology attempts to overcome the inherent limitations of the MD force field model by comparison with experiments, whereas the other contributions are focused on speeding up a calculation, but entirely within the MD simulation itself.

G. Everything else

Not everything fits into simple categories and that is especially true in this field, including attempts to improve geometry optimization, to analyze the statistics behind benchmark datasets, and applications to larger biopolymers. In fact, there are many, many more possible applications of data-enabled chemistry, many of which are not included in this Special Topic and so are beyond the scope of this editorial.

Pernot and Savin[68] perform an in-depth study of the methods currently being used to benchmark approximations against datasets, an important topic as ever larger datasets are being generated. They question the summary statistics typically reported, such as RMSE or MAE, showing that because the error distributions are not simple, little can be inferred about error probabilities from these numbers alone. They advocate more informative measures and show their usefulness.

The position of the LUMO and the width of the optical gap in polymers for solar cells are important for power conversion efficiency. Jørgensen *et al.*[69] perform first-principles calculations on about 4000 monomers and show that a grammar variational AUTOENCODER using a simple string representation makes quite accurate predictions, reducing the cost of a search by up to a factor of 5. Afzal *et al.*[70] model the refraction index of organic polymers by combining first-principles calculations with ML to predict packing fractions of the bulk polymers.

Again, along the lines of solving a material- and property-specific problem, Pilania *et al.*[71] study the effect of lanthanide dopants in inorganic scintillation counter materials. They use ML on some key experimentally measured parameters and combine the results with high-throughput electronic structure calculations to perform screening for materials that exhibit optimized levels of the dopant relative to the gap of the host material.

Another important problem is that of geometry optimization, sometimes at a high level of theory. Schmitz and Christiansen[72] use GPR to optimize geometries using numerical gradients. They use lower levels of electronic structure calculations, such as Hartree-Fock or MP2, and then calculate differences to higher level theory. The interpolation introduces errors of no more than microHartrees.

In a similar vein, Sørensen *et al.*[73] also perform geometry optimization but on materials at an approximate DFT level. They find that UNSUPERVISED LEARNING can be used to categorize atoms in many diverse partially ordered surface structures of anatase titanium oxide. They also perform gradient-based minimization of a summed cluster distance resulting from this analysis which allows escape from meta-stable basins and so helps find global minima more quickly.

On the other hand, in a totally different system and regime, Botlani *et al.*[74] use MD to simulate dynamic allostery, in which regulator-induced changes in protein structure are comparable to thermal changes. Thus the data must be mined to find patterns in a very high dimensional space to identify mechanisms. UNSUPERVISED CLUSTERING shows that regulator binding strongly alters the protein's signalling network, not by changing connections between amino acids as one might naively imagine, but rather by changing the connectivity between clusters.

Antimicrobial peptides interact with simple phospholipid membranes, which is relevant for rational drug design. Cipcigan *et al.*[75] introduce new tools for analyzing the k-mer spectrum encoded in antimicrobial databases and ways to visualize membrane binding and permeation of helical peptides.

IV. SUMMARY

We hope you have found this editorial a useful guide to the important content, the papers in our Special Topic. We end with some remarks about the nature of the field. ML has been scoring some impressive successes in various areas of human activity. There is tremendous hope for similar successes in applications to physical sciences. However, progress in this direction requires discovering more subtle rules than in many other arenas. So it takes time for researchers to find the best ways to apply ML to their problems. But practical chemists and materials scientists can now create a dazzling array of different molecular structures and alloys. Once the progress reported here moves beyond development and proof-of-principle, perhaps we can look forward to new materials and drugs designed with ML methods that build on human intuition but apply it to more possibilities than a human could ever imagine. We shall see.

ACKNOWLEDGMENTS

K.B. acknowledges NSF 1464795. M.R. acknowledges funding from the EU Horizon 2020 program Grant No. 676580, The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence. O.A.v.L. acknowledges funding from the Swiss National Science Foundation (Nos. PP00P2_138932 and 310030_160067). This work was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. The guest editors sincerely thank the staff and editors of J. Chem. Phys. for putting this Special Topic together and all the authors for their input into this editorial.

