Electronic Structure

OPEN ACCESS

ROADMAP

Roadmap on Machine learning in electronic structure

RECEIVED
30 September 2021

REVISED
22 December 2021

ACCEPTED FOR PUBLICATION
21 February 2022

PUBLISHED
19 August 2022

H J Kulik[1,*], T Hammerschmidt[2,*], J Schmidt[3,*], S Botti[4,*], M A L Marques[3,*], M Boley[5], M Scheffler[6,*], M Todorović[7,8], P Rinke[8,*], C Oses[9], A Smolyanyuk[9], S Curtarolo[9,*], A Tkatchenko[10], A P Bartók[11,*], S Manzhos[12,*], M Ihara[12], T Carrington[13], J Behler[14], O Isayev[15], M Veit[16], A Grisafi[16,17], J Nigam[16], M Ceriotti[16,*], K T Schütt[18,19,*], J Westermayr[20], M Gastegger[18], R J Maurer[20], B Kalita[21], K Burke[21,22,*], R Nagai[23,24], R Akashi[24], O Sugino[23,24], J Hermann[25,35], F Noé[25,26,27,35], S Pilati[28,29,*], C Draxl[30], M Kuban[30], S Rigamonti[30], M Scheidgen[30], M Esters[9], D Hicks[9], C Toher[9,36,*], P V Balachandran[31,37], I Tamblyn[32], S Whitelam[33], C Bellinger[34] and L M Ghiringhelli[6,*]

[1] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America
[2] ICAMS, Ruhr-Universität Bochum, Germany
[3] Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, 06120 Halle (Saale), Germany
[4] Institut für Festkörpertheorie und -optik, Friedrich-Schiller-Universität Jena, Max-Wien-Platz 1, 07743 Jena, Germany
[5] Department of Data Science and AI, Monash University, Melbourne, Australia
[6] The NOMAD Laboratory at the FHI of the Max-Planck-Gesellschaft and IRIS Adlershof of the Humboldt Universität, Berlin, Germany
[7] Department of Mechanical and Materials Engineering, University of Turku, FI-20014, Turku, Finland
[8] Department of Applied Physics, Aalto University, FI-00076, Aalto, Finland
[9] Center for Autonomous Materials Design, Duke University, Durham NC 27708, United States of America
[10] Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg
[11] Department of Physics and Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry, CV4 7AL, United Kingdom
[12] Tokyo Institute of Technology, Japan
[13] Queen's University, Canada
[14] Institut für Physikalische Chemie, Universität Göttingen, Germany
[15] Department of Chemistry, Carnegie Mellon University, United States of America
[16] Laboratory of Computational Science and Modeling, Institute of Materials, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[17] PASTEUR, Département de Chimie, Ecole Normale Supérieure, 75005 Paris, France
[18] Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany
[19] Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
[20] Department of Chemistry, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom
[21] Department of Chemistry, University of California, Irvine, United States of America
[22] Department of Physics and Astronomy, University of California, Irvine, United States of America
[23] Institute for Solid State Physics, The University of Tokyo, Japan
[24] Department of Physics, The University of Tokyo, Japan
[25] Department of Mathematics and Computer Science, FU Berlin, Arnimallee 6, 14195 Berlin, Germany
[26] Department of Physics, FU Berlin, Arnimallee 14, 14195 Berlin, Germany
[27] Department of Chemistry, Rice University, 6100 Main St., Houston, TX 77005-1827, United States of America
[28] School of Science and Technology, Physics Division, University of Camerino, I-62032 Camerino (MC), Italy
[29] INFN-Sezione di Perugia, 06123 Perugia, Italy
[30] Humboldt-Universität zu Berlin (HU Berlin), Germany
[31] Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA 22904, United States of America
[32] Department of Physics, University of Ottawa, Canada
[33] Lawrence Berkeley National Laboratory, United States of America
[34] National Research Council, Digital Technologies, Ottawa, Canada
[35] Microsoft Research, Cambridge, United Kingdom
[36] Department of Materials Science and Engineering and Department of Chemistry and Biochemistry, University of Texas at Dallas, Richardson TX 75080, United States of America
[37] Department of Mechanical and Aerospace Engineering, University of Virginia, Charlottesville, United States of America VA 22904
* Authors to whom any correspondence should be addressed.

E-mail: hjkulik@mit.edu, thomas.hammerschmidt@rub.de, jonathan.schmidt@student.uni-halle.de, silvana.botti@uni-jena.de, miguel.marques@physik.uni-halle.de, scheffler@fhi-berlin.mpg.de, patrick.rinke@aalto.fi, stefano@duke.edu, alexandre.tkatchenko@uni.lu, Albert.Bartok-Partay@warwick.ac.uk, manzhos.s.aa@m.titech.ac.jp, joerg.behler@uni-goettingen.de, olexandr@olexandrisayev.com, michele.ceriotti@epfl.ch, kristof.schuett@tu-berlin.de, kieron@uci.edu, sugino@issp.u-tokyo.ac.jp, frank.noe@fu-berlin.de, sebastiano.pilati@unicam.it, claudia.draxl@physik.hu-berlin.de, ghiringhelli@fhi-berlin.mpg.de, cormac.toher@utdallas.edu, pvb5e@virginia.edu and isaac.tamblyn@uottawa.ca

Keywords: machine learning, electronic structure, computational materials science, density-functional theory

Abstract

In recent years, we have been witnessing a paradigm shift in computational materials science. In fact, traditional methods, mostly developed in the second half of the XXth century, are being complemented, extended, and sometimes even completely replaced by faster, simpler, and often more accurate approaches. The new approaches, that we collectively label by machine learning, have their origins in the fields of informatics and artificial intelligence, but are making rapid inroads in all other branches of science. With this in mind, this Roadmap article, consisting of multiple contributions from experts across the field, discusses the use of machine learning in materials science, and share perspectives on current and future challenges in problems as diverse as the prediction of materials properties, the construction of force-fields, the development of exchange correlation functionals for density-functional theory, the solution of the many-body problem, and more. In spite of the already numerous and exciting success stories, we are just at the beginning of a long path that will reshape materials science for the many challenges of the XXIth century.

Contents

1. Predicting material properties 4
1.1. Using machine learning to accelerate computational materials design 4
1.1.1. Status 4
1.1.2. Current and future challenges 5
1.1.3. Advances in science and technology to meet challenges 6
1.1.4. Concluding remarks 6
Acknowledgments 7
1.2. Machine learning for material-property prediction 7
1.2.1. Status 7
1.2.2. Current and future challenges 7
1.2.3. Advances in science and technology to meet challenges 8
1.2.4. Concluding remarks 9
Acknowledgments 9
1.3. Predicting thermodynamically stable material 9
1.3.1. Status 9
1.3.2. Current and future challenges 11
1.3.3. Advances in science and technology to meet challenges 11
1.3.4. Concluding remark 11
1.4. Learning rules for materials properties and functions 11
1.4.1. Status 11
1.4.2. Current and future challenges 12
1.4.3. Advances in science and technology to meet challenges 13
1.4.4. Concluding remarks 14
Acknowledgments 14
1.5. Deep learning for spectroscopy 14
1.5.1. Status 14
1.5.2. Current and future challenges 15
1.5.3. Advances in science and technology to meet challenges 15
1.5.4. Concluding remarks 17
Acknowledgments 17
1.6. Machine learning for disordered systems 17
1.6.1. Status 17
1.6.2. Current and future challenges 17
1.6.3. Advances in science and technology to meet challenges 18
1.6.4. Concluding remarks 19

Acknowledgments 20
2. Construction of accurate force fields and beyond 20
2.1. Machine learning for molecular quantum simulations 20
2.1.1. Status 20
2.1.2. Current and future challenges 20
2.1.3. Advances in science and technology to meet challenges 21
2.1.4. Concluding remarks 21
Acknowledgments 21
2.2. Bayesian machine learning for microscopic interactions 21
2.2.1. Status 21
2.2.2. Current and future challenges 22
2.2.3. Advances in science and technology to meet challenges 23
2.2.4. Concluding remarks 23
2.3. Spectroscopically accurate potential energy surfaces (SAPES) from machine learning 23
2.3.1. Status 23
2.3.2. Current and future challenges 24
2.3.3. Advances in science and technology to meet challenges 24
2.3.4. Concluding remarks 25
Acknowledgments 25
2.4. High-dimensional neural network potential energy surfaces in chemistry and materials science 25
2.4.1. Status 25
2.4.2. Current and future challenges 25
2.4.3. Advances in science and technology to meet challenges 26
2.4.4. Concluding remarks 27
Acknowledgments 27
2.5. Transferable neural network force fields 27
2.5.1. Status 27
2.5.2. Current and future challenges 27
2.5.3. Advances in science and technology to meet challenges 29
2.5.4. Concluding remarks 29
Acknowledgments 29
2.6. Integrated machine learning models: electronic structure accuracy beyond local potentials 29
2.6.1. Status 29
2.6.2. Current and future challenges 30
2.6.3. Advances in science and technology to meet challenges 30
2.6.4. Concluding remarks 31
Acknowledgments 32
3. Solving the many-body problem with machine learning 32
3.1. Unifying machine learning and electronic structure methods 32
3.1.1. Status 32
3.1.2. Current and future challenges 33
3.1.3. Advances in science and technology to meet challenges 33
3.1.4. Concluding remarks 34
Acknowledgments 34
3.2. Using machine learning to find new density functionals 34
3.2.1. Status 34
3.2.2. Current and future challenges 35
3.2.3. Advances in science and technology to meet challenges 35
3.2.4. Concluding remarks 36
Acknowledgments 36
3.3. Machine learning Kohn–Sham exchange–correlation potentials 36
3.3.1. Status 37
3.3.2. Current and future challenges 37
3.3.3. Advances in science and technology to meet challenges 38

3.3.4. Concluding remarks ... 38
Acknowledgments ... 38
3.4. Deep-learning quantum Monte Carlo for molecules ... 38
3.4.1. Status ... 38
3.4.2. Current and future challenges ... 39
3.4.3. Advances in science and technology to meet challenges ... 40
3.4.4. Concluding remarks ... 40
Acknowledgments ... 41
3.5. Disordered quantum systems ... 41
3.5.1. Status ... 41
3.5.2. Current and future challenges ... 41
3.5.3. Advances in science and technology to meet challenges ... 41
3.5.4. Concluding remarks ... 42
Acknowledgments ... 43
4. Big data for machine learning ... 43
4.1. Challenges and perspectives for interoperability and reuse of heterogenous data collections ... 43
4.1.1. Status ... 43
4.1.2. Current and future challenges ... 43
4.1.3. Advances in science and technology to meet challenges ... 44
4.1.4. Concluding remarks ... 44
Acknowledgments ... 45
4.2. The AFLOW framework for computational materials data and design ... 45
4.2.1. Status ... 45
4.2.2. Current and future challenges ... 46
4.2.3. Advances in science and technology to meet challenges ... 47
4.2.4. Concluding remarks ... 48
Acknowledgments ... 48
5. Frontier developments of machine learning in materials science ... 48
5.1. Adaptive learning strategies for electronic structure calculations ... 48
5.1.1. Status ... 48
5.1.2. Current and future challenges ... 49
5.1.3. Advances in science and technology to meet challenges ... 49
5.1.4. Concluding remarks ... 50
Acknowledgments ... 50
5.2. Reinforcement learning ... 50
5.2.1. Status ... 50
5.2.2. Current and future challenges ... 51
5.2.3. Advances in science and technology to meet challenges ... 51
5.2.4. Concluding remarks ... 52
5.3. Interpretability of machine learning models in physical sciences ... 52
5.3.1. Status ... 52
5.3.2. Current and future challenges ... 52
5.3.3. Advances in science and technology to meet challenges ... 53
5.3.4. Concluding remarks ... 53
Acknowledgments ... 54
Data availability statement ... 54
References ... 54

1. Predicting material properties

1.1. Using machine learning to accelerate computational materials design

Heather J Kulik
Massachusetts Institute of Technology

1.1.1. Status

Computational materials discovery efforts with density functional theory (DFT) and machine learning (ML) have matured in the past decade. Here, I focus on open-shell transition-metal complex discovery, which has

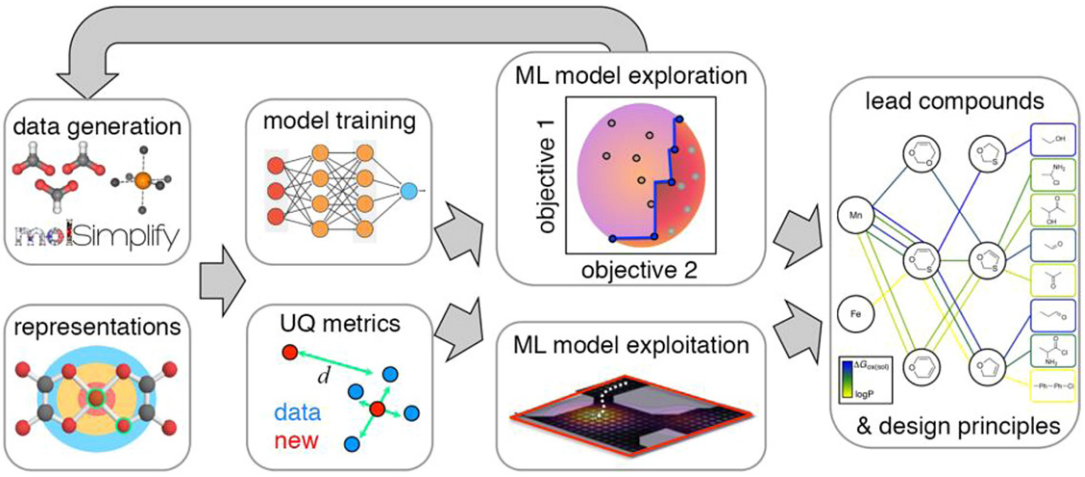

Figure 1. Approach showing data generation and ANN ML model training for exploration and exploitation in computational materials discovery.

unique challenges, owing to the vastness of compound space spanned by their ligand chemistry, isomers, coordination number, spin/oxidation state, and charge [1]. Specialized workflows for DFT high-throughput screening of transition-metal chemistry (e.g., molSimplify, https://molsimplify.mit.edu) are key for data generation. For this class of materials, semi-empirical and force field methods are not predictive. Tailored representations [2] are also essential for predictive ML artificial neural network (ANN), models on modest data sets (ca. 300–1000 complexes). Sparse, graph-based, metal-focused representations encode the metal-dominance of transition-metal complex properties for properties such as spin splitting, redox potential, and catalyst energetics [2]. For closed-shell complexes, standard whole-molecule descriptors used in organic chemistry may also be suitable [3]. Once trained, both ML model exploitation [4] and exploration [5, 6] can accelerate chemical discovery efforts (figure 1). Trained models can be *exploited* to enumerate or optimize properties with a genetic algorithm (GA) in a discrete chemical space by leveraging uncertainty quantification (UQ) metrics such as ensemble variance or distances in the model's latent/feature space [7]. The GA fitness function can be the combined property score with a penalty for high-uncertainty, distant points. This approach ensures the prediction errors on discovered complexes are close to test set errors by only making predictions where the models is confident, and lead compounds can be validated with DFT. Alternatively, in active learning (AL) with ML model *exploration* (see also section 5.1), we acquire points that are both promising and uncertain for model retraining, for example, with expected improvement in efficient global optimization [5]. This approach is useful when multi-objective optimization requires a large (ca. millions of complexes) search space and the best leads are unknown. These methods enhance ML model accuracy at an improving Pareto front at each generation. Because DFT calculations are carried out at each step, the improvement of the model can be assessed as can its optimism about the compound space. With this approach, design rules and leads are discovered in weeks instead of decades that a parallelized, random search with DFT would require [5].

1.1.2. Current and future challenges

Despite the promise and rapid progress in this field, compelling materials spaces with correlated electronic structure in open-shell transition metal complexes introduce additional concerns:

(a) *Electronic structure method accuracy*. The hierarchy of systematically improvable accuracy established for small-molecule organic chemistry fails for transition-metal chemistry. Although DFT is widely applied, imbalances in delocalization and static correlation error make the choice of exchange–correlation (XC) functional [1] system-dependent (figure 2). Single-reference correlated quantum chemistry methods can fail due to the multi-determinantal nature of open-shell systems [8]. Detection of multi-reference character in screening (e.g., with MultirefPredict, https://github.com/hjkgrp/MultirefPredict) is necessary. However, multi-reference methods require careful parameter selection, including active space and orbitals. Robust data for benchmarking may be unavailable due both to uncertainty in experimental measurements and differences between the computational and experimental setup.

(b) *Efficient and robust data acquisition*. Calculations may fail to converge, to optimize to a stationary point, or to produce a robust result, leading to wasted computational effort [9]. Automated workflows should avoid generating erroneous data that hinder the learning task for ML models. Given the nuanced computational cost and performance considerations in transition-metal chemistry, automated tools must be

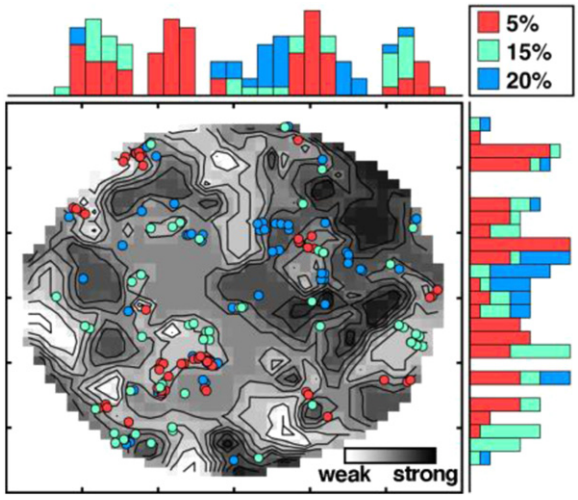

Figure 2. Dimensionality reduction of candidate mononuclear octahedral spin crossover (SCO) complexes (i.e., with near-degenerate high- and low-spin states) with mid-row (i.e., Cr–Co) 3*d* transition metals. The leads (discrete circles) change with functional choice (% exact exchange, colored as in inset). Reprinted with permission from [1]. Copyright (2019) American Chemical Society.

imparted with the expert knowledge to fine-tune electronic structure method parameters, choices, and cost vs accuracy trade-offs.

(c) *Anthropogenic bias in dataset construction*. The set of compounds that have been synthesized, characterized, and reported in the experimental literature carries significant human biases and omit failed outcomes. Given the array of choices for enumerating inorganic materials, hypothetical sets can introduce their own biases and are sensitive to the rules or building blocks used for enumeration. ML models that learn design principles from these sets, whether generative in nature or through exploration of a discrete set of compounds, will be influenced by these biases.

(d) *Multi-faceted criteria in materials design*. Computational materials design frequently focuses on optimization of one energy-based criterion, such as a band gap or descriptor of catalytic activity. In practice, a large number of other criteria such as cost, stability, synthesizability, solubility, and toxicity are equally important but have received less attention. This is both because these quantities are harder to predict and because it may be challenging to identify *a priori* the biggest impediment to experimental realization of a computationally designed material.

1.1.3. Advances in science and technology to meet challenges

ML-accelerated computational discovery is expected to benefit most from synergistic integration of advancements in artificial intelligence (AI) with related areas of computational chemistry. Although electronic structure method accuracy is system-specific, statistical models of optimal parameter choice, including DFT functionals, multi-reference character [8], or active spaces in multi-reference theories [10] will enable improvement. ML models that encode more flexible parameters in current theories or supersede analytical forms have the potential to advance accuracy beyond current methods. The integration of UQ [7] and sensitivity analysis into both the electronic structure and ML predictions will bring robustness to computational discovery. Semi-supervised models that leverage a combination of labelled and unlabelled data will address the challenge posted by divergent property landscapes with varying theory choice [1, 8] (figure 2). Natural language and image processing extraction of large experimental data sets will provide both larger benchmark sets and knowledge of the repeatability/uncertainty from independent experimental property measurements (see also section 4.1). Continued increases in algorithmic and hardware efficiency will also increase how much data can be generated and improve the fidelity of this data. Nevertheless, models that can recapitulate expert decisions will guide how to best use computational resources, including prediction of when calculations will succeed or fail. Increased development and use of human-guided ML will be important. While bias is challenging to overcome, critical assessments of the relationship between synthesized and hypothetical materials, including through improved representations that adequately encode similarity and data distribution characteristics will enable biases to be recognized and acknowledged. Tight integration between experiment and computation driven by autonomous tools and improved generative models will enable ML-accelerated discovery to address the multitude of unknowns associated with the design of practical materials.

1.1.4. Concluding remarks

Initial efforts of integrating ML into computational workflows suggest substantial promise for augmenting and accelerating the traditional trial-and-error approach to computational materials discovery. Certain regions of chemical space, such as those that contain the correlated open-shell transition-metal centers in coordination complexes and metal–organic frameworks, are both the most promising for functional materials design and the most fraught with outstanding challenges. Tighter integration between computational scientists and experimental efforts as well as incorporation of advanced AI into software workflows are expected to enable extension of the current efforts to tackle realistic, multi-faceted design challenges. In doing so, it is anticipated that it will become increasingly feasible to carry out autonomous discovery of new functional materials in days to weeks instead of years or decades.

Acknowledgments

HJK acknowledges generous support by the Office of Naval Research under Grant Numbers N00014-17-1-2956, N00014-18-1-2434, and N00014-20-1-2150, DARPA Grant D18AP00039, the Department of Energy under Grant Numbers DE-SC0018096 and DE-SC0012702, the National Science Foundation under Grant Numbers CBET-1704266 and CBET-1846426, an AAAS Marion Milligan Mason Award, and a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. The author also thanks Adam H Steeves for providing a critical reading.

1.2. Machine learning for material-property prediction

Thomas Hammerschmidt

Ruhr-Universität Bochum, Germany

1.2.1. Status

Many aspects of our society rely on highly optimized materials. The demands on the materials range from structural stability in extreme conditions to high functional performance with long life time. Common goal of the materials design is to optimize the performance by tuning material chemistry and processing. The immediate questions are: (i) which crystal structure and microstructure form for a given chemical composition under which conditions? (ii) Which performance can be expected? (iii) how stable is the performance during service? (iv) How can processing optimize performance and life time? Traditional empirical optimization leaves vast regions of chemical space and processing unexplored. The promise of ML for material-property prediction is to optimise known materials and to design new materials for specific target properties.

The most fundamental material property is the equilibrium crystal structure that a combination of chemical elements will form. Early attempts to predict compound formation used one-dimensional descriptors that arrange the elements of the periodic table along a string [11]. Compounds can then be represented as structure maps that cast the data set of experimentally observed compounds in low-dimensional representations, an approach later confirmed by mining DFT data [12]. More fine-grained information of the chemical elements like electronic configuration, covalent radius, or electro-negativity, can be utilized for classification of structural stability with more complex structure maps [13] or in feature vectors for regression learning of DFT data. A further refinement is to utilize also information of the interatomic interaction, either in terms of the mere geometric arrangement of the atoms or e.g. in terms of electronic-structure based descriptors [14] based on bond-order potentials. Current ML of material properties typically combines atom/bond information with chemistry information, see figure 3 as example [15]. Such atomic-scale descriptors can be upscaled to some degree, e.g., to complex entities like grain boundaries [16]. However, for predicting macroscopic material properties like plastic deformation, many length and time scales need to be bridged. ML at these scales is based on information from, e.g., micromechanical simulations [17].

1.2.2. Current and future challenges

The prediction of material properties by ML has enormous potential to deliver a wealth of materials for the benefit of our society, particularly for the energy, environment and communication sectors. The possibilities range from optimized known materials over unexpected new material classes to inversely-designed materials. Still, the challenges are manifold:

(a) At the atomic level we have witnessed enormous progress in the last years regarding the application of data-science techniques to results of quantum-mechanical (QM) calculations. Most ML models, however, use descriptors that are agnostic of the underlying physics, i.e. of the interatomic bond between the chemical elements. This is instead picked up implicitly by extensive sampling of the potential energy surface (PES). The consequence is a degree of data hunger that poses a considerable challenge for the exploration of high-dimensional chemical and structural spaces.

(b) ML macroscopic material properties requires different concepts. Taking structural materials as example, the relevant information units are the microstructure elements and their formation and evolution during fabrication, processing and operation. This involves the combinatorial diversity of chemical compositions

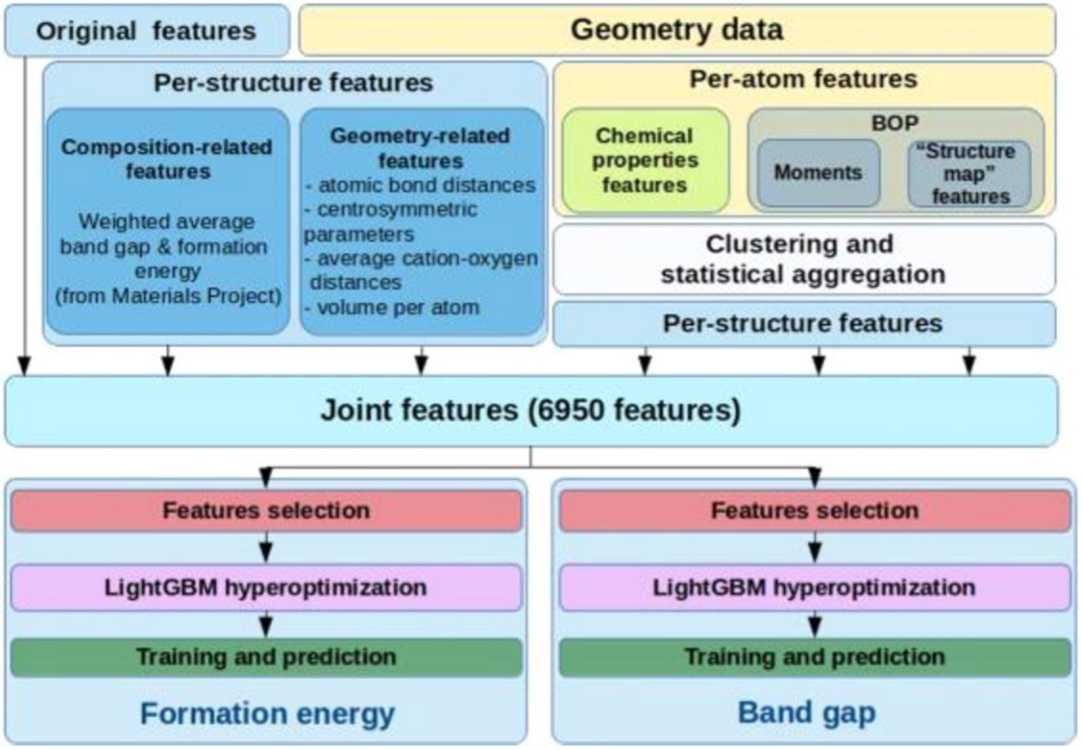

Figure 3. Example of ML architecture for predicting structural and functional material properties of transparent conductors, reproduced from [15], CC BY 4.0. The feature engineering constructs a feature vector with information on chemistry and atomic structure at a per-atom and at a per-structure level. The feature vector is contracted in the feature selection step and separate models are trained for the formation energy and the bandgap.

and the geometrical complexity of the microstructure. From an electronic-structure perspective, a main challenge is the extrapolation from calculations for comparably small simulation cells to microstructure elements (e.g., dislocations, interfaces, precipitates) with sufficient chemical and geometrical complexity. At the macroscopic level the challenges are the representation of microstructure elements and the identification of suitable descriptors as well as the generation of meaningful artificial data for learning the time evolution.

(c) A related challenge is the gap between the data from electronic-structure calculations and the data from experiments on macroscopic material properties. Taking mechanical deformation as example, it is known that the modification of stacking-fault energies by alloying elements affects the macroscopic plastic deformation, see e.g. reference [18]. However, it is not clear at the moment how to join data of electronic-structure calculations of stacking-fault energies with experimental stress–strain data in order to machine-learn the influence of chemical composition on the mechanical strength of a microstructured material.

1.2.3. Advances in science and technology to meet challenges

Meeting the above challenges for predicting material properties by ML requires advances in the combination of correlative data science with causative physical models. This requires methodological developments of physical models and of their connection to data-science techniques.

Major impact on the efficiency, robustness and interpretability of ML models can be expected from advances in the construction of chemistry/physics-aware descriptors, in the biasing of sampling techniques towards the chemically/physically relevant phase-space, in the implementation of ML constraints/guidance from domain knowledge (available as physical parameters or as physical models), and in the development of ML models with robust transferability.

This may include, e.g., extensions of the periodic table of the elements towards dictionaries of local building blocks of chemistry as developed in molecular sciences [19]. For predicting the properties of bulk materials, such schemes would need to be advanced to handle long-ranged effects of crystalline and microstructured systems. Another potentially viable route to utilize domain knowledge of the interatomic interaction would be to equip electronic-structure based descriptors of the local atomic environment with parameters that are specific for the chemical bond, e.g., pairwise Hamiltonians from down-folding DFT eigenstates to a tight-binding minimal basis [20] (figure 4).

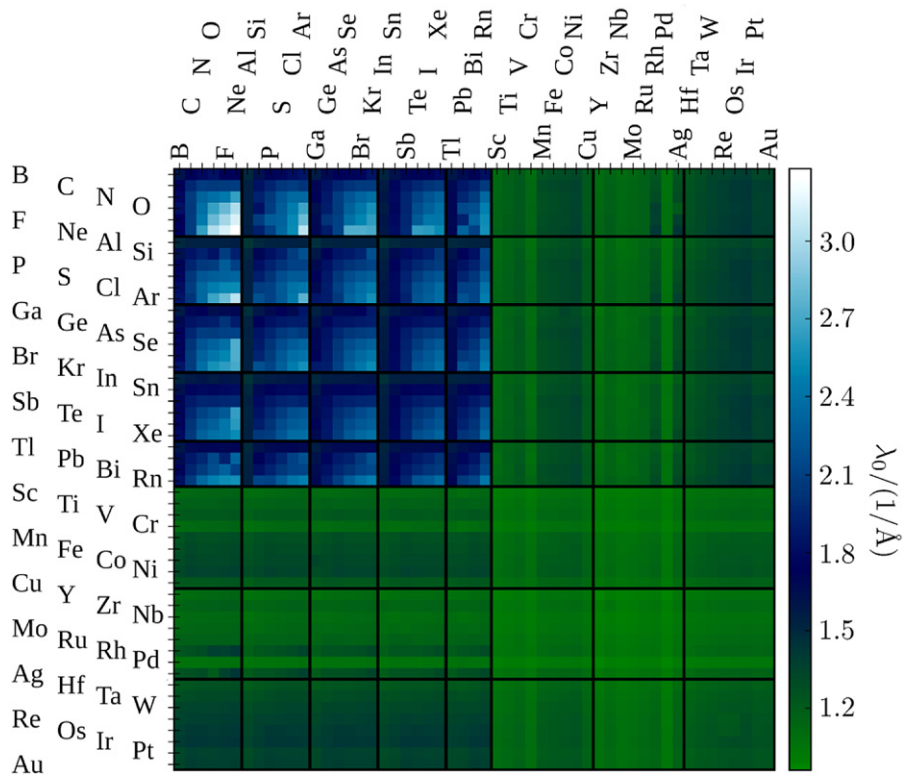

Figure 4. Parameters of a tight-binding bond-model obtained by downfolding DFT eigenstates of two-atomic molecules to a minimal basis across the periodic table, reprinted (figure) with permission from [20], Copyright (2021) by the American Physical Society. The combination with electronic-structure based descriptors of the local atomic environment [14], effectively encodes domain-knowledge of the interatomic interaction in the descriptor that can be exploited for ML material properties across chemical space.

A prerequisite for combining the results of electronic-structure calculations with experimental data on macroscopic material properties is the ML-ready preparation of the data sets in a common logical framework. This will allow further advances in identifying correlations in the joint data at different time and length scales.

1.2.4. Concluding remarks

The application of ML for the prediction of materials properties has just started but already now the tremendous success is revolutionizing the design of materials. High precision and vast chemical screenings have been demonstrated for properties that are directly linked to time- and length scales of electronic-structure calculations. Some of the central challenges at the moment are the combination of physical insight and data-science techniques, the up-scaling of the electronic-structure calculations to macroscopic material properties and the joint learning with experimental data. Taking these steps will enable us to switch gear from computing the property of a material to computing the material for a property.

Acknowledgments

Financial support by the Deutsche Forschungsgemeinschaft (DFG) through Project C1 of the collaborative research centre SFB/TR 103 'From Atoms to Turbine Blades—A Scientific Basis for a new Generation of Single-Crystal Superalloys' is acknowledged.

1.3. Predicting thermodynamically stable material

Jonathan Schmidt[1], Silvana Botti[2] and Miguel A L Marques[1,*]

[1]Martin-Luther-Universität Halle-Wittenberg, Germany

[2]Friedrich-Schiller-Universität Jena, Germany

E-mail: miguel.marques@physik.uni-halle.de

1.3.1. Status

The discovery of new materials is one of the key drivers for progress in material science. We are obviously interested in predicting composition and crystal structure of stable materials that can be synthesized in a lab. A particularly important concept in materials design is the distance of a candidate system to the convex hull of thermodynamic stability E_{Hull} (see figure 5), as this quantity measures if the compound would rather decompose to reaction products with lower energy. Finding crystal phases on the convex hull (or close to it) is unfortunately an extremely difficult task. In fact, just enumerating all possible combinations of chemical

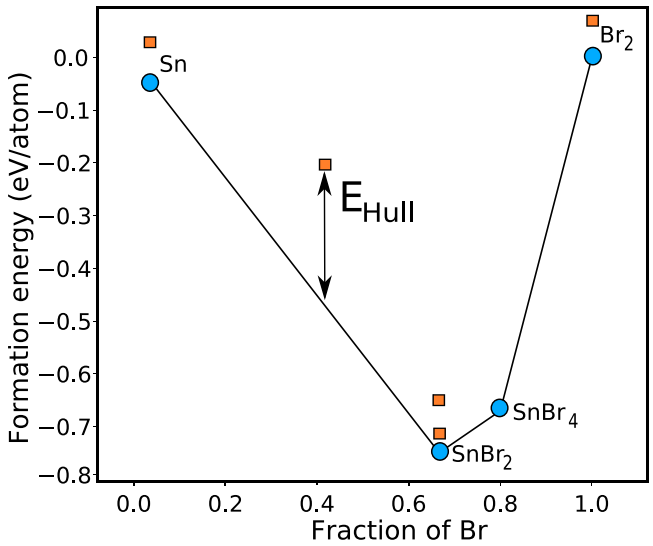

Figure 5. Convex hull (line connecting the blue circles) for the binary system Sn_xBr_{1-x}. Unstable systems are marked with orange squares.

elements in all different crystal structures is a problem with factorial complexity. Moreover, most of the resulting compounds are highly unstable, and therefore of little or no interest. There are two opposite approaches to tackle the problem of thermodynamic stability. (i) one can search for the most stable crystal structure(s) at a given chemical composition [21] or (ii) one can search for the chemical elements that stabilize a certain structure prototype. This latter method is applied in high-throughput studies based on DFT. Although successful, these approaches are highly data-intensive. In this context, the emergence of ML has shown tremendous potential to speed up materials discovery. Early works [22–25] were based on traditional ML algorithms, such as kernel ridge regression (KRR) or randomized tree ensembles. These works were already quite successful, allowing for a speed-up factor of five [23] or more.

In most cases the target property predicted was the formation energy [22, 24, 26, 27] which defines stability with respect to decomposition in elementary substances. In principle, the distance to the convex hull can be calculated from the formation energies but it can also be predicted directly. Contradictory results in literature [28–30] do not allow to conclude yet whether the former or the latter choice leads to better predictions.

Some authors [22–24] used only compositional information, such as e.g., the electronegativities, atomic numbers, positions of the constituting elements in the periodic table, etc, to define the input system. Crystal structure information based on the Voronoi tessellations, invariant with respect to the volume of the unit cell, was incorporated in reference [25]. Purely composition-based models are limited to predict only one structure, as they cannot differentiate between different crystal prototypes with the same composition. As a consequence, only one prototype can be included in the training data, leading to a restriction of available data by orders of magnitude. Ward *et al* demonstrated [25] that additional training data of other prototypes improve prediction accuracy in calculating both formation energy and distance to the convex hull. It is important to note, however, that in high-throughput studies the relaxed geometry of the compounds is in general not available, so this information should not be used as input feature of the machines.

For the last years we have witnessed an evolution from simple ML algorithms to a variety of deep neural networks (NNs) [26, 27, 30–32] with far superior performance when sufficient data is available. The performance gap between handcrafted features and representations learned by deep NNs is not surprising, as this can be expected from experience gathered in other fields of science [33]. This second generation of prediction models can again be split into composition-based and structure-based models. Different works benefit from different representations of the composition: reference [31] uses a simple vector, while reference [27] represents the composition as a fully connected graph of the elements. According to an in-depth comparison of the various implementations [28], the latter is the most accurate model purely based on composition for the prediction of formation energies.

Structure-based models stem from the crystal-graph convolutional NNs of reference [26]. These message passing networks [32] rely on representing crystals as a graph where each atom forms a node and the edges contain a representation of the bonds. Periodicity results in loops in the graph. After the publication of the original idea [26] a large number of works [30, 32] have applied message passing networks with different update functions to predict several material properties. Structure based deep learning models not only profit from the additional training data but also from the complete knowledge of the crystal structure. Therefore while the

accuracy of these networks and their prediction ability is superior [28], the reported errors are not directly pertinent to high-throughput searches. For example, Park and Wolverton [30] reports roughly eight times larger errors when using non-relaxed structures as input. Nevertheless, the speed-up achieved during high-throughput searches based on the improved crystal-graph convolutional NNs of reference [30] is excellent and around one order of magnitude better than in earlier composition-based works.

One recent development is the inclusion of uncertainty in the prediction [27]. Uncertainty estimates allow for AL and for an informed decision on the candidate materials predicted by the models. Goodall and Lee [27] also reports an improvement in the prediction error when using a loss function that includes information on the aleatoric variance. The epistemic variance is usually approximated through a Monte-Carlo estimate, either by using an ensemble of models or by including active dropout layers. There is one large group of models that is based on experimental instead of theoretical data. Unfortunately, the number of experimentally realized systems is very limited, and is dwarfed by the available theoretical data.

1.3.2. Current and future challenges

All theoretical solid-state databases, and consequently all discussed ML models, rely on the Perdew–Burke–Ernzerhof functional [34]. Since the development of this density-functional, 25 years of research have brought progress that has been largely unused. Upgrading to a new functional such as SCAN [35], that provides better formation energies, requires recalculating all convex hulls, a computationally expensive but perfectly feasible task with today's supercomputers.