[1] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," Nature **521**(7553), 452–459 (2015).
[2] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science **349**(6245), 255–260 (2015).
[3] T. M. Mitchell, *Machine Learning* (McGraw Hill, New York, 1997).
[4] M. Rupp, "Machine learning for quantum mechanics in a nutshell," Int. J. Quantum Chem. **115**(16), 1058–1073 (2015).
[5] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Prentice Hall, 2009).
[6] *Chemoinformatics*, edited by J. Gasteiger and T. Engel (Wiley-VCH, Weinheim, 2003).

[7]C. Selassie and R. P. Verma, "History of quantitative structure-activity relationships," in *Burger's Medicinal Chemistry*, 7th ed., Drug Discovery and Development Vol. 1, edited by D. J. Abraham and D. P. Rotella (Wiley, 2010).

[8]G. Schneider, "Virtual screening: An endless staircase?," Nat. Rev. Drug Discovery **9**(7), 273–276 (2010).

[9]R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning and materials informatics: Recent applications and prospects," Nat. Partner J. Comput. Mater. **3**, 54 (2017).

[10]*Predicting Structured Data*, edited by G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and V. Vishwanathan (MIT Press, Cambridge, 2007).

[11]J. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction* (Springer, New York, 2007).

[12]B. Settles, *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning Vol. 18 (Morgan & Claypool, 2012).

[13]B. Schölkopf and A. Smola, *Learning with Kernels* (MIT Press, Cambridge, 2002).

[14]T. Hofmann, B. Schölkopf, and A. Smola, "Kernel methods in machine learning," Ann. Stat. **36**(3), 1171–1220 (2008).

[15]B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Conference on Computational Learning Theory (COLT 1992), Pittsburgh, Pennsylvania, USA, July 27–29, 1992* (ACM, 1992), pp. 144–152.

[16]B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comput. **10**(5), 1299–1319 (1998).

[17]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Data Mining, Inference, and Prediction (Springer, New York, 2009).

[18]C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, 2006).

[19]C. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1996).

[20]*Neural Networks: Tricks of the Trade*, 2nd ed., Lecture Notes in Computer Science Vol. 7700, edited by G. Montavon, G. B. Orr, and K.-R. Müller (Springer, Berlin, Germany, 2012).

[21]I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).

[22]K. Pearson, "On lines and planes of closest fit to systems of points in space," Philos. Mag. **2**(11), 559–572 (1901).

[23]I. Jolliffe, *Principal Component Analysis*, 2nd ed. (Springer, New York, 2004).

[24]C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker, "Learning curves: Asymptotic values and rate of convergence," in *Advances in Neural Information Processing Systems 6 (NIPS 1993), Denver, Colorado, USA, November 29–December 2*, edited by J. D. Cowan, G. Tesauro, and J. Alspector (Morgan Kaufmann, 1993).

[25]V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. (Springer, 2001).

[26]K.-R. Müller, M. Finke, N. Murata, K. Schulten, and S. Amari, "A numerical study on learning curves in stochastic multilayer feedforward networks," Neural Comput. **8**(5), 1085–1106 (1996).

[27]A. Mullard, "The drug-maker's guide to the galaxy," Nature **549**(7673), 445–447 (2017).

[28]P. Kirkpatrick and C. Ellis, "Chemical space," Nature **432**, 823 (2004).

[29]O. A. von Lilienfeld, "Quantum machine learning in chemical compound space," Angew. Chem., Int. Ed. **57**(16), 4164–4169 (2018).

[30]G. Y. Yang, J. Wu, S. G. Chen, W. J. Zhou, J. Sun, and G. H. Chen, "Size-independent neural networks based first-principles method for accurate prediction of heat of formation of fuels," J. Chem. Phys. **148**(24), 241738 (2018).

[31]M. Eickenberg, G. Exarchakis, M. Hirn, S. Mallat, and L. Thiry, "Solid harmonic wavelet scattering for predictions of molecule properties," J. Chem. Phys. **148**(24), 241732 (2018).

[32]F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, "Alchemical and structural distribution based representation for universal quantum machine learning," J. Chem. Phys. **148**(24), 241717 (2018).

[33]N. Lubbers, J. S. Smith, and K. Barros, "Hierarchical modeling of molecular energies using a deep neural network," J. Chem. Phys. **148**(24), 241715 (2018).