So far, we discussed various algorithmic approaches to predict the thermodynamic stability. Of course, the amount and quality of data also plays a major role in ML. Early works [22, 23] usually used custom training sets calculated for each high-throughput study. Nowadays, most models are trained either on the materials project database [36] or the open quantum materials database [37]. Unfortunately, these databases are not fully compatible with one another due to different calculation parameters. The former is relatively small and includes mostly stable or almost stable materials. The latter is larger and contains a more varied distribution of materials. The largest database is by far the automatic FLOW (AFLOW) database [38] with 3.3 million compounds. There are furthermore few publicly available datasets, each containing several hundred thousands entries [23, 39]. Combining all these data should allow for the use of datasets that will be one order of magnitude larger than current ones, enabling consequently more reliable predictions of thermodynamics stability. This is a timely, but unfortunately far from trivial task. Concerning the algorithmic development, the obvious challenge is to harness the accuracy of message-passing models to predict stable materials by developing versions that do not rely on relaxed geometries.

1.3.3. Advances in science and technology to meet challenges

The developments in ML algorithms to discover new stable materials follow a path similar to the one of nearly all other fields of AI during the last 70 years. In fact, models that leverage computation more efficiently and are able to learn from large amounts of data have replaced, and by now dominate, algorithms based on human intuition and understanding. This is true in the fields of image recognition, natural language processing, strategy games (Go, chess, StarCraft II), etc. Sutton called this the 'bitter lesson' [40]. Consequently, we should focus on developing methods and datasets that let us use effectively the enormous potential that modern computing offers. On the one hand, this requires maximizing the amount of data. Unfortunately, at the moment, the majority of calculations performed in high-throughput studies is thrown away. For example, DFT results for unrelaxed structures can and should also be kept for training ML models. This would require developing structure sensitive models that can take advantage of the extra data and circumvent calculations of relaxed crystal structures.

1.3.4. Concluding remark

Finally, and maybe most importantly, we have to realize that all developed models are just tools that *should* be combined and further applied to explore the chemical space of possible compounds, instead of being left unused after being developed.

1.4. Learning rules for materials properties and functions

Mario Boley[1] and Matthias Scheffler[2]

[1]Monash University, Australia

[2]The NOMAD Laboratory at the FHI of the Max-Planck-Gesellschaft and IRIS Adlershof of the Humboldt Universität, Berlin, Germany

1.4.1. Status

In materials science and engineering, one is typically searching for materials that exhibit exceptional performance for a certain function, and the number of these materials is extremely small. Thus, statistically speaking, we are interested in the identification of 'rare phenomena', and the scientific discovery typically resembles the

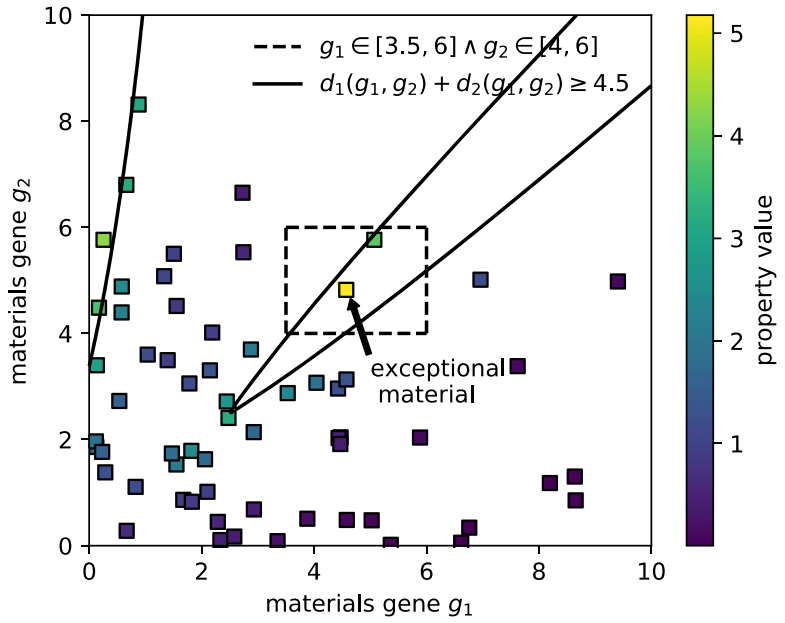

Figure 6. By mapping materials (depicted as squares) into spaces defined by relevant 'materials genes' we can identify regions where materials exhibit desired exceptional properties. Depending on the employed AI methods, such regions can be given by simple Boolean conditions (dashed line) or in terms of more complex analytic 'descriptor' functions (here, d_1, d_2) of the genes (solid line). Typically, the space of relevant genes will have a much higher dimensionality than two.

proverbial hunt for the needle in a haystack. Let us illustrate this with a 'classical' example, i.e. searching for materials that are very robust, highly transparent, and at the same time have a high heat conductivity. In the immense space of structural and chemical materials, there is one strong high-performance candidate: carbon in the diamond structure. Hardly any other material comes close. And from a thermodynamic perspective, this material is not even stable but metastable. As we understand the mechanisms behind the mentioned properties, we trust the conclusion that diamond is the exceptional champion of the issued search. But how can we reliably find materials that exhibit exceptional performance for functions in general, for example, for catalysis, photovoltaics, or batteries? All searches face the following situation [41]:

- The number of possible materials is practically infinite.
- The electronic and atomistic processes that rule a desired materials function are many, and their concerted action is typically highly complex and intricate, resulting an immense number of possibly relevant mechanisms.
- The number of data that are 'clean' (comprehensively characterized and high-quality) and *relevant* for the function of interest are typically very low.

Under these daunting conditions we aim to identify the *rules* that govern the rare phenomena corresponding to particularly exceptional materials. Such rules describe regions in materials spaces that are relevant for the function of interest (see figure 6). In analogy to biology, the basic physico-chemical parameters entering these rules may be called 'materials genes', as they are related to processes that trigger, actuate, or facilitate, or hinder the property of interest. In particular, we are interested in such regions that (1) contain exclusively or at least predominantly materials with desired properties and (2) are described in a way that allows us to efficiently sample from them new synthesizable materials. Publicly shared materials databases and AI methods have enabled encouraging progress [41] towards this goal (see figure 7 as an example) [42]. However, critical challenges remain.

1.4.2. Current and future challenges

Most available data science and ML methods are fundamentally unsuited for the required identification of rare phenomena. Firstly, they typically aim to fit a global model to the available data by minimizing the 'regularized' average error. This focus on average global performance not only puts importance on accurately modelling the hay instead of the needles. Even worse, regularization means to deliberately avoid modelling the extraordinary for the sake of avoiding overfitting. Secondly, as pointed out by Ghiringhelli *et al* [43], off-the-shelve methods cannot reliably identify meaningful and trustworthy rules that describe exceptional materials, because they implicitly or explicitly rely on descriptors (also called representations) of materials that are either too restrictive (because they are hand-picked) or too unrestrictive (e.g., in the case of deep learning) and thus model 'non-physical' relations likely unrelated to the materials genes of relevance.

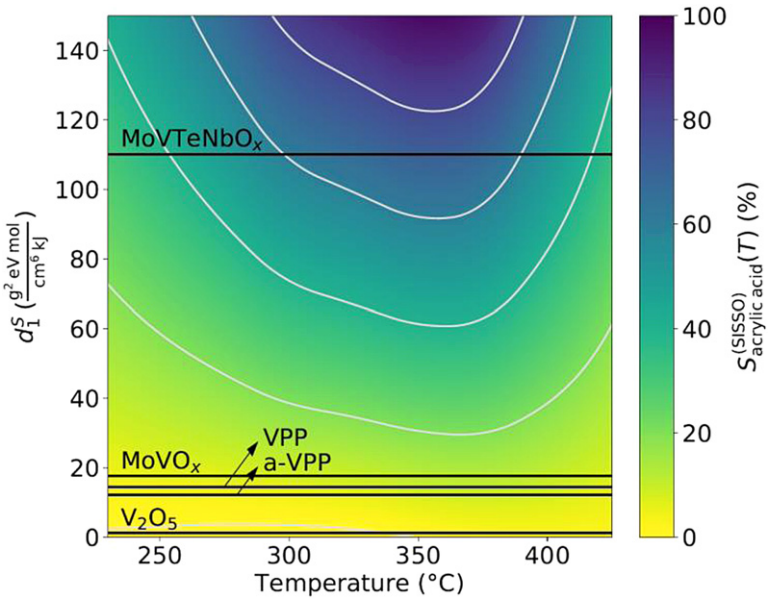

Figure 7. Map of catalysts given by the SISSO model for the selectivity of propane conversion to acrylic acid. For details see reference [42]. The desired high selectivity situation is colored dark blue. The materials used for deriving the descriptors are indicated by the black lines. The function $d_1{}^S$ is quite complex, identifying the macroscopy material, i.e. its basic properties (e.g. composition) as well as its porosity etc. $VPP_{modified}$ is a suggested material that would result by changing some materials parameters of the VPP material. In general, however, the relationship 'real materials' $\rightarrow d_1{}^S$ is not efficiently invertible.

Using symbolic regression and compressed sensing, the sure independence screening and sparsifying operator (SISSO) approach [44] alleviates this problem by identifying descriptors consisting of typically only a few analytical functions of relevant materials genes. Based on its physical plausibility and robust empirical performance, we can say with some confidence that this approach successfully identifies rules satisfying our first criterion: the description of regions that predominantly contain desired materials. A remaining problem lies in the second requirement: our ability to efficiently sample interesting novel materials. Rejection sampling can be employed to generate candidates if the considered materials class is small, e.g., binary systems restricted to a few crystal structures. However, this does not scale to the vast design spaces relevant for general searches. The central challenge is that SISSO similar to other commonly used descriptors are not efficiently invertible. While representing materials through their genes enables us to discover reliable rules, many points in gene space do not correspond to real materials, and this complicates the direct generation of new candidates from a specific region.

1.4.3. Advances in science and technology to meet challenges

An important alternative approach to rule identification is subgroup discovery (SGD) [45]. Similar to SISSO, SGD also describes non-linear relations between materials genes and properties. However, in contrast to SISSO, the SGD rules are given as Boolean conjunctions of conditions on individual genes. This means that the described regions in gene space are simple axis-parallel (hyper-)rectangles, which makes it easier to generate novel materials from them: while, as above, most combinations of the gene values may not correspond to real materials, axis-parallel conditions allow to decompose the generation process into simpler steps by considering conditions on decoupled genes independently.

Unfortunately, currently available SGD methods are, not designed to describe rare phenomena. They are based on ideas from confirmatory statistics (significance testing) to derive final conclusive results from a given dataset. To assure results that are significant for the data at hand, they prioritize the detection of relatively frequent phenomena. Fortunately, in the context of materials science, this extremely conservative approach of one-shot correctness can be relaxed. Since we have computational methods that can obtain accurate new data with reasonable efficiency, we can aim for an approach where pattern discovery and first-principles methods work in unison to facilitate rapid scientific discovery.

Borrowing ideas from Thompson sampling and Bayesian optimization [46], such rule discovery methods should propose rules that are reasonable candidates to describe the rare material champions and then obtain new simulated data from the proposed regions to validate or falsify this proposal. By repeating this process, we iteratively arrive at new regions where desired materials are more and more likely to be found. Instead of

one-shot correctness, this approach aims to identify the desired rare phenomenon as soon as possible in this iterative process by optimizing an exploration/exploitation trade-off[38].

This compelling vision provides a clear agenda of statistical and algorithmic problems to tackle: firstly, we need a sound selection mechanism for hypotheses about rare phenomena that appropriately compromises between the value of a rule and the likelihood that it can be confirmed by future data. Secondly, we need efficient algorithms that find optimal regions based on this selection mechanism.

1.4.4. Concluding remarks

In summary, publicly shared materials data and AI code, as provided by the NOMAD AI Toolkit [47], as well as physically plausible representations based on materials genes (like the ones used in SISSO and SGD) have facilitated progress towards identifying rules that describe desired materials. So far, however, all approaches are lacking either the ability to consistently describe only promising materials or the ability to efficiently generate them—at least at the ultimately required scale. To advance further, challenging statistical and algorithmic problems have to be solved, but there are promising starting points: the combination of Bayesian approaches to multiple hypothesis testing [48] as well as the versatile branch-and-bound approach [49] to discrete optimization stands a good chance to enable the envisioned methods. However, due to their reliance on adaptively generated new data, their development will require a concentrated interdisciplinary effort between materials and data science.

Acknowledgments

We acknowledge Luca Ghiringhelli, Lucas Foppa, Claudia Draxl, Wray Buntine, and Daniel Schmidt for insightful discussions and, in particular, thank Luca Ghiringhelli and Lucas Foppa for critically reading the manuscript. This work received funding from the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 951786), the NOMAD CoE, and ERC: TEC1P (No. 740233) as well as from the Australian Research Council (DP210100045).

1.5. Deep learning for spectroscopy

Milica Todorović[1,2] and Patrick Rinke[2]

[1]University of Turku, Finland

[2]Aalto University, Finland

1.5.1. Status

Spectroscopy is a fundamental tool in materials research, characterisation and discovery, and has consequently become a major objective of ML tasks. Here, deep learning based on NNs is a particularly powerful approach. NNs are universal approximators since they have the ability to represent almost arbitrarily complex relationships, as found in spectroscopy between materials properties and spectra, given the right architecture, enough neurons, layers (depth) and training data. Deep learning has celebrated first successes in spectroscopy by correlating the electronic structure and spectral properties of materials to their atomic structure[39] [50–52], functional properties [53, 54] and synthesis parameters [55, 56].

Deep learning for spectroscopy pursues two parallel goals (figure 8(a)): *spectra prediction* (typical in computational studies) and *property inference* (typical in experimental approaches). Successful NN spectra predictions allow us to cut down on the time and resources behind computational or experimental spectroscopy. Trained on available input (e.g., atomic structure or materials attributes) and output (e.g., spectra or spectroscopic quantities) pairs, the NN can make output predictions for new input instantaneously, without further resource requirements [50–52] and directly, without first computing the PES as discussed in section 2.3 of this roadmap.

In property inference tasks, data input and output are reversed to echo spectroscopic applications. NNs predict materials structure and properties from spectral input, or classify the inputs into different categories. Spectroscopy input can come in the form of spectra or spectral images. This approach to deep learning spectroscopy has, for example, been applied to extract structure information from core level [55, 56], nuclear magnetic resonance [57], vibrational [58] and Raman [59] spectroscopy, to identify cancerous cells or microbial pathogens from Raman and infrared spectra [53, 54] and to detect faulty photovoltaic modules from electroluminescence images [60].

The first deep learning spectroscopy attempts were made 30 years ago [61, 62], but the influx of modern deep NN architectures has provided a notable research boost in the last 3 to 4 years. To unlock the full potential of deep learning for spectroscopy, several challenges have to be overcome. NNs could then become

[38] Here, 'exploration' refers to sampling from regions where one is still uncertain about materials performance, and 'exploitation' refers to sampling from regions with relatively strong and certain materials performance.

[39] Atomic motion can be included by incorporating it in the spectral training data through, e.g., molecular dynamics or electron–phonon coupling.

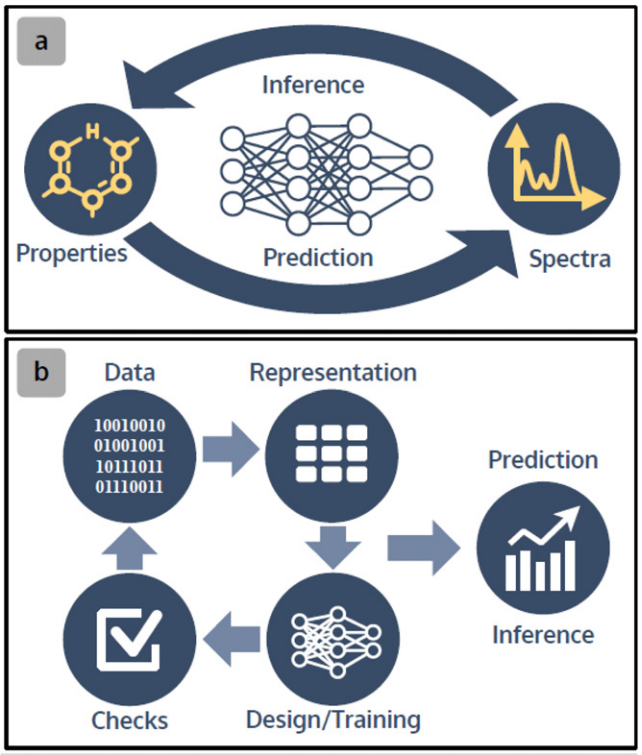

Figure 8. Deep learning spectroscopy paradigm (a) and workflow (b).

a staple in theoretical spectroscopy for fast and accurate spectra generation enabling high-fidelity and high-throughput excited state research. They could be directly integrated into spectroscopic hardware from workbench instruments to large scale infrastructures (e.g. synchrotrons) to aid diagnostics and data analysis and facilitate data-driven science [63].

1.5.2. Current and future challenges

A typical deep learning spectroscopy workflow is shown in (figure 8(b)). Each step from data acquisition, choice of materials representation, NN design and training, to testing and prediction, presents its own challenges. To advance the current state-of-the-art, we must address the issues of raw *data availability*, material *representation* and its *invertibility*, as well as *model interpretability*, *uncertainty* and *scalability* (figure 9).

Deep learning networks typically contain a large number of neurons, with parameters that must be learned during training. Although deep architectures provide the NNs with the flexibility to learn the complex relationships encountered in spectroscopy, parameter fitting requires extensive training data. Open data sets and data infrastructures are emerging in the natural sciences and engineering [60], but spectroscopy data is scarce. The challenge of data-hungry NNs needs to be addressed by *data availability* (or better data abundance), as well as more data efficient network architectures and training protocols.

A more conceptual challenge is related to materials representations, data frameworks that encode material microstructure and properties into the NN. While representation design is an active research field, it is unclear which representation types produce the most accurate and transferable deep learning models. Further *invertibility* problems arise when a representation is inferred from spectra instead of materials properties [56, 57]. This presents an obstacle to inverse prediction tasks, which now require an additional reconstruction step to retrieve the desired properties from the deep learned representation.

From the technical viewpoint, *model scalability*, *interpretability* and *uncertainty* stand in the way of rapid development. Deep learning spectroscopy requires big data, placing a burden on our computational infrastructures in both data computation and model fitting tasks. The learning process in NNs is arithmetical and abstract. *Interpretability* relates to the human desire to extract physical insight from NN models, and gain a better understanding of deep learning so we can make systematic improvements. NNs are also lacking an intrinsic measure of model *uncertainty* to indicate confidence in any individual prediction. Equipping deep learning approaches with additional information about the model, such as uncertainty, would allow us to systematically improve both spectroscopy datasets and learning quality.

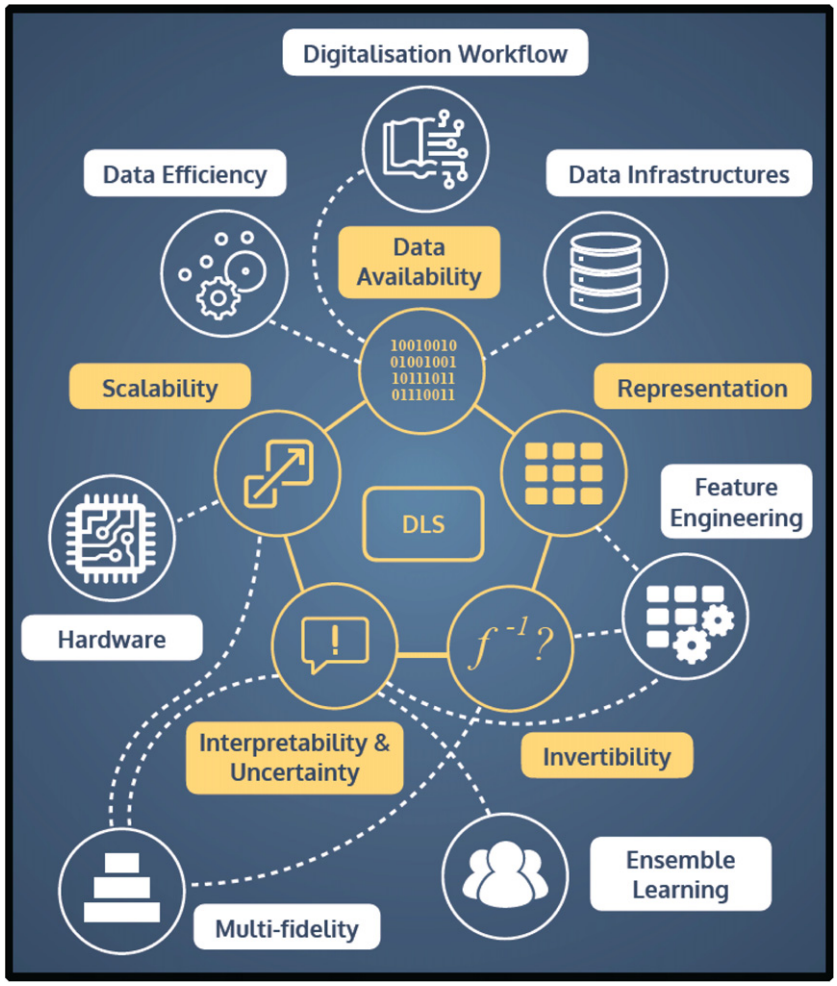

Figure 9. Deep learning spectroscopy challenges (in yellow) and scientific and technological advances (in white) to address them.

1.5.3. Advances in science and technology to meet challenges

Emerging solutions to the outstanding problems are illustrated in (figure 9). To overcome the current data scarcity in deep learning spectroscopy, ongoing simulation work is contributing open-access datasets [64], with experimental data delivered by open-source digitalization workflows developed for, e.g. multidimensional photoemission spectroscopy [65]. In the future, curated spectroscopy datasets should be made available to the community by open-science data infrastructures [63], and data acquisition workflows should be directly integrated into the instrumentation to facilitate routine data digitization in spectroscopy.

Multi-fidelity ML techniques, including transfer learning (TL), have the potential to address both *data availability* and *scalability* issues. In these hierarchical approaches, learning is based on ample but approximate low fidelity data and refined with costly high quality data points. Advances in multi-fidelity applications [66] promise to accelerate spectroscopy research: abundant data from a data-rich spectroscopy technique could be used to reduce the number of required acquisitions from a resource-intensive experiment or computational method.

The challenges associated with materials *representation* and the *invertibility* of deep learning can be mitigated by feature engineering. Feature engineering refers to the design of data representations for optimal learning, a subject of active research. Incorporating domain knowledge and constraints (e.g. invariances, uniquenesses or invertibilities) into this process would facilitate smaller network architectures and faster learning. Moreover, automated feature generation (by, e.g., a preceding NN) could produce even more compact representations or reveal features that were previously hidden to human researchers.

From the technical viewpoint, *model scalability*, *interpretability* and UQ can be addressed by innovative NN design. TensorFlow or PyTorch for Python or Flux for Julia provide examples of well-developed deep learning libraries that can facilitate the implementation of more complex learning frameworks. These deep learning software libraries, coupled with upcoming GPU architectures and hybrid GPU/CPU computing platforms, will allow us to build on current studies towards larger datasets and novel applications.

With the help of ensemble learning, we are finally gaining insight into deep learning. Using the same data to train multiple models at the same time reveals model variability and thus *uncertainty*, which can be exploited

to improve datasets and enhance learning. Ensemble learning, feature engineering, NN architecture design and multi-fidelity learning provide us with a method portfolio for tracking information uptake and processing in NNs, ultimately facilitating the *interpretability* of deep learning models.

1.5.4. Concluding remarks

Deep learning spectroscopy has become an exciting research field, brimming with innovative ideas and approaches that are employed across different types of spectroscopy. We are fast approaching generalised and transferable pre-trained models for fast predictions and industrial pre-screening. Through ongoing work, we will be able to correlate computational spectra to experimental data, facilitating the interpretation of spectroscopy signals and accelerating applications.

In the future, robust, possibly pre-trained and easy to use deep learning spectroscopy software needs to be developed for non-experts and integrated into data infrastructures or spectroscopy instruments or facilities. In time, the accumulation of studies across different research fields could make it possible to establish correlations between different types of spectroscopies. All spectroscopic responses of materials are governed by the same QM foundations. Accessing this complementarity with deep learning will allow us to combine the strengths of different spectroscopy techniques and usher in a new era in materials characterisation.

Acknowledgments

We acknowledge support from the Academy of Finland via the Novel Applications of Artificial Intelligence in Physical Sciences and Engineering Research program (Project No. 316601) and the Flagship programme Finnish Center for Artificial Intelligence (FCAI) as well as from COST Action 18234, supported by COST (European Cooperation in Science and Technology).

1.6. Machine learning for disordered systems

Corey Oses, Andriy Smolyanyuk and Stefano Curtarolo*

Duke University, United States of America

E-mail: stefano@duke.edu

1.6.1. Status

Disordered materials—characterized by extreme structural and chemical disorder—remain a critical focus for research and development. Metallic glasses couple enhanced mechanical properties—such as greater strength and corrosion resistance than their crystalline analogs—with plastic-like processability, advancing applications such as precision gears, sporting goods, and medical devices [67–69]. Chalcogenide glasses can exhibit rapid amorphous-crystalline transitions, with corresponding optical contrast changes that are useful for phase-change memory devices [70]. High-entropy solid-solutions, having several components near equimolar concentrations, offer excellent strength/ductility combinations, robust thermodynamic stability, and often properties surpassing those of the constituents [71].

ML and data-driven surrogate modeling have enabled many recent discoveries in disordered systems. Perim *et al* proposed a spectral glass-forming-ability descriptor based on the energy distribution of distinct structural phases (figure 10(a)), inspired by Greer's ansatz of necessary structural confusion during cooling [67]. The work was extended in two other studies: (i) a generalization of the descriptor for ternary compositions (figure 11(a)) [68] and (ii) the creation of an automatic phase diagram reader analyzing the eutectic angle—a proxy for its depth—for 200 chemistries with 385 eutectics (figure 11(c)) [72]. Ren *et al* employed the feedback between ML and high-throughput experiments, incorporating synthesis path information, to guide the discovery of metallic glasses (figure 11(d)) [69]. Recently, Kusne *et al* combined databases, ML, and experiments through a closed-loop autonomous framework and directed it to the Ge–Sb–Te ternary system, resulting in the discovery of a new phase-change memory material [70].

Another spectral descriptor, the entropy-forming-ability, was proposed by Sarker *et al* to quantify accessibility of random configurations in solid solutions. It led to the discovery of six high-entropy, high-hardness carbides [73]. The descriptor was extended within the Lederer–Toher–Vecchio–Curtarolo approach incorporating random configurations into a mean-field statistical model where order parameters predict the order-disorder transitions [74]. Rickman *et al* used canonical-correlation analysis and a GA to find new high-hardness multi-component alloys (figure 11(b)) [75]. Grabowski *et al* developed an approach to compute vibrational free energies of multi-component systems accounting for anharmonicity that combines thermodynamic integration and an ML potential, outperforming existing approaches in efficiency and accuracy [76].

1.6.2. Current and future challenges

The *search space* for disordered systems is ever-growing. The '$N + 1$' theorem demonstrates that, statistically, the tendency to form ordered compounds is overtaken by the configurational entropy associated with

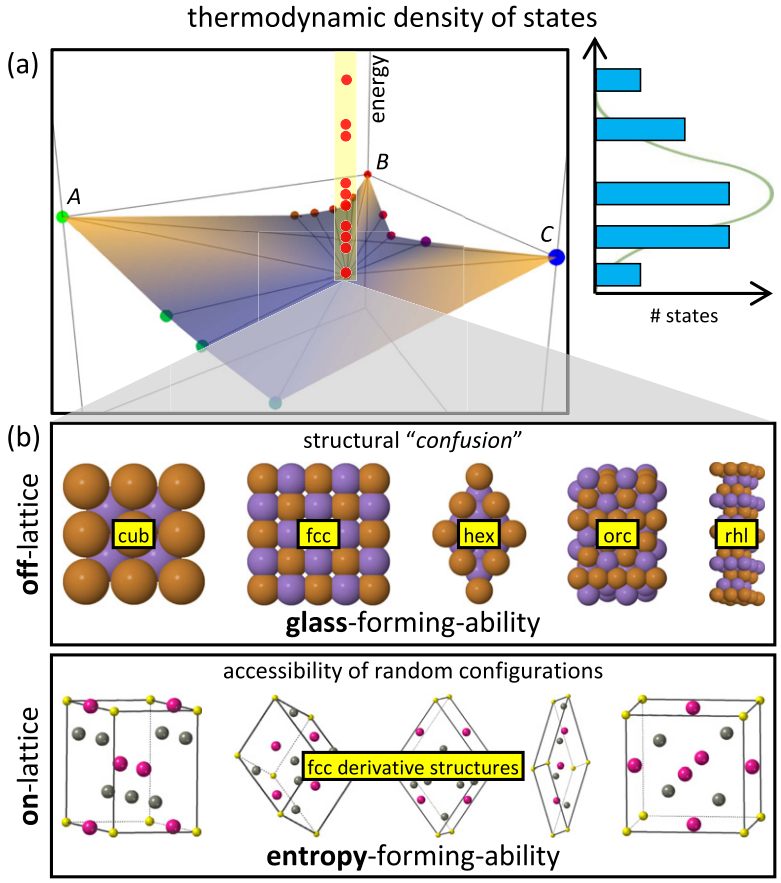

Figure 10. Illustration of the thermodynamic density of states motivating the glass-forming-ability and entropy-forming-ability descriptors as formulated in references [67, 73], respectively. (a) The thermodynamic density of states is the energy distribution of unique structural phases having a fixed composition. (b) If the set of structural phases are of different lattices (off-lattice), the distribution captures structural '*confusion*' during solidification, quantifying the glass-forming-ability [67]. Alternatively, if the structural phases are enumerated derivative structures (on-lattice), representing e.g. random ordered approximates, the distribution captures the accessibility of these configurations, quantifying the entropy-forming-ability [73].

a rising number of species, making disorder unavoidable [71]. Because the size of the search space renders trial-and-error experimentation and computational analyses difficult (even in the most efficient and high-throughput workflows [69–71, 73, 75, 76], see for example AFLOW in section 4.2), there is a need for effective and interpretable entropy-, kinetic- and synthesizabilty descriptors. Glasses are particularly challenging, as their formation is strongly influenced by processing [69] and they lack an underlying lattice on which to build configurational thermodynamics [67].

Understanding properties at *operating conditions* is also critical [70, 74, 75]. Overcoming a zero-temperature formalism requires calculation of the vibrational free energy [67]. The increasing chemical complexity is a major obstacle for computational accuracy ($\approx$1 meV atom^{-1}) as the number of parameters needed to fit reliable ML potentials quickly becomes prohibitively large [76].

The quality and availability of *data* controls the rate at which predictive models can be constructed. Much of the relevant data is published in non-standard tables and graphs, such as phase diagrams having labels difficult to interpret in an automatic fashion (figure 11(c)) [72]. Beyond accessibility, approaches relying on experimental data, while valuable, are often limited in scope, having narrow domains of applicability with regards to chemistry and stoichiometry [67, 73, 74]. Experimental data is also biased toward positive results (e.g. formation of a single phase), whereas 'negative' results (phase decomposition) are often not published [73]. Generally, ML models are excellent interpolators and poor extrapolators, calling into question whether they are suitable for the task of true-knowledge discovery. For the vast search space of glass formers and high-entropy materials, the construction of sufficiently trained and interpretable ML models remains an ambitious challenge considering the abundance of data required to capture the full set of chemistries, stoichiometries, and kinetic processes.

1.6.3. Advances in science and technology to meet challenges

Construction of open-access databases, both experimental and computational, will accelerate the development of ML and surrogate models, improving accuracy and widening applicability. Data accessibility and content are crucial. Application programming interfaces (APIs) enable automatic retrieval of data, allowing for rapid

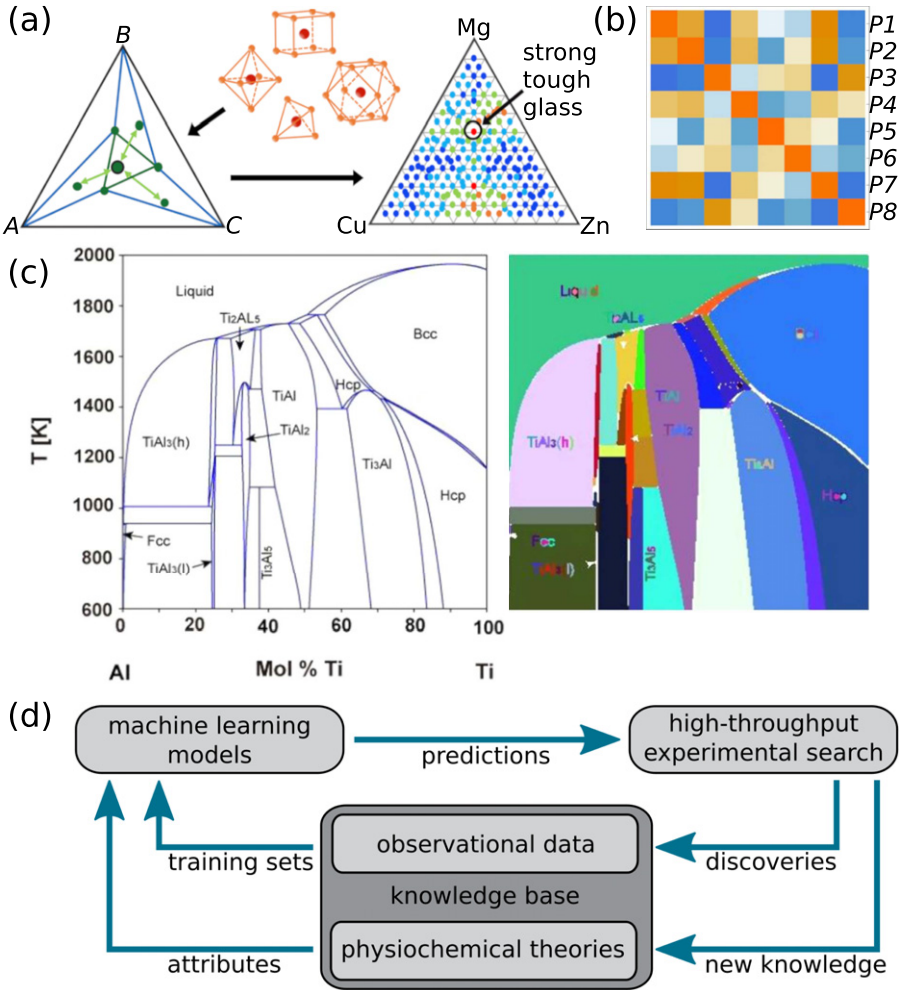

Figure 11. Various examples of ML applications to disordered systems: (a) an illustration of how the structural confusion concept determining glass-forming-ability was generalized for ternary compounds [68], (b) a visualization of the correlation matrix of features (predictors *P*1 to *P*8) used to predict hardness [75], (c) an illustration of discernible features captured by the pattern recognition software developed to analyze phase diagrams [72], and (d) an active-learning workflow between ML and high-throughput experiments [69]. (a) Reprinted from [68], Copyright (2019), with permission from Elsevier. (b) Reproduced from [75], CC BY 4.0. (c) Reproduced from [72], CC BY 4.0. (d) Reproduced from [69], CC BY-NC 4.0.

(re)training of models as algorithms and parameters are optimized and new data is made available. Standardization of simple query syntax and data structures will facilitate integration of data from multiple sources. Metadata provides necessary context for measurements and calculations—allowing researchers to integrate datasets—and should include details such as temperature/pressure/compositional ranges, classification criteria, models/equations, grid densities, and calculation parameters. Efforts should be made toward exposing graphical data in machine-processable formats (e.g., phase diagrams, x-ray diffraction patterns), and making available data not typically published ('negative' results) that would help validate models.

The curse-of-dimensionality—a reference to the overwhelming number of feature combinations conceivable—ensures that intelligent descriptor development (e.g., via surrogate features) will continue to play a vital role in modeling over brute-force feature enumeration. Quantifying concepts/insights such as Greer's 'structural confusion' and Turnbull's 'deep eutectic' in glasses, while also incorporating thermodynamic descriptions, will expedite discoveries.

Above all, integration of active-learning workflows is expected to have the biggest impact in modeling. A bidirectional feedback mechanism between ML models and experiments/calculations has shown great promise in accelerating materials discovery and property calculation [69–76]. In one case, a science-over-the-network infrastructure automates most aspects of the prediction-to-experimental-validation workflow [70], allowing for each trial to inform the next until the target is achieved. In another case, a model predicting properties of a material (e.g., an interatomic potential) employs an extrapolation-grade to assess whether a new input (configuration) deviates too far from the training set, indicating the need to expand the training set and triggering a subsequent rebuilding of the model [76]. The approach offers a systematic path to discovery and desired predictive power while avoiding the need to build arbitrarily large training sets.

1.6.4. Concluding remarks

Structural and chemical disorder provides access to unexpected properties, useful for many valuable technological applications. Still, its direct modeling remains challenging. Yet, advancements in ML are narrowing the gap. A combination of automation, development of databases/APIs, and new infrastructure linking ML models with high-throughput experiments have given rise to active-learning workflows. Human input is reduced to optional expert intervention, critical as delocalization becomes prominent. Active-learning inherently overcomes the extrapolation limitations of ML, exposing deficiencies in training data at each step and self-correcting with new measurements/calculations. The approach has shown to be the most effective way of generating new data, improving models, and exploring large spaces—like those where future, better-performing glass formers and high-entropy materials are expected to reside.

Acknowledgments

The authors thank Ohad Levy, Yoav Lederer, Donald W Brenner, and Xiomara Campilongo for fruitful discussions. Research sponsored by DOD-ONR (N00014-17-1-2090, N00014-17-1-2876).

2. Construction of accurate force fields and beyond

2.1. Machine learning for molecular quantum simulations

Alexandre Tkatchenko
University of Luxembourg

2.1.1. Status

The employment of ML approaches is transforming the field of molecular simulations (MS). This is particularly true for QM modelling, given the high computational cost of explicit first-principles calculations for solving the Schrödinger equation (SE) for systems of interacting nuclei and electrons. The widely quoted dilemma of MS consists of selecting an approximate QM method that provides sufficient accuracy and yet is computationally tractable to carry out sufficiently long molecular dynamics simulations for a system of interest. The ultimate goal of developing quantum machine learning (QML) approaches is to abolish this dilemma and achieve the accuracy of high-level QM methods in MS at the computational cost comparable to classical mechanistic force fields. As a community, we are still far from achieving this goal, nevertheless many seminal contributions in the past decade have pushed the QML field to the forefront of molecular simulations [77, 78]. For example, QML methods can now identify new phases in amorphous materials [79], allow carrying out molecular dynamics of medium-sized molecules with essentially exact QM forces [80], and offer unprecedented statistical insights into chemical environments [19, 81, 82]. Up to now, most of these applications were done under idealized conditions (small molecules in vacuum or solids under controlled conditions of temperature and pressure). Future work should concentrate on enabling tighter embedding of molecular simulations and ML methods [83], combining QM and statistical mechanics via ML algorithms, developing universal ML approximations for covalent and non-covalent molecular forces, and developing algorithms for targeted exploration of large chemical spaces of reactants, products, and transition states. Obviously, all of these advances should be continuously assessed on growing community-curated datasets of validated microscopic and macroscopic properties. The most remarkable aspect of ML techniques is that their statistical view on molecular properties often enables asking new questions and obtaining novel insights into MS. For example, ML analysis of large swaths of chemical space leads to discoveries of molecules with unexpected properties [82], offers hints for new chemical reaction mechanisms [84], or even suggests new physicochemical relations [85]. Such novel discoveries are often made by interdisciplinary teams of researchers that are able to combine their knowledge of physical laws and constraints, chemical intuition, and sophisticated ML algorithms.