[34]M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, "WACSF—weighted atom-centered symmetry functions as descriptors in machine learning potentials," J. Chem. Phys. **148**(24), 241709 (2018).

[35]K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, "Machine learning of molecular properties: Locality and active learning," J. Chem. Phys. **148**(24), 241727 (2018).

[36]C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, "Constant size descriptors for accurate machine learning models of molecular properties," J. Chem. Phys. **148**(24), 241718 (2018).

[37]T. Bereau, R. A. DiStasio, Jr., A. Tkatchenko, and O. A. von Lilienfeld, "Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning," J. Chem. Phys. **148**(24), 241706 (2018).

[38]G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials," J. Chem. Phys. **148**(24), 241730 (2018).

[39]S. Fujikake, V. L. Deringer, T. Hoon Lee, M. Krynski, S. R. Elliott, and G. Csányi, "Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures," J. Chem. Phys. **148**(24), 241714 (2018).

[40]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys. **148**(24), 241733 (2018).

[41]O. T. Unke and M. Meuwly, "A reactive, scalable, and transferable model for molecular energies from a neural network approach based on local information," J. Chem. Phys. **148**(24), 241708 (2018).

[42]K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**(24), 241722 (2018).

[43]N. Di Pasquale, S. J. Davie, and P. L. A. Popelier, "The accuracy of *ab initio* calculations without *ab initio* calculations for charged systems: Kriging predictions of atomistic properties for ions in aqueous solutions," J. Chem. Phys. **148**(24), 241724 (2018).

[44]M. A. Wood and A. P. Thompson, "Extending the accuracy of the SNAP interatomic potential form," J. Chem. Phys. **148**(24), 241721 (2018).

[45]C. Zeni, K. Rossi, A. Glielmo, N. Gaston, F. Baletto, and A. De Vita, "Building machine learning force fields for nanoclusters," J. Chem. Phys. **148**(24), 241739 (2018).

[46]J. E. Herr, K. Yao, R. McIntyre, D. Toth, and J. Parkhill, "Metadynamics for training neural network model chemistries: A competitive assessment," J. Chem. Phys. **148**(24), 241710 (2018).

[47]Y.-J. Zhang, A. Khorshidi, G. Kastlunger, and A. A. Peterson, "The potential for machine learning in hybrid QM/MM calculations," J. Chem. Phys. **148**(24), 241740 (2018).

[48]V. Quaranta, M. Hellström, J. Behler, J. Kullgren, P. D. Mitev, and K. Hermansson, "Maximally resolved anharmonic OH vibrational spectrum of the water/ZnO($10\bar{1}0$) interface from a high-dimensional neural network potential," J. Chem. Phys. **148**(24), 241720 (2018).

[49]C. Qu and J. M. Bowman, "High-dimensional fitting of sparse datasets of CCSD(T) electronic energies and MP2 dipole moments, illustrated for the formic acid dimer and its complex IR spectrum," J. Chem. Phys. **148**(24), 241713 (2018).

[50]T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, "Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions," J. Chem. Phys. **148**(24), 241725 (2018).

[51]A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington, Jr., and S. Manzhos, "Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy," J. Chem. Phys. **148**(24), 241702 (2018).

[52]N. Artrith, A. Urban, and G. Ceder, "Constructing first-principles phase diagrams of amorphous Li_xSi using machine-learning-assisted sampling with an evolutionary algorithm," J. Chem. Phys. **148**(24), 241711 (2018).

[53]J. Schmidt, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the stability of ternary intermetallics with density functional theory and machine learning," J. Chem. Phys. **148**(24), 241728 (2018).

[54]X. Li, F. S. Curtis, T. Rose, C. Schober, A. Vazquez-Mayagoitia, K. Reuter, H. Oberhofer, and N. Marom, "Genarris: Random generation of molecular crystal structures and fast screening with a Harris approximation," J. Chem. Phys. **148**(24), 241701 (2018).

[55]S. Kiyohara and T. Mizoguchi, "Searching the segregation configuration at the grain boundary by a Monte Carlo tree search," J. Chem. Phys. **148**(24), 241741 (2018).