2.1.2. Current and future challenges

The main challenge of QML is to develop universal models that are able to predict arbitrary QM properties of molecules and solids (total energy, atomic forces, multipoles, polarizabilities, gaps) while being as data efficient as possible. The traditional ML approach of using big data to increase performance is helpful but insufficient given that nucleoelectronic systems have many symmetries and invariances that need to be satisfied, and in fact substantially help, when predicting their QM properties. For example, many existing QML models are rather successful when predicting extensive properties (atomization/cohesive energy or polarizability), but they are much less accurate for predicting intensive electronic properties (electronic gaps, excitation energies). This creates a new dilemma for further development of QML methods: the ML models need to incorporate more physical knowledge ('quantumness'), while also being fast to evaluate and applicable to increasingly larger and more complex systems, as well as to a wider set of electronic properties.

Another pressing challenge is that ML-driven molecular simulations should strive toward achieving realistic complexity. Investigations using highly accurate QM methods normally require overly simplified model systems while more realistic model systems necessitate less accurate but computationally efficient MS 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 approximations.

Many further challenges exist that have led or will lead to mutual *bidirectional* cross-fertilization between ML and MS. The power of this path is that solving a burning problem in MS 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 natural sciences typically requires novel ML models and unforeseen scientific innovations, and this can lead to interesting insights that are not necessarily limited to molecular simulations.

2.1.3. Advances in science and technology to meet challenges

ML is a relatively new technology compared to decades of developments of QM and statistical mechanics techniques in the field of MS. Hence, many complementary directions are being explored at the moment, some of which lead to important advances. For example, hundreds of different representations (a necessary input to any ML model) have been proposed to model interatomic interactions in molecules and solids. Most of the available representations trade efficiency vs quality of the description. This situation can be compared to the proliferation of different density-functional approximations (DFA) for electronic-structure calculations. Eventually, the community should agree on a reasonably small set of useful and practical representations.

An emerging idea is to directly learn computationally efficient model Hamiltonians for electronic interactions based on correlated wavefunctions, DFA, tight-binding, molecular orbital techniques, and/or the many-body dispersion method. ML can predict Hamiltonian parameters and the QM 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.

One important aspect that QML approaches enable is providing a novel perspective on exploring increasingly larger chemical spaces for designing molecules and materials with desired properties. However, any such exploration requires reliable QM data for a larger set of systems of interest. The recent emergence of comprehensive datasets (such as NOMAD, Materials Project, GDB, among others) is very welcome and this path of creating validated, easily accessible, and trustworthy data should be further pursued.

2.1.4. Concluding remarks

Molecular simulations have been significantly advanced with ML approaches. However, many challenges remain and solving them will require coming up with creative interdisciplinary approaches combining quantum and statistical mechanics, chemical knowledge, and sophisticated ML tools, firmly based on growing datasets that cover increasingly broader domains of the vast chemical space.

Many advances in this field required mixed teams with members educated in different aspects of physics, chemistry, mathematics, and computer science. Going forward, this field 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 research contribution to the exciting and emerging field of ML-driven molecular quantum simulations.

Acknowledgments

AT acknowledges the European Research Council (ERC) and Luxembourg's Fonds National de la Recherche (FNR) for their generous funding.

2.2. Bayesian machine learning for microscopic interactions

Albert P Bartók

University of Warwick, United Kingdom

2.2.1. Status

A long-held promise of atomic simulation has been to serve as the ultimate toolset to predict physical properties and interpret experimental phenomena, thus capable of *ab initio* materials and molecule design. While methodological, software and hardware developments have significantly increased the capabilities of QM packages, time and length scales necessary to capture multiscale phenomena are still out of the boundaries of first-principle calculations. Traditionally, interatomic potentials have been proposed to represent microscopic interactions in a computationally efficient way. The design of such potentials is typically based on theoretical considerations, containing a small number of free parameters, which can be found from matching the

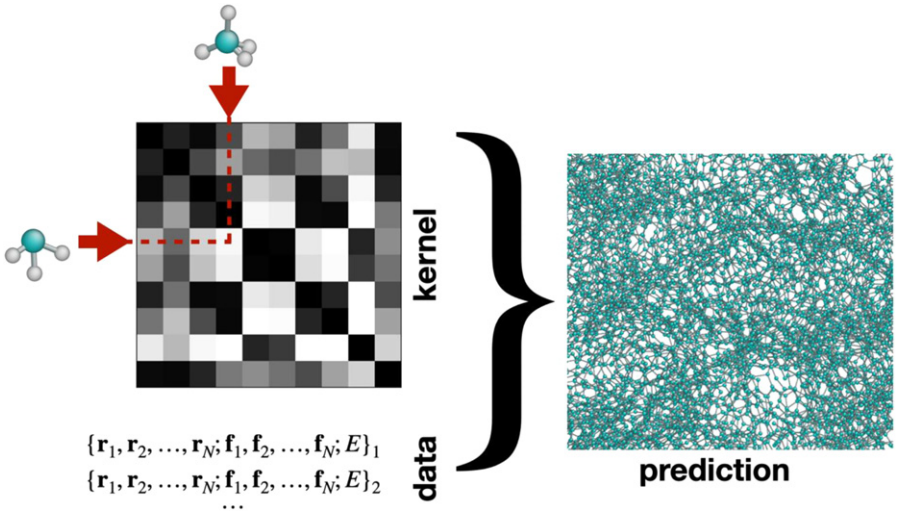

Figure 12. Schematic representation of the GAP framework. *Ab initio* data collected on systems tractable by QM calculations and GPR are used to fit models which can be evaluated for significantly larger configurations without loss of accuracy.

behaviour of the model to limited experimental or computed data. However, predictive capabilities of simulations based on the fixed functional forms employed by traditional interatomic potentials, have proved severely limited in all but the simplest cases, necessitating more flexible approaches.

The demand for adaptive models, capable of describing multiple bonding situations of the same material simultaneously, suggests a significantly expanded parameter space, which necessitates larger amount of data to determine the parameterisation. Thanks to the availability of reliable and efficient *ab initio* software packages, microscopic data, as opposed to macroscopic observables, can be generated in abundance and utilised in fitting flexible interatomic potential models. This had been recognised well before ML has become ubiquitous, and formulated, for example, as the force-matching embedded atom model [86], the ReaxFF force-field [87], or PESs for molecular systems.

ML interatomic potentials pushed this idea to the extreme, by disposing of most of the physics-based considerations of the functional form, replacing it with a non-parametric regressor that imposes little or no constraints on the mathematical form of the interaction, and relies chiefly on data. Gaussian process regression (GPR) is a Bayesian technique that imposes a prior in the form of a distribution of functions and uses data as evidence to provide predictions [88]. Gaussian approximation potentials (GAP), see figure 12 [89] represent a practical realisation of a ML potential, based on a combination of sparse GPR and a purpose-built kernel called smooth overlap of atomic positions [90], which have proved highly successful in molecular and materials science modelling.

2.2.2. Current and future challenges

The highly flexible form of GAP is both a blessing and a curse in that the physical behaviour of the model is derived from data, therefore adequate data coverage is essential to constrain the interaction function at all atomic environments that are sampled at conditions of the simulator's interest. Even though it has been demonstrated that general-purpose potentials for single-component systems can be generated [91], multi-component systems remain challenging due to the significantly increased complexity of the configurational space. To automate the data collection, iterative and AL approaches have been investigated, but it is important to note that the computational cost of fitting GAPs increases linearly with database sizes. Although the cost of evaluating a GAP is controlled by the sparsification applied on the data set, the number of representative points necessary to achieve an accurate fit is also expected to increase with the complexity of the PES. It is reasonable to assume that generating GAP models of more complex materials, and especially disordered phases, may become impractically expensive due to the (i) amount of *ab initio* data that needs to be computed; (ii) GAP fitting procedure; (iii) GAP evaluation; or a combination of these.

Long range electrostatic and dispersion interactions, resulting from charge transfer and polarisation, pose another challenge to ML potentials that are primarily optimised to capture the energetics of localised, chemical bonding. Due to screening effects, the effective range of electrostatics may be significantly reduced, and therefore the majority of the interaction may be captured by a local model, but this approach is, in general, detrimental to transferability.

GPR has the advantage that the posterior distribution of the model is available, providing not only a prediction for the mean, but the variance as well. In practice, however, the error estimate computed from the

predicted variance has only been found to be quantitative for simple, low-dimensional fits, such as two- and three-body interaction terms [92]. A robust error prediction would significantly aid the reliability of GAP and other GPR based potentials, and also enable further automation of the potential generation process. Further studies are required to understand the failure of the error prediction.

2.2.3. Advances in science and technology to meet challenges

Methodological improvements will need to address the current performance limitations of the GAP framework when applied on complex, multicomponent materials. Recent studies demonstrated that feature selection techniques can greatly reduce the computational cost of the evaluation of the potential, at a cost of a modest and controllable loss of accuracy. The general ML community has achieved significant advances in sparsification technology [93], which should be evaluated in the context of interatomic potentials and adopted where the advantages are evidenced.

Fitting GAP potentials is computationally expensive, and for larger data sets memory requirements necessitate specialised hardware. Developments, either methodological or concerning the software implementation, should be directed to utilise standard parallel architectures, thereby speeding up fits and eliminate the need for large memory machines. Not only would this democratise the fitting process, but it would allow exhaustive hyperparameter optimisation, resulting in more robust and transferable potentials.

Currently two aspects of PESs, locality and smoothness, are explicitly built into the Bayesian prior of the GAP framework. Neither is fully general, nor do they capture all the common features of atomic interactions. Incorporation of more physical priors would improve transferability of ML potentials, reduce the amount of data that is required for training and potentially increase the computational efficiency of evaluating the models. A unified Bayesian model for long-range electrostatics would incorporate such a prior, resulting in improved efficiency, accuracy and transferability. Similarly, short range repulsion from the Pauli exclusion principle is a fundamental property of atomic interactions, but currently it is either learned from the data or treated via a pair term, which is fitted separately. A prior encoding our physical understanding would be the Bayesian solution.

Finally, these improvements of the GAP methodology and software would also revolutionise the workflows for database generation. Currently a lengthy iterative process is required to generate the necessary data [94], often exploring highly unphysical configurations. Elimination of incorrect bias from suboptimal hyperparameters also depends on using large amounts of fitting data. A more restrictive, but physical prior would alleviate this reliance on large databases, leading to quicker and more reliable fitting protocols.

2.2.4. Concluding remarks

Much beyond a proof-of-principle concept, ML interatomic potentials have matured to be utilised as tools to gain, hitherto impossible, quantitative understanding of microscopic processes and to make accurate macroscopic predictions [95]. Challenges posed by complex potential energy landscapes can be addressed by further developments of the ML framework, leading to more transferable and cheaper models that can be fit from a small amount of data in an automated way. Closer integration of physical priors into the formalism can be viewed as a step towards traditional interatomic potentials, but without the loss of generality and rigorous mathematical treatment of regularisation.

2.3. Spectroscopically accurate potential energy surfaces (SAPES) from machine learning

Sergei Manzhos,[1] Manabu Ihara[1] and Tucker Carrington[2]

[1]Tokyo Institute of Technology

[2]Queen's University

2.3.1. Status

Solving the SE describing the motion of the nuclei, and in particular calculating vibrational (e.g. infrared) spectra, presents a stringent test for ML potentials, as spectra are sensitive to the global quality of the PES. Comparing computed and experimental observables is the ultimate test of the quality of a PES, and thereby the usefulness ML fitting methods. Errors at test points are less informative. To compute spectra with the desired sub-cm^{-1} accuracy, PES errors must be much smaller (on the order of a cm^{-1}) than those admissible in MD and quantum reaction dynamics calculations (where PES errors of hundreds of cm^{-1} are not uncommon) [96].

Some methods of solving the SE require a potential in sum-of-products (SOP) form [97, 98]. (NN, see section 2.4 for a more general account of NN potentials) allow achieving a SOP naturally by using exponential neurons [99]. The accuracy is competitive with alternative SOP schemes such a *potfit* [97], even when using fewer terms [100]. The first NN spectroscopically accurate potential energy surfaces (SAPESs) were produced by Manzhos and Carrington in 2006 and the methods can now be used routinely [96, 101, 102]. NN (and other methods) have been combined with permutationally invariant polynomials (PIP) to ensure symmetry.

It is important to ensure the correct symmetry to achieve a SAPES, although symmetry can always be restored by averaging at symmetrically equivalent points. In reference [101], for the first time, SAPESs for CH_4 were constructed from the same *ab initio* data with different ML methods (NN, PIP-NN, interpolating moving least squares) and full-dimensional variational calculations were used to assess the spectroscopic accuracy of the PES. All methods resulted in PES errors (at test points) and spectrum errors of the same order of magnitude (both on the order of several cm^{-1}). PIP-based methods gave a lower PES error whereas NN gave a spectrum slightly closer to the experiment. Structures and harmonic frequencies were practically the same.

More recently, GPRs were shown to produce spectroscopically accurate PESs from less data than required by a NN for the same accuracy [103]. Combining both NN and GPR with a n-mode representation/high dimensional model representation (HDMR) [104] can improve ML fits from sparse data and achieve SAPES.

2.3.2. Current and future challenges

Today, several ML methods (notably NN, GPR) can achieve SAPESs, with similar errors, in a routine and black-box way for molecules with five or more atoms. However, truly comparative (using the same data and computing spectra with accurate methods) studies of SAPESs are still scarce and more such studies are needed. SAPES for larger systems remain a challenge but the challenge is shifting from building the PES to developing methods for accurately computing spectra. The comparative study on methane [101] showed good accuracy of the PES and the spectrum with all methods, but comparative data on other and less symmetric five-atomic molecules are still outstanding. The comparative study of NN vs GPR also suggests that the PES error may be much (by a factor of 50) larger than the spectrum error [103]. Judging the quality of a PES on the basis of errors at a set of test points can be misleading. More comparisons of not only PES errors achieved with different methods but of resulting spectra are needed to determine the best way of fitting a PES and typical required test point errors.

This is related to another challenge—data distribution (sampling). Smart point selection schemes have brought significant advantages to reactive PESs [96]. There are indications that significant benefits can be reaped from point optimization when computing spectra [105, 106], but for SAPES this is yet to be explored and used in applications. As sampling of multidimensional PESs is necessarily sparse, methods to avoid or detect 'holes' (that significantly deteriorate the spectrum) are desirable.

A major challenge for SAPES construction remains molecules on surfaces or nanoparticles, which are of importance to technologies such as fuel cells, industrial and photo-catalysis etc. Accurate computational spectroscopy has been largely absent from this field notably because of the lack of SAPESs, even though it is desired, in particular, for accurate species assignment. Reported PES fitting errors for molecules on surfaces are relatively high ($\gg$10 cm^{-1}), which are compounded with the low accuracy of the underlying *ab initio* methods (typically DFT with a GGA functional).

Recently, powerful black-box NN based methods have emerged that allow mapping between structure (including atom types as well as positions) and properties that can *also* be used for PES construction [107]; however, their performance for spectroscopy is still not explored.

2.3.3. Advances in science and technology to meet challenges

To fully utilise the potential of ML in constructing SAPESs, further developments in methods of computational spectroscopy are needed that will allow calculations on five- and more-atom systems with exact kinetic energy operators (KEOs) and arbitrary coupling. To construct SAPESs from very sparse data, combining a PES representation with lower-dimensional functions, either via HDMR or via dimensionality reduction, and ML are very promising [104], in particular, to reduce the risk of 'holes'.

GPR has recently emerged as a powerful tool with several advantages over NN, achieving better accuracy with fewer data (or requiring fewer data for the same accuracy). Being a non-parametric method with which it is much easier to avoid overfitting (and 'holes'), it also can deal with high-dimensional data (although it becomes costly with more than ~10 000 training data). Expanding the use of GPR can help address the challenge of SAPES for molecule-surface systems precisely because it allows using fewer, and therefore higher-accuracy, data.

In the next several years, ML combined with more accurate vibrational spectroscopy computational methods will be applied in solid state and on interfaces. Collocation [108] will make it possible to get accurate spectra by considering a number of degrees of freedom without any limitations on the degree of coupling and the KEO, and vibrational self-consistent field, because it is easy to apply, will provide more accurate spectra than a harmonic approximation, in particular, by using an n-mode representation of the PES. ML will be used either to build the entire PES or component functions in an HDMR representation. Experimental spectra of molecules on surfaces are poorly resolved (~1 cm^{-1}) and probe only low-lying states, which reduces the required accuracy of the PES. It is also possible to avoid building SAPESs for surfaces and other difficult cases and instead to use collocation and compute the potential at all the collocation points [108]. Most ML SAPESs used supervised

ML. Unsupervised approaches are promising especially in the area of selecting optimal sampling points and are awaiting in-depth exploration when applied for SAPES.

2.3.4. Concluding remarks

PES fitting is the bridge between the calculation of *ab initio* points and the application of a method for computing a vibrational spectrum. Having a fitted PES significantly reduces the number of required *ab initio* points. It is now possible, and more importantly, easy to make SAPESs using black-box ML fitting methods. Previously, it was necessary to develop physically motivated fitting functions for each problem. This requires knowing much about the molecule for which one wants to fit a PES. Using ML methods obviates the need to tinker with a fitting function and makes it almost trivial to build the bridge. This reduces the task of computing a spectrum to choosing an *ab initio* method, running quantum chemistry calculations and then choosing a dynamical method and computing the spectrum.

Acknowledgments

TC is supported by the Natural Sciences and Engineering Research Council of Canada.

2.4. High-dimensional neural network potential energy surfaces in chemistry and materials science

Jörg Behler
Universität Göttingen

2.4.1. Status

Machine learning potentials (MLP) have become an important tool for atomistic simulations in chemistry and materials science, because they can provide energies and forces with the accuracy of electronic structure methods at a small fraction of the computational costs. The first MLPs have been introduced about 25 years ago by Doren and co-workers [109] employing ANNs. This first generation of NN potentials, which has been explored by several groups in the following decade, demonstrated the high accuracy of MLPs but was still restricted to low-dimensional systems depending only on a few degrees of freedom. MLP became applicable to high-dimensional condensed systems containing thousands of atoms through the introduction of high-dimensional neural network potentials (HDNNPs) by Behler and Parrinello in 2007 [110], which represented the first example of a second-generation MLP. In this approach, which is common to many modern MLPs, the total energy is constructed as a sum of environment-dependent atomic energies that in case of HDNNPs are delivered by a set of atomic NNs. The underlying assumption about the locality of the atomic interactions works surprisingly well for many systems, as long as the considered chemical environments are sufficiently large. Still, in many cases long-range electrostatic interactions are important. These have been included in the third generation of MLPs by making use of environment-dependent atomic charges, for instance expressed by a second set of atomic NNs [111]. These charges are then used to compute long-range electrostatic interactions by explicitly evaluating Coulomb's law. Nevertheless, third-generation HDNNPs are still local and do not allow to take global dependencies of the electronic structure such as non-local charge transfer or even changes in the total charge of the system into account. These phenomena can be included in the fourth generation of MLPs employing global charge equilibration techniques. A first method applicable to ionic systems has been the charge equilibration NN technique [112], which has recently been combined with HDNNPs to yield a fourth-generation 4G-HDNNP [113] that is applicable to a wide range of systems (figure 13).

2.4.2. Current and future challenges

In the past two decades, methodical advances have substantially extended the applicability of high-dimensional NN potentials. Still, several challenges remain. A first challenge is the further incorporation of physical knowledge, with the inclusion of electrostatic interactions in third- and fourth-generation HDNNPs being a first step. For instance, currently a lot of work is in progress to also incorporate dispersion interactions, which represent a comparably small but important contribution to the potential-energy surface of many systems and can also be rather long-ranged. Several approaches are possible to include dispersion interactions, which can either be assigned to the third- or fourth-generation. It should be noted that both of these generations include long-range interactions without truncation, while the central quantities like charges, or dispersion coefficients, have a local or non-local dependence, respectively. Further interesting extensions could involve the charge density or atomic spins, which might in the long-term perspective open the possibility to construct HDNNPs for the simultaneous description of several electronic states.

Another challenge is the validation of HDNNPs and MLPs in general. While ML methods can reproduce available data very accurately, they often have very limited extrapolation capabilities, and thus the knowledge about the range of validity of a given potential is of vital importance. The central problem is that the validation is most challenging in the absence of reliable reference data, while just in this situation quality control is essential. Therefore, improved methods for detecting unreliable predictions are needed. Estimates of the reliability

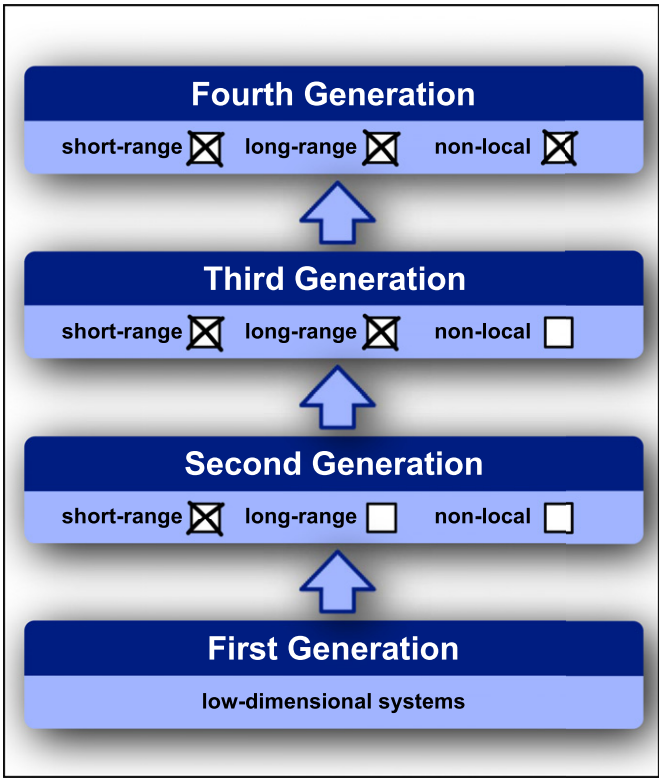

Figure 13. Generations of NN potentials.

and the relevance of novel atomic configurations encountered in atomistic simulations can be made based on ensembles of NNs [114]. Such AL strategies are a very important field of research [115], and connect the validation challenge to the challenge of constructing suitable reference data sets. These sets should be as small as possible to enable an efficient construction of the potential, while a large diversity of structures is needed to achieve transferable potentials. The identification of the atomic configurations needed to cover the relevant part of the potential-energy surface remains a crucial aspect of the development of all types of MLPs.

2.4.3. Advances in science and technology to meet challenges

The construction of MLPs is a very interdisciplinary field, which benefits from advances in many different areas. Along with the progress in the construction of more reliable potentials as outlined above, the accuracy of the underlying electronic structure calculations is becoming increasingly important, since MLPs cannot be more accurate than the underlying data. While DFT calculations at the level of the generalized gradient approximation are still dominant for condensed systems, it has been recognized that the level of hybrid functionals would be desirable for many systems. The substantially higher costs of these functionals require further advances in the efficiency of modern DFT codes as well as in computer hardware. Hence, the construction of the reference data sets will remain the computational bottleneck in the development of HDNNPs.

In contrast to the comparably mature field of electronic structure calculations, the technology of ML algorithms, which are nowadays penetrating every aspect of life, is advancing very rapidly. Many modern software tools and libraries are now available and lower the barrier for entering the field of MLP development by facilitating the construction of potentials. In this context it is important to note that the classification scheme of MLPs into generations is not fully applicable to all types of MLPs, including also some flavors of NN potentials. An example is represented by message passing NNs like atoms-in-molecules network (AIMNet) [82], which pass information about the atomic environments through the system. Consequently, the interaction range that can be described is related to the number of passing steps does not depend on a fixed cutoff radius as employed for instance in HDNNPs.

Another challenge concerns the development of suitable descriptors to characterize the atomic configurations, which has been a fundamental problem of early NN potentials. With the introduction of second-generation MLPs a breakthrough has been achieved [116], which resulted in descriptors compatible with the mandatory rotational, translational and permutational invariances of the PES. Although many different types of descriptors are available nowadays meeting these requirements, some fundamental limitations like the unfavourable scaling with the complexity of configuration space in terms of the number of chemical elements

remain unsolved. Therefore, with increasing possibilities to construct large data sets, a general solution of this scaling problem is now becoming more and more urgent.

2.4.4. Concluding remarks

In summary, the development of high-dimensional NN potentials, like the development of MLPs in general, is a rapidly growing field which has not yet reached its peak. Starting with first potentials suitable for rather small molecular systems, over the years NN potentials have been extended to high-dimensional systems containing thousands of atoms, now including long-range interactions based on atomic charges taking non-local charge transfer and even different global charge states into account. All these developments have enabled simulations of increasingly complex systems in almost all fields of chemistry, materials science, and even biomolecular systems. Several challenges remain, like the construction of representative and high-level reference data, the validation of the obtained potentials, and the derivation of improved descriptors for chemically more diverse systems. In particular the inclusion of physical knowledge recently has received a lot of attention, and many new interesting developments can be expected in the coming years.

Acknowledgments

Funding by the Deutsche Forschungsgemeinschaft (Be3264/12-1, project number 405479457) is gratefully acknowledged.

2.5. Transferable neural network force fields

Olexandr Isayev

Carnegie Mellon University

2.5.1. Status

In the area of *ab initio* molecular simulations, DFT calculations have become a workhorse of computational organic chemistry. But we face a dilemma: standard computational algorithms for N-electron systems require $O(N^2)$ storage and $O(N^3)$ arithmetic operations. This $O(N^3)$ complexity is a critical bottleneck that limits capabilities to study larger realistic chemical systems and longer time scales relevant to the biological experiments. One solution to these problems is the development of empirical potentials built with ML methods [117]. The ML potentials have seen remarkable progress during recent years and have proven their ability to accurately predict energies and forces of molecules when trained on a properly developed dataset.

Behler and Parrinello introduced the idea of HDNNPs [110]. In HDNNPs, the total energy of the system is computed based on atomic contributions: $E_{\mathrm{tot}} = \sum_{i=1}^{N_{\mathrm{atom}}} NN_i(G_i)$ where G_i depends on the atomic coordinates and the local environment within a given cutoff distance. The cutoff distance limits the interatomic interactions with the neighboring atoms that are contributing to the structural fingerprints in the form of many-body symmetry functions G_i. The G_i are constructed for every atom and are used as input vectors to the atomic NN, which computes the energy contribution of the atoms to the total energy E_{tot}. HDNNPs are trained to describe one molecular system at a time. Therefore, one of the main issues of HDNNPs is transferability. The potential needs to be retrained for every new application.

This problem has been addressed with the development of new methods that provide general-purpose models like Accurate NeurAl networK engINe for Molecular Energies or ANI. In the ANI model, Smith *et al* developed a modified symmetry functions G_i (Justin Smith symmetry functions or JSSFs) that allowed overcoming these limitations for organic molecules [118]. ANI-1x model uses iterative AL procedure with training to a large and diverse dataset of molecules [119]. The initial ANI models were developed for neutral organic molecules consisting of four elements (HNCO). Subsequently, ANI-2x models were extended to seven elements (CHNOSFCl) [120] and even nine [82]. Overall, the ANI methodology provides a systematic approach for generating atomistic potentials (figure 14). It drastically reduces the human effort required for fitting a force field and automates their development. Using an NNP does not require one to choose a functional form.

2.5.2. Current and future challenges

Most NNPs, including ANI models (figure 15(a)), are inherently local in how they describe chemistry. Adding missing long-range interactions is needed for an accurate description of realistic chemical systems. One route to do this is to predict atomic point charges for modeling the long-range Coulomb potential.

Using multi-modal training, one can predict atomic charges together with energies and forces. The 'AIMNet' architecture (figure 15(b)) was inspired by the quantum theory of atoms in molecules. The AIMNet lifts multiple limitations in NNPs. It encodes long-range interactions and learnable representations of chemical elements. Several alternative approaches were also proposed in SchNet [121] and HIPNN [122] models. The AIMNet model utilizes the idea of multi-modal learning, making a simultaneous prediction of different atomic properties based on one common layer. This layer is enforced to capture the relationships across multiple learned modalities and serves as a joint latent representation of atoms in the molecule.

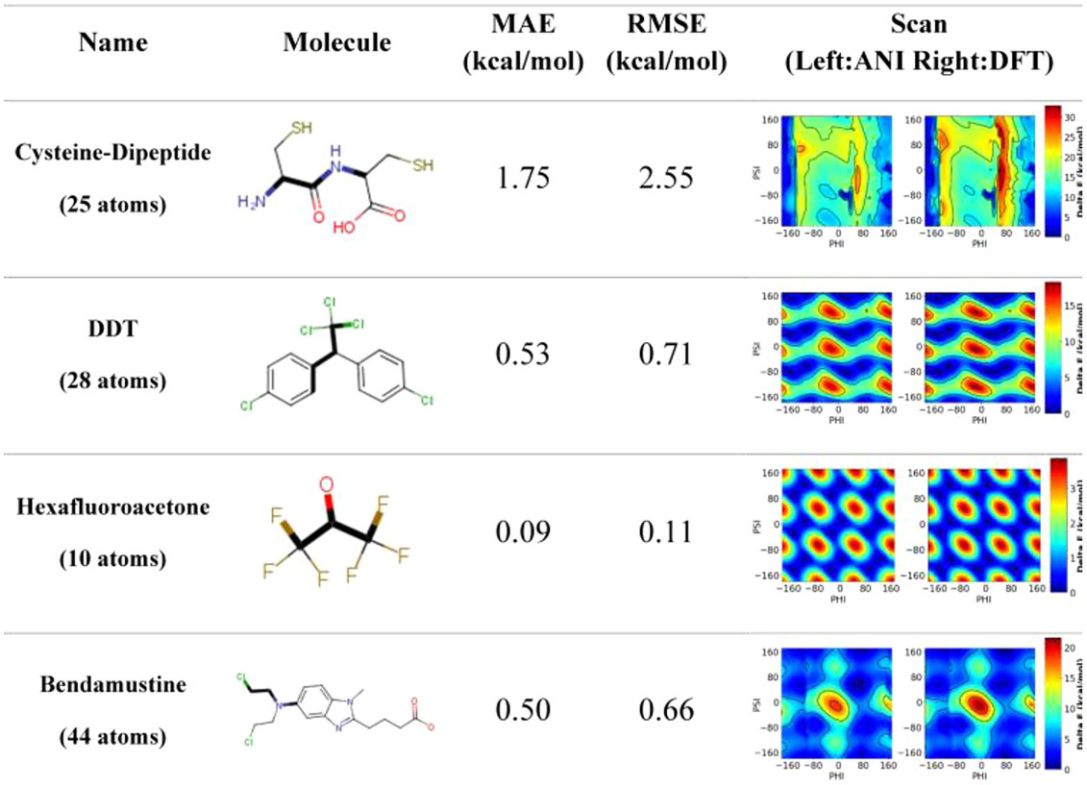

Figure 14. An example of the accuracy of the ML PES scans. Relaxed 2D torsion profiles for ANI-2x (left) and DFT (right). Two dihedrals (shown in bold) were rotated about one another to generate the corresponding confirmations. The bonds composing the scanned dihedrals are highlighted in bold. The middle columns show the MAE and RMSE respectively, of the relative energies in kcal mol^{-1} between ANI and DFT. Reprinted with permission from [120]. Copyright (2020) American Chemical Society.

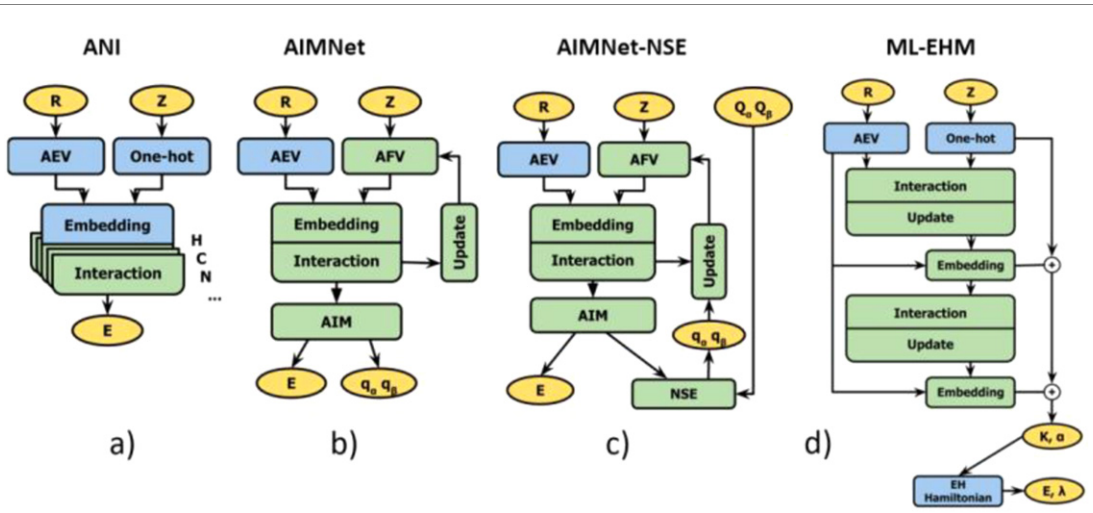

Figure 15. Several NN architectures for atomistic force fields. (a) ANI; (b) AIMNet; (c) AIMNet-NSE; and (d) ML-EHM. The yellow blocks show input data (coordinates R, atomic numbers Z, total molecular charge Q) and output (energies E, spin-polarized charges q, EHM diagonal matrix elements, α and empirical fitting coefficient K). The green blocks indicate NN blocks for training and the blue blocks show mathematical transformations. Reprinted with permission from [124]. Copyright (2021) American Chemical Society.

Most NNPs have so far been trained on only either closed-shell or open-shell structures and therefore cannot correctly describe effects of spin and multiplicity. As the first step in this direction, recent work introduced the AIMNet-NSE (neural spin-charge equilibration), figure 15(c) architecture to learn a transferrable potential for organic molecules in arbitrary charge states [123]. Conceptually the neural spin equilibration (NSE) module serves as a neural charge- and spin-equilibration scheme by redistributing spin-charges through the iterative procedure and making energy prediction based on the distribution of alpha and beta spin densities. In contrast to the standard geometric descriptors, the AIMNet-NSE model incorporates adaptable *electronic*

information into ML models. It could be applied as a fast and reliable method to compute multiple properties like ionization potential, electron affinity, spin-polarized charges, and a wide variety of conceptual DFT indexes.

Another direction of NNPs model development is focused on capturing the correct physical behavior by combining physical models with ML [124]. This so-called physics-aware AI models promise to improve generalization by forcing ML models to obey physical laws and symmetries. The simplest of such models could be ML combined with the extended Hückel method or ML-EHM (figure 15(d)). ML-EHM predicts a set of molecule- and environment-dependent Hamiltonian elements to predict Frontier orbitals and energies approaching DFT accuracy.

2.5.3. Advances in science and technology to meet challenges

One of the major concerns of ML force field development is the reference data used for training. The quantity of data that can be used for training is limited due to the high computational cost of QM. Therefore, many models are developed to address one specific application. This severely hinders the applicability of NNPs in practice. This issue might be mitigated with advanced training strategies that take advantage of AL and TL. These algorithms can help not only to decrease required reference data but also improve the accuracy of NNPs. Training ML models for every QM method is also impractical. Developing multi-theory ML models and *data fusion* is a critical bottleneck in constructing robust ML-accelerated QM methods. TL can be used to retrain an existing model with additional training data to extend the domain of applicability.

Most of the NNPs available in the literature provide only deterministic predictions and cannot model uncertainties. It is important distinguishing between at least two different types of uncertainty, often referred to as aleatoric and epistemic. Epistemic uncertainty results from the lack of knowledge about the system and could be addressed with the accumulation of more training data. In contrast, the aleatoric or statistical uncertainty is associated with a model. The incorporation of probabilistic methods and Bayesian NNs will help to capture inheriting model uncertainty.

An *explainable* ML model is also essential to understand, appropriately trust, and effectively develop a proper physical model. Thus ML/AI models are expected to incorporate physics knowledge in their design and architecture. This includes conservation laws, causality, symmetry, geometrical and topological properties, constraints, and more. We envision that novel approaches will be able to interpret and highlight those physical priors learned by the models. Next-generation analysis and design tools will help domain scientists think about new ideas and find underlying physical laws in visual and straightforward, yet interpretable ways.

2.5.4. Concluding remarks

Recent years showed substantial progress in NNP development and their applications toward a variety of molecular systems. They are promising to change the way how force fields are constructed. Atomistic ML potentials offer accuracy comparable with QM methods but many orders of magnitude faster in many cases. NNPs are already used to find reliable conformational energies for molecules, re-parametrizing existing force fields, protein-ligand free energy calculations. But this power comes with great responsibility. Reflecting on the famous quote attributed to Derek Lowe, '*It is not that machines are going to replace chemists. It is that the chemists who use machines will replace those that do not*'. We are currently witnessing a transformation of chemical sciences into a novel data-driven field. This requires deep methodological and cultural change coupled to educational and workforce development programs at the professional, graduate, undergraduate, and even high school levels.

Acknowledgments

The work performed by OI was made possible by the Office of Naval Research (ONR) through support provided by the Energetic Materials Program (MURI Grant No. N00014-21-1-2476). This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the US Department of Energy (DOE) Office of Science. This research is part of the Frontera computing project at the Texas Advanced Computing Center. Frontera is made possible by the National Science Foundation award OAC-1818253.

2.6. Integrated machine learning models: electronic structure accuracy beyond local potentials

Max Veit, Andrea Grisafi*, Jigyasa Nigam and Michele Ceriotti

École Polytechnique Fédérale de Lausanne, Switzerland

∗ Present address: PASTEUR, Département de Chimie, Ecole Normale Supérieure, 75005 Paris, France

2.6.1. Status

Electronic structure calculations have progressed to a level of accuracy which makes modeling of atomic-scale systems from first principles truly predictive. By computing energy and forces corresponding to the ground-state Born–Oppenheimer PES they enable molecular dynamics simulations that explore the structural

landscape, and assess the stability of different configurations. What is more, electronic structure methods provide a wide spectrum of properties available either as a by-product of the calculation or as a post-processing step, so that the prediction of functional properties, electronic responses, and experimental observables are available with similar accuracy and transferability to those achieved for the ground-state energetics.