[56]A. Seko, H. Hayashi, and I. Tanaka, "Compositional descriptor-based recommender system for the materials discovery," J. Chem. Phys. **148**(24), 241719 (2018).
[57]T. M. Dieb, Z. Hou, and K. Tsuda, "Structure prediction of boron-doped graphene by machine learning," J. Chem. Phys. **148**(24), 241716 (2018).
[58]N. Mardirossian and M. Head-Gordon, "Survival of the most transferable at the top of Jacob's ladder: Defining and testing the ωB97M(2) double hybrid density functional," J. Chem. Phys. **148**(24), 241736 (2018).
[59]J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka, "Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits," J. Chem. Phys. **123**(6), 062201 (2005).
[60]R. Nagai, R. Akashi, S. Sasaki, and S. Tsuneyuki, "Neural-network Kohn-Sham exchange-correlation potential and its out-of-training transferability," J. Chem. Phys. **148**(24), 241737 (2018).
[61]M. Levy and F. Zahariev, "Ground-state energy as a simple sum of orbital energies in Kohn-Sham theory: A shift in perspective through a shift in potential," Phys. Rev. Lett. **113**, 113002 (2014).
[62]H. Ji and Y. Jung, "A local environment descriptor for machine-learned electronic structure theory," J. Chem. Phys. **148**, 241742 (2018).
[63]J. Hollingsworth, L. Li, T. E. Baker, and K. Burke, "Can exact conditions improve machine-learned density functionals?," J. Chem. Phys. **148**, 241743 (2018).
[64]J. Seino, R. Kageyama, M. Fujinami, Y. Ikabata, and H. Nakai, "Semi-local machine-learned kinetic energy density functional with third-order gradients of electron density," J. Chem. Phys. **148**(24), 241705 (2018).
[65]L. Boninsegna, F. Nüske, and C. Clementi, "Sparse learning of stochastic dynamical equations," J. Chem. Phys. **148**(24), 241723 (2018).
[66]C. Wehmeyer and F. Noé, "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics," J. Chem. Phys. **148**(24), 241703 (2018).
[67]Y. Matsunaga and Y. Sugita, "Refining Markov state models for conformational dynamics using ensemble-averaged data and time-series trajectories," J. Chem. Phys. **148**(24), 241731 (2018).
[68]P. Pernot and A. Savin, "Probabilistic performance estimators for computational chemistry methods: The empirical cumulative distribution function of absolute errors," J. Chem. Phys. **148**(24), 241707 (2018).
[69]P. B. Jørgensen, M. Mesta, S. Shil, J. M. García Lastra, K. W. Jacobsen, K. S. Thygesen, and M. Schmidt, "Machine learning-based screening of complex molecules for polymer solar cells," J. Chem. Phys. **148**(24), 241735 (2018).
[70]M. A. F. Afzal, C. Cheng, and J. Hachmann, "Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers," J. Chem. Phys. **148**(24), 241712 (2018).
[71]G. Pilania, K. J. McClellan, C. R. Stanek, and B. P. Uberuaga, "Physics-informed machine learning for inorganic scintillator discovery," J. Chem. Phys. **148**(24), 241729 (2018).
[72]G. Schmitz and O. Christiansen, "Gaussian process regression to accelerate geometry optimizations relying on numerical differentiation," J. Chem. Phys. **148**(24), 241704 (2018).
[73]K. H. Sørensen, M. S. Jørgensen, A. Bruix, and B. Hammer, "Accelerating atomic structure search with cluster regularization," J. Chem. Phys. **148**(24), 241734 (2018).
[74]M. Botlani, A. Siddiqui, and S. Varmaa, "Machine learning approaches to evaluate correlation patterns in allosteric signaling: A case study of the PDZ2 domain," J. Chem. Phys. **148**(24), 241726 (2018).
[75]F. Cipcigan, A. P. Carrieri, E. O. Pyzer-Knapp, R. Krishna, Y.-W. Hsiao, M. Winn, M. G. Ryadnov, C. Edge, G. Martyna, and J. Crain, "Accelerating molecular discovery through data and physical sciences: Applications to peptide-membrane interactions," J. Chem. Phys. **148**, 241744 (2018).
[76]T. S. Hy, S. Trivedi, H. Pan, B. M. Anderson, and R. Kondor, "Predicting molecular properties with covariant compositional networks," J. Chem. Phys. **148**, 241745 (2018).