Unfortunately, the high computational cost, and its steep scaling with the number of electrons included in the simulation, limit the time and length scales that are accessible to simulations. The last decade has witnessed the emergence of ML techniques to address these limitations, and to bring the accuracy of first-principles methods to the types of simulations that are needed to model complex materials and molecules in realistic conditions. Using the atomic identities and positions as inputs, appropriately processed to incorporate fundamental symmetries and physical insights, ML techniques make it possible to fit structure-property relations using a very flexible functional form and a limited number of reference calculations. Once trained, the model can be used to inexpensively predict the same kind of properties for any set of new, yet similar, structures, paving the way to the calculation of thermodynamic observables that can directly be compared with experiments.

Most of the established approaches, including the ones discussed in sections 2.1–2.5 of this Roadmap, focus on the prediction of interatomic potentials [125]. As ML potentials are employed to make predictions of more and more experimentally accessible materials observables, however, it is becoming increasingly important to predict properties beyond just the PES. Without access to the full spectrum of electronic and functional properties, ML falls short of being a complete replacement for electronic-structure calculations.

2.6.2. Current and future challenges

The most successful ML schemes share several common ideas. The use of translation and rotation invariant descriptors of the local configurations mimics the invariance of the potential to these symmetry operations. Furthermore, an additive decomposition of the energy in contributions from atom-centered environments, localized by a relatively short-ranged cutoff, enables transferability between different system sizes and improves greatly the data efficiency of the training step. This decomposition, although justifiable in light of the nearsightedness of electronic matter, undermines their ability to capture classical long-range effects such as electrostatic interactions and polarization phenomena, as well to describe non-local quantum effects such as dynamical electronic correlations.

Overall, it is still very difficult for ML models to replicate the ability of first-principles methods to predict properties beyond the potential. One major challenge is that the common wisdom that the community has developed to guide the construction of a good ML potential may not apply to other properties. Another, more fundamental issue is that several properties have structure beyond that of a rotationally-invariant scalar. Tensors and scalar fields, for instance, require a framework that reflects their covariance with respect to rotations and/or translations. Likewise, spectral properties such as the electron density of states (DOS) or the dielectric response (see also section 1.5) require simultaneous learning of multiple target observables, which also calls for a model that is adapted to the structure of the target data. Being able to predict properties with non-trivial geometric and algebraic nature opens the way to make better use of the ingredients of the electronic-structure calculation, either as a learning target or as an integral part of the learning architecture.

In addition to the theoretical hurdles, there are still many technical challenges still hindering widespread adoption of ML for general properties. A main priority is the development of software packages that treat all properties on an equal footing and allow fitting and predicting them in tandem with the potential. The generation of training data and the optimization of the computational cost of these calculations, are closely related issues that will require a concerted effort across the community.

2.6.3. Advances in science and technology to meet challenges

Many active research lines aim to close the gap between the capabilities of electronic structure calculations and their data-driven counterparts. The main strategy that they have in common is to adapt either the atomistic features that are used as input, or the mathematical structure of the model itself to reflect the underlying physics of the problem and the specific structure of the target property.

Rotationally covariant quantities, such as tensors and scalar fields, can be decomposed into a minimal basis of irreducible spherical tensors, which can be learned with a corresponding set of covariant structural features [126] or by building models endowed with a covariant architecture [127]. The efficient evaluation of equivariant features that describe high-orders of interatomic correlations [128], and can achieve remarkable levels of accuracy even using a simple linear model, is a promising research direction.

A second trend involves using electronic properties both as regression targets and as inputs—blurring the lines between electronic-structure calculations and data-driven models. Not only has it become possible to obtain accurate predictions of the ground-state electron density [129]: a new generation of orbital-free density-functional approaches has been proposed using the density as an input to learn accurate electronic

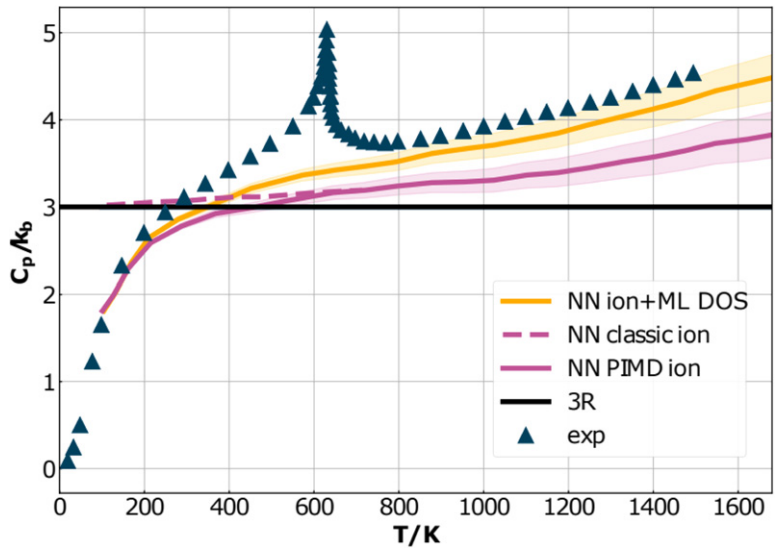

Figure 16. Prediction of the heat capacity of nickel by integrating a ML PES, to sample the nuclear motion, with a ML model of the electronic DOS. The model gives experimentally-accurate predictions by including both nuclear and electronic degrees of freedom. Note, however, that the heat capacity peak at the Curie temperature is not reproduced—showing that the ML model could be further improved by including magnetic effects. Figure reprinted with permission from [133], Copyright (2021) by the American Physical Society.

energies [130]. The data-driven treatment of other ingredients of an electronic structure calculation, such as the two-center Hamiltonian matrix elements [131] (see also section 3.1) or the functional ansatz for the wavefunction [132] (discussed in section 3.4), incorporates ML into methods that span the whole spectrum of quantum chemical techniques. These predictions are also useful to compute physical observables: the electronic contributions to the thermophysical properties of materials [133] (figure 16) and the electrostatic potential and the interaction between molecular fragments [129] can be easily obtained from the electronic DOS and the charge density.

To increase the accuracy of both potentials and property models, one can no longer avoid incorporating long-range physics. This can be done by using models with an explicit physical structure, e.g. by computing the electrostatic energy of the system based on the prediction of local charges and multipoles [134]. Such approaches have the advantage of including long-range electron correlation by virtue of enforcing the correct physics, as was recently shown in the case of the molecular dipole moment [135], or indeed by the charge equilibration approach discussed in section 2.4. An alternative strategy to reach the same goal involves including long-range correlations directly into the featurization. These models based on long-range features have the advantage of being more flexible, without restriction to any particular model or target property. For example, the multi-scale long-distance equivariants [136] use simultaneously an atom-density to describe the local structure and an artificial potential generated by it (figure 17) to capture non-local behaviour with an interpretable asymptotic limit.

Last but not least, software and data repositories must also be adapted to this new generation of integrated models, providing better interoperability with electronic-structure packages, efficient implementations of increasingly complicated featurizations and regression schemes, and standardized storage of properties such as electron density and wavefunctions.

2.6.4. Concluding remarks

ML models have made great strides in reproducing and predicting the thermodynamic properties of materials at finite-temperature by approximating and sampling the QM PES. Integrated schemes that predict *any* property accessible from electronic-structure calculations, and that unify ML predictions and physics-based steps, combine the best characteristics of the two approaches, further extending the reach of atomistic simulations. The fundamental challenge consists in finding the balance between the level of physical information that is incorporated directly in the model and the data-driven flexibility needed to capture unexpected effects. The description of long-range physics and of complex properties such as densities, tensors and matrix elements, provide compelling examples of the potential of generally applicable, physics-inspired, and mathematically sound ML schemes for atomic-scale modeling.

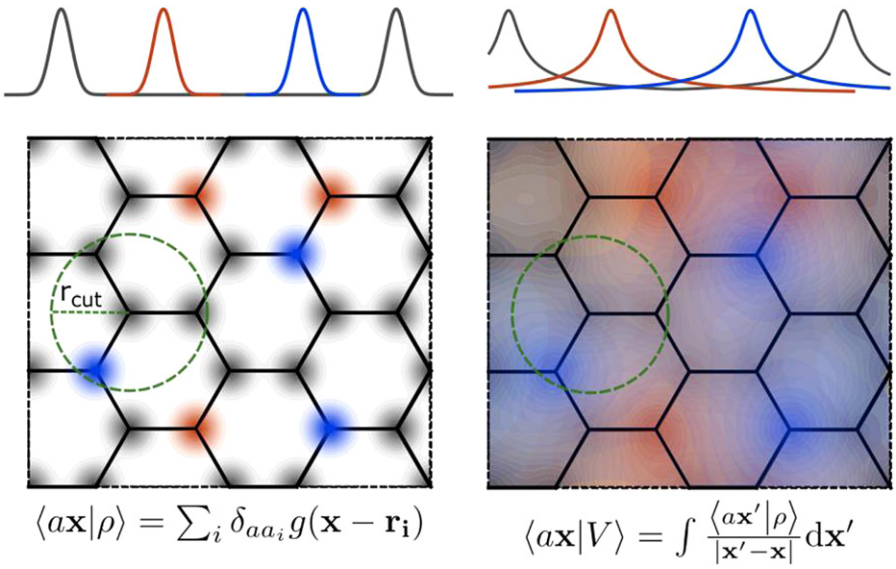

Figure 17. Construction of a long-range representation of an atomic structure. (Left) An atomic density is constructed by smearing atoms into Gaussian functions; the resulting density is expanded on a basis within a given localized cutoff about each atom, missing long-ranged interactions. (Right) The field to be expanded is instead constructed from the Coulomb potential generated by the (fictional) smeared atomic density, bringing true long-range information directly into the structural representation. Reproduced from [136] with permission of The Royal Society of Chemistry.

Acknowledgments

The authors would like to acknowledge support from the NCCR MARVEL, funded by the Swiss National Science Foundation (SNSF) (Grant Agreement ID 51NF40-182892).

3. Solving the many-body problem with machine learning

3.1. Unifying machine learning and electronic structure methods

Kristof T Schütt[1,2], Julia Westermayr[3], Michael Gastegger[1] and Reinhard J Maurer[3]

[1]Technische Universität Berlin, Germany

[2]Berlin Institute for the Foundations of Learning and Data, Germany

[3]University of Warwick, United Kingdom

3.1.1. Status

The prediction of PESs and chemical properties with ML has become an established procedure for accelerating electronic structure methods. While those models do not explicitly capture the electronic degrees of freedom of the system, there has been a recent surge of ML being used in all aspects of quantum chemistry, such as predicting the electron density, Hamiltonians and wavefunctions. These developments unlock the potential for unified approaches that use ML as an integral part of electronic structure methods [137].

Physical knowledge is increasingly being built into atomistic ML models. This includes not only fundamental constraints, such as rotational and translational symmetries and energy conservation, but also representations adopted from electronic structure methods. Examples for the latter include the use of Hartree–Fock molecular orbitals for the prediction of higher levels of theory [138]. Similarly, MP2 features have been used for the prediction of coupled-cluster amplitudes [139]. In this instance, ML provides a starting guess to accelerate convergence at the higher level of theory. Beyond that, incorporating physical regularities in ML models facilitates the representation of electronic structure. For example, the NN SchNOrb (figure 18) predicts Hamiltonians in local atomic orbital representations common to most quantum chemistry codes [83]. Thus, electronic structure data can serve as input to ML and the predictions can be fed back into quantum chemistry software. Density-functional tight-binding has been fused with ML to learn Hamiltonians [140, 141] and repulsive energy contributions with improved accuracy and transferability [142].

Finally, there have been several approaches to solve the SE using a NN representation of the wave function. For example, PauliNet [132] yields highly accurate correlation energies by using variational quantum Monte Carlo (QMC) in combination with the NN potential SchNet (see section 3.4). These examples demonstrate the potential of hybrid ML-electronic structure methods to not only accelerate but enable calculations at high accuracy, without requiring unattainable amounts of training data.

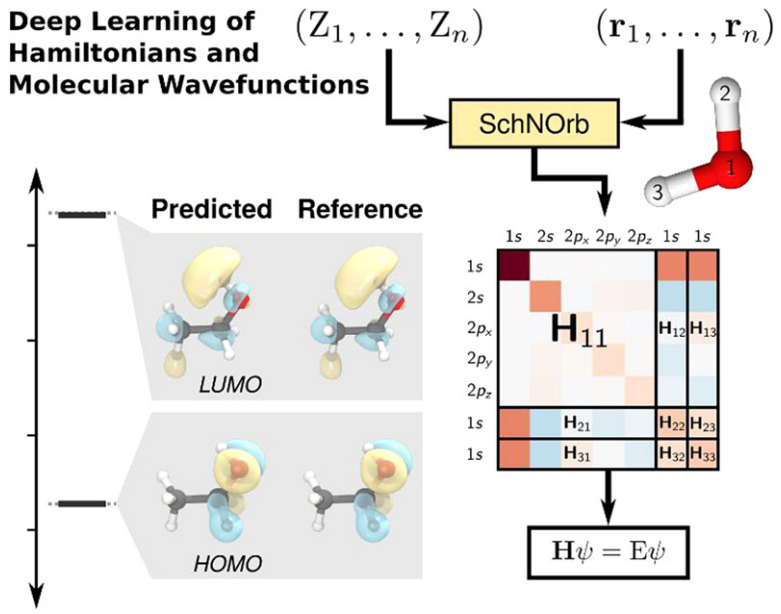

Figure 18. The SchNOrb NN predicts ground-state Hamiltonians in atomic orbital representations from atomic configurations. This can be used to predict molecular orbital energies, wave functions and to provide input to quantum chemistry software [83].

3.1.2. Current and future challenges

The development of ML-enhanced electronic structure methods requires a thorough understanding of the capabilities and limitations of both components for optimal symbiosis. In order for NNs to benefit from incorporating physical knowledge, it needs to be represented appropriately. For example, while nuclear charges appear to be helpful to characterize atom types, high-dimensional embeddings have turned out to be more effective in practice. Currently, the two major challenges of unifying ML and electronic structure methods are obtaining efficient ML representations of electronic structure and generating suitable reference data.

To overcome the main bottlenecks of electronic structure codes, ML surrogates must be able to represent central quantities such as electron density, wave functions and multi-centre/multi-electron integrals. Future ML representations of these quantities may depart from common existing basis representations due to the differing requirements of ML models compared to electronic structure theory [131]. For example, while calculated properties such as orbital or state energies can be non-smooth with respect to nuclear positions, this is highly problematic for ML approaches and requires careful consideration of representations that deliver smooth functions in configuration space. Other issues arise when properties are calculated from wavefunctions of different states. Since the wavefunction is only determined up to an arbitrary phase, sign changes need to be controlled against a reference [143].

Another challenge is the prohibitive cost of computations to generate accurate reference data. In such cases, divide and conquer approaches need to be used, partitioning the system using, e.g., a combination of different levels of theory in a multi-scale strategy or by breaking the large system down into more manageable fragments. Local atom or fragment-centred representations of electronic structure synergize well with such schemes and offer the possibility of transferability and linear scaling with system size. However, it is not yet clear how they can account for long-range, collective, or symmetry effects on system scales not accessible to the reference method. The latter aspect is particularly problematic for the representations of systems that exhibit different electronic and spin configurations, for example transition metal complexes or excited states [143].

3.1.3. Advances in science and technology to meet challenges

A necessary condition for the deep integration of ML and electronic structure codes is the definition of fast, compatible interfaces with formalised communication and data standards. In this context, the modularization of quantum chemistry packages is a recent development that needs to be further pursued (figure 19) [137]. While quantum codes are often implemented in Fortran, deep learning frameworks such as PyTorch or TensorFlow rely on C++ backends and the flexibility of Python. To achieve complete integration, automatically differentiable electronic structure codes will be required for end-to-end optimization of ML components.

A promising route towards more transferable and scalable electronic structure codes is their combination with basis representations predicted by ML. Those can be constructed either by pre-processing electronic structure reference data [131, 138], or by performing integrated representation learning. Previous design choices need to be revisited and explored in combination with ML. This relates particularly to the trade-off between

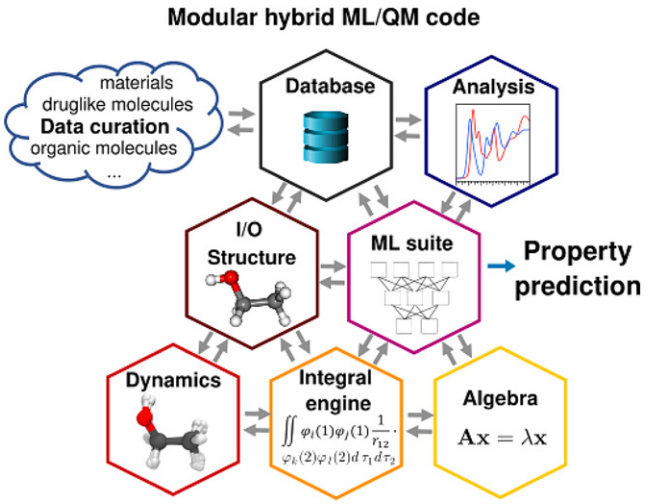

Figure 19. Schematic layout of highly modular electronic structure codes that are deeply integrated with elements of ML. Reproduced from [137]. CC BY 4.0.

basis set size and the complexity of interaction integrals. Decisions that would have previously been discarded due to computational infeasibility may be enabled by ML and lead to faster, more accurate solutions. New ML methods must be able to better deal with the non-smoothness of certain properties, such as excited-state potentials, or non-unique properties with arbitrary phase. These steps may eliminate the need for manual pre-processing and enable the seamless integration of ML algorithms into electronic structure code.

For the efficient acquisition of reference data, further improvements are required in so-called life-long learning, i.e., the continued training of models, which can be prone to overfitting or forgetting of previously acquired knowledge. Ideally, data acquisition happens transparently, i.e., explicit calculations are carried out automatically when the training domain of the ML component is left. This necessitates further development of methods for fast and reliable error estimates. Finally, the whole procedure should be integrated with global data repositories, to optimally use computing resources. To overcome system size restrictions of the electronic structure reference, new multi-scale approaches must be developed, e.g., by embedding local ML representations into physically motivated global frameworks. Here, expertise from established fragmentation methods and multi-scale strategies will prove invaluable.

3.1.4. Concluding remarks

Recent developments have shown that ML methods are able to model fundamental properties of electronic structure, either in real-space or in abstract basis representations. The challenges for the coming years relate to finding physically meaningful representations within ML models that can break through the scaling limitations of conventional electronic structure theory. To achieve this, ML methods will need to go beyond the reference electronic structure basis by constructing efficient representations that correctly capture physical boundary conditions while retaining favourable computational scaling properties. Finally, fast and transferable unified approaches need to incorporate physical constraints into ML models while staying flexible enough to learn expressive representations from data.

Acknowledgments

This work was funded by the Austrian Science Fund (FWF) [J 4522-N] (JW), the Federal Ministry of Education and Research (BMBF) for the Berlin Center for Machine Learning/BIFOLD (01IS18037A) (KTS), and the UKRI Future Leaders Fellowship programme (MR/S016023/1) (RJM). MG works at the BASLEARN-TU Berlin/BASF Joint Lab for Machine Learning, co-financed by TU Berlin and BASF SE.

3.2. Using machine learning to find new density functionals

Bhupalee Kalita[1] and Kieron Burke[1,2]

[1]University of California

[2]University of California

3.2.1. Status

DFT has provided low-cost alternatives to direct solution of the SE for almost a century [144]. The Kohn–Sham (KS) scheme [145], in which only the XC energy needs to be approximated as a functional of the density, has greatly improved accuracy while maintaining low computational cost. Today, about 30% of supercomputer

use is devoted to solving these equations, but there are hundreds of different human-designed XC approximations in use, each producing different predictions. Almost all begin using the density and its gradient (semilocal approximations). Materials science is dominated by simple standard functionals, often designed using exact conditions, while chemistry mostly uses approximations designed only for molecules, but often achieving higher accuracy.

Four prominent limitations come to mind. Most DFT calculations are for weakly correlated systems, and there is tremendous desire to improve their accuracy without significant computational cost. Second, DFT has well-known generic failures, such as self-interaction error or poor energetics for strongly correlated systems, such as a stretched H_2 molecule [146]. Most XC approximation fail to produce a realistic binding energy curve without breaking spin. As one goes from two atoms to four and many, the difficulties grow and can be related to the failure of DFT approximations to capture Mott–Hubbard physics [147]. Third, theorems prove that, in principle, one can avoid solving the KS equations if one has a sufficiently accurate approximation for the KS kinetic energy, but here the limitations are even greater, due to the need to extract accurate densities and total energies. Finally, ground-state DFT yields only ground-state energies and densities, but there is also tremendous need to predict response properties. Here we focus only on the ground state.

ML has already helped with functional development. In prescient work a quarter-century ago, Tozer *et al* [148] found a semilocal approximation by training a NN to optimize a fit to KS potentials. Moreover, Bayesian methods were used to analyse DFT errors in 2005 [149]. More recently, Snyder *et al* [150] used KRR combined with a principal component analysis of training densities, to create a KS kinetic energy functional reaching chemical accuracy, albeit in a very simple model. And Nagai *et al* [151] showed that, by training a NN on both the densities and energies of just a few molecules, one could create semilocal approximations comparable in accuracy to those of humans and generalizing to a broad range of molecules.

The field of using ML to design functionals is in its infancy, and improvements in speed, accuracy, and applicability of DFT are beckoning. Any such improvements that can be implemented in standard codes will have enormous scientific impact.

3.2.2. Current and future challenges

ML is promising for improving DFA to overcome the limitations listed above, and progress is likely in all three areas.

First, there are many ingredients already in use for making XC approximations, including dispersion corrections, fractions of exact exchange (both global and range separated), random phase approximation, etc. Can ML be used to find the 'best' combination of these ingredients? More fundamentally, how do we define 'best'?

Second, ML allows the possibility of constructing completely non-local functionals, using information about the density at every point in space, either with KRR or NN's. This can be used to find the exact functional for strongly correlated systems, as in reference [152], especially if full differentiable programming techniques are used. Here, by using the KS equations as a regularizer, a full dissociation curve for (one-dimensional) H_2 was constructed from just two data points alone, suggesting tremendous potential for generalizability. However, such a functional, defined on the whole R-space, cannot be applied to arbitrary systems, so what features must be included to make it work more generally?

Third, ML can produce pure density functionals, which could bypass the need to solve the KS equations. This was demonstrated for small molecules, producing an ML functional that yields accurate densities and energies for malonaldehyde and resorcinol MD simulations [153], and for water in condensed phase [154]. But, as above, such functionals cannot be expected to generalize well, and so must be retrained for each new species, unlike standard DFT.

In figure 20, we employ the KS regularizer (KSR) from reference [152] to calculate the binding curve of 1D H_3 and show its attributes at $R = 4$ Bohr. The KSR is chemically accurate even when the bond is stretched and predicts the density with negligible error. A recent study [155] provides an example of implementation of differentiable DFT in 3D. A similar extension of the work in reference [152] can effectively provide a stable solution for strongly correlated matter. However, much work remains to test these algorithms to answer what would be the degree of generalization and what could be done to improve them further.

3.2.3. Advances in science and technology to meet challenges

ML has revolutionized many aspects of everyday life, from movie selection to facial recognition. Over the past 10–15 years, there have been significant attempts to use it in physical sciences and especially in electronic structure theory. The most notable success has been the development of force fields, both in chemical and configuration space [156].

But the development of DFA is still a black art, requiring an unholy alliance of physical (or chemical) intuition, deep knowledge of theory, and some very carefully chosen data. A major difficulty is to build ML

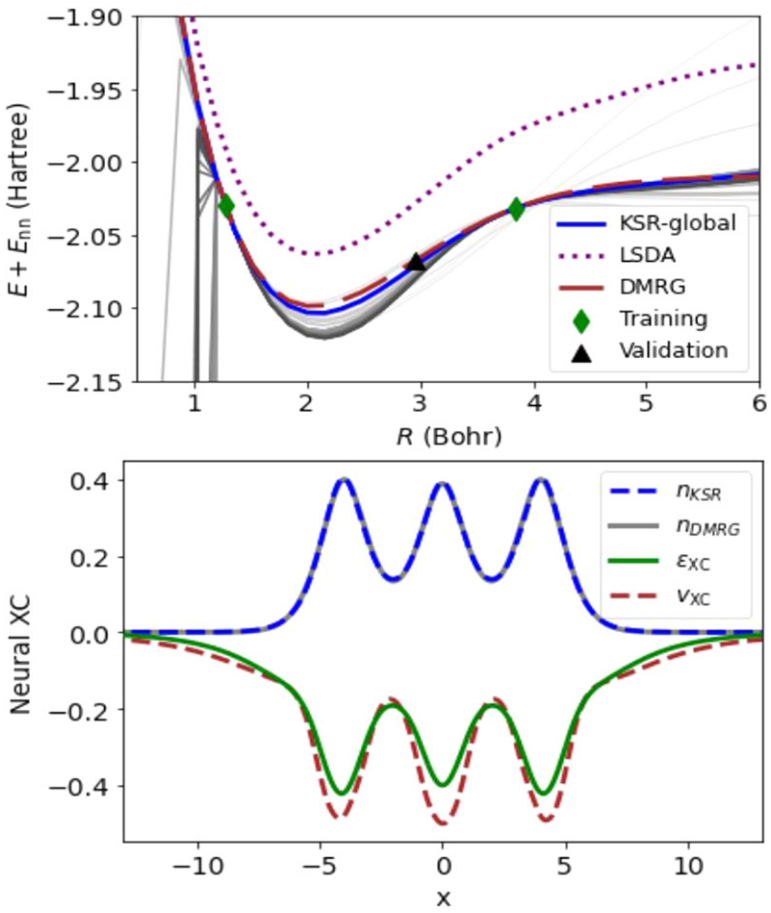

Figure 20. (a) One dimensional H_3 dissociation energy curve created with the KSR-global function from reference [152]. This model was trained with just two configurations. The changes in predictions as the model evolves from underfitted to overfitted are shown by the darkening shades of grey. The optimal parameters, determined from a single validation configuration, yield the chemically accurate blue curve. E_{nn} is the nuclear interaction energy, density matrix renormalization group (DMRG) is essentially exact, and LSDA is the result of the local spin density approximation. (b) The density, XC energy and XC potential of H_3 at 4 Bohr, calculated using the optimal parameters.

models that respect all the implicit (and explicit) rules in DFT that humans know (often only intuitively) so that the models extrapolate appropriately to new materials and new molecules. With our traditional XC approximations, when we run a KS calculation on an entirely new problem, we have a strong sense of how accurate we expect it to be, and certain intuitive consistency tests, such as trying a different functional, even if we cannot put quantitative error bars on our predictions. If we can use ML to design better functionals, overcoming any of the three challenges mentioned previously, such ML-designed functionals will permanently alter the computational landscape.

Much has been said and written about the potential for quantum computers to transform electronic structure calculations. It is certainly true that, once a sufficiently large error-correcting machine is widely available, there are several strongly correlated problems that they might solve for us. But unless there are extreme speedups in routine classical computations, DFT will long continue as the workhorse for the 99% of problems (or aspects of these problems) that DFT works well for.

3.2.4. Concluding remarks

The applications of ML to functional design are still in their infancy. There is no general-purpose XC approximation designed by ML in use or available in most codes. It will take more effort and research to understand what the best way is to apply ML techniques (likely NN's) to develop better approximations, including ones that can be systematically improved with increases in training data. ML could produce either faster or more accurate functionals for present applications or extend the reach of practical DFT calculations to encompass strongly correlated systems. The future looks bright but has not arrived yet.

Acknowledgments

KB acknowledges NSF Grant No. CHE 1856165 and BK acknowledges NSF Grant No. DGE 1633631.

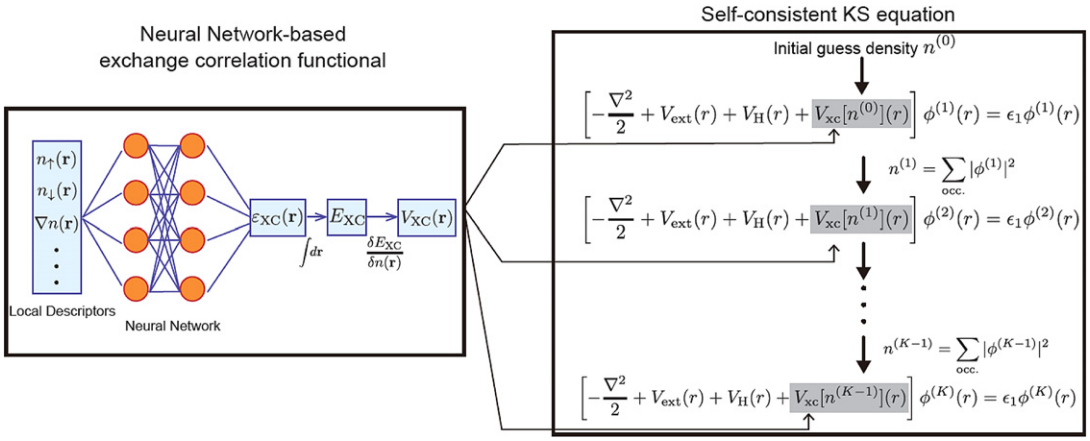

Figure 21. Schematic figures of ML XC functional (left) and its usage in KS equation (right). The same ML projection is repeatedly substituted for the self-consistent KS cycle. Reproduced from [151]. CC BY 4.0.

3.3. Machine learning Kohn–Sham exchange–correlation potentials

R Nagai[1,2], R Akashi[2] and O Sugino[1,2]
[1]The University of Tokyo
[2]The University of Tokyo

3.3.1. Status

The method of KS DFT [145] is the current standard for the first principles calculation of electronic structures for its reasonable computational cost. Improvement of the XC functional, which governs the accuracy of results, has long been a challenging issue. Being a functional of the distribution of charge density, its exact form is beyond calculable expression. Its practical approximate forms have so far been constructed using the physical conditions, such as asymptotic behavior and scaling relations, as a guiding principle, but the accuracy has still been a problem [157, 158]. This is due to the lack of physical conditions needed for the construction of complex functionals beyond the local or semilocal ones and is also due to difficulty in analytically interpolating the asymptotes inside the nonuniform density region typical of real materials. There is much room for improvement if one can directly refer to the nonuniform density and use a powerful interpolation scheme for functional development.

The ML scheme is expected to overcome this difficulty. By using the extremely flexible ML model for the XC functional and tuning the model parameters to reproduce the density-to-energy relation of real materials, one can obtain a computable form that applies to the nonuniform density cases. The data to be reproduced (training data) are, for example, those from the accurate and costly calculation. The modern ML approach to DFT has been first initiated for an orbital-free formalism [150], whereas it has later been extensively applied to the more familiar KS method [151, 152, 159, 160] (figure 21). In the latter case, the use of the KEO has been shown to suppress numerical instability when applied to systems not included in the training data [159]: we here focus on this.

3.3.2. Current and future challenges

The ML models such as NN in principle enable us to implement the fully nonlocal exact functional with arbitrary accuracy, but in practice, we are faced with obstacles. To train the XC potential applicable to real three-dimensional materials, one may use electronic structure data (energy and density) of representative systems such as molecules and solids. However, gathering accurate data is a demanding task. Precise experimental observation of electronic density is still difficult, and therefore one has to employ theoretical calculations for generating the training data. Since our goal is to improve accuracy over the existing functionals, the training data should be generated from methods that are more accurate than those of the standard functionals. Some wavefunction theory methods such as the coupled-cluster and QMC methods meet this requirement at an affordable computational cost. However, the cost of their application to large systems like solids becomes yet formidable.

Even if an accurate electronic structure dataset is obtained, another class of difficulty arises in designing the training scheme. Usual supervised learning uses pair of input and output data for the training. For the training of XC potential, the ‘input’ and ‘output’ correspond to the density and XC potential, respectively. To prepare the training XC potential with the training electronic density, we need to solve the numerically difficult inverse KS problem [161]. Furthermore, the XC potential thus obtained has a difficult property to treat in the training.

Since it is close to the exact one, the value of the potential at any spatial point is dependent on the whole density distribution; the exact XC hole is fully nonlocal. Though training of the fully nonlocal ML potential has been demonstrated in 1D model systems [159], such a form is not applicable to other systems; when the sizes of target systems are different, the density distribution cannot be input into the same trained ML model. Transferrable design of (semi)local XC potential is thus desirable. Moreover, determining the nonlocal functional with many parameters requires a large amount of training data.

3.3.3. Advances in science and technology to meet challenges

With the limited availability of the training dataset, a recent study [151] have demonstrated an efficient way to train the ML functional. It showed that the semilocal form trained with accurate electronic density data in a few molecules can yield practical accuracy for various molecules. This transferability is due to the fact that the electronic state at a point is mainly affected by those within a short-range. By limiting the functional form to semilocal, every spatial point gives different density-potential relation. A large amount of data is, especially in 3D, thus available even in a single molecule, thanks to which the training of the huge number of model parameters is enabled. Furthermore, the kinetic operator term in the KS equation is shown to suppress the error coming from the non-smooth shape of the trained ML functional [159]. Utilizing those properties enables efficient extraction of the essential properties of the XC functional from a limited number of molecules, which is computationally desirable.

In the same study [151], a novel approach has been initiated which avoids the problem of getting the training XC potential data. There, each training iteration consists of the whole solving procedure of the KS equation, and the obtained energy and density are compared with the reference density data. The ML parameters are optimized to decrease the loss function defined by the density and energy. In this procedure, the XC potential is not directly referred to. Optimization was executed with the simulated annealing, which is basically a random walk and does not use a gradient of the loss function. Later, Li *et al* have implemented the solution of the KS equation itself as a differentiable program, where back-propagation of the loss function, i.e. differentiation of the KS procedure, is implemented [152]. The indirect training of XC potential with the density paves a feasible way to improve ML density functionals exceeding conventional ones.

3.3.4. Concluding remarks

The ML methods open a novel approach for constructing the XC potential referring to the realistic density regime, in contrast to the conventional functional construction that refers to the asymptotes. For further improvement, there are challenges to overcome: efficient collection of electronic structure data and development of effective training methods. To overcome them, insight into the Frontier technology of ML is important as well as accumulated knowledge of materials science. The semilocal property of electronic effect, the kinetic operator as 'regulator [159]', and integration of the KS equation into the training can be listed as effective ways for further advances toward the functional with ultimate accuracy. As DFT has advanced as a general framework for studying interacting many-body systems, it will be fruitful to exchange knowledge independently developed in other fields such as classical particles [162].

Acknowledgments

This work was supported by KAKENHI Grant No. JP20J20845 from Japan Society for the Promotion of Science. Part of the calculation was performed at the Supercomputer System B and C at the Institute for Solid State Physics, the University of Tokyo.

3.4. Deep-learning quantum Monte Carlo for molecules

Jan Hermann[1] Frank Noé[1,2,3]

[1]FU Berlin, Germany

[2]FU Berlin, Germany

[3]Rice University, United States of America

3.4.1. Status

Most chemical and physical properties of molecules and materials are accurately described by the nonrelativistic and time-stationary electronic SE. As the computational cost of its exact solutions increases exponentially with the number of electrons, N, the main challenge is finding approximations that strike a good balance between accuracy and computational cost.

The past decade has seen a surge in ML approaches to predict the outcome of quantum chemistry calculations by training kernel machines or NNs on datasets of molecules (see sections 2.1–2.6). In contrast, this section focuses on *ab initio* ML, which aims to find the solution of a problem specified by self-consistency relations or a variational formulation.

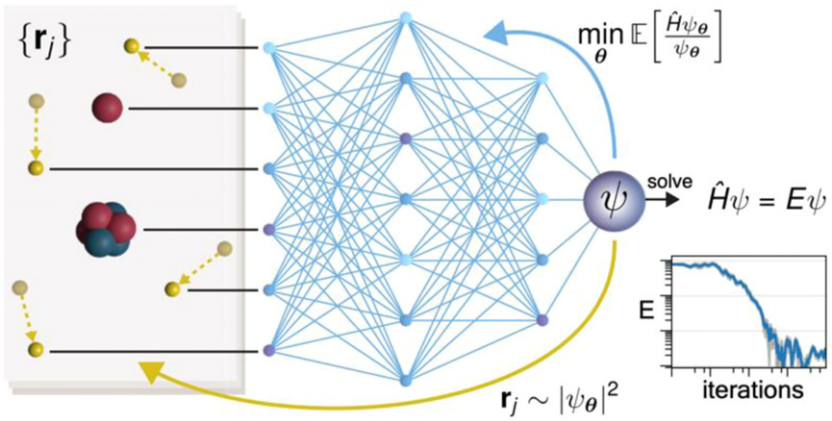

Figure 22. Deep QMC solves the SE by alternating between optimizing the parameters of the NN ansatz via the variational principle (blue) and sampling from the square of the wave function to generate data (yellow). Reprinted by permission from Springer Nature Customer Service Centre GmbH: [Springer Nature][Nature Chemistry] [166] (2020).

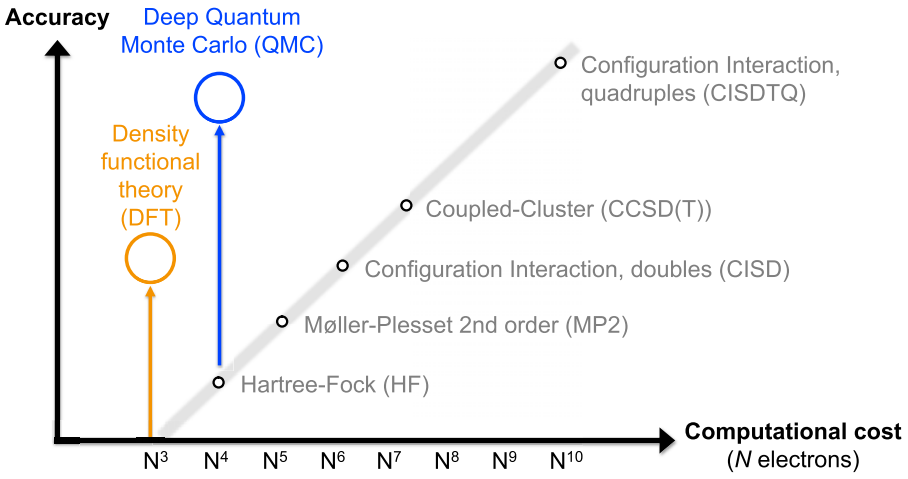

Figure 23. Pople diagram of quantum chemistry with DFT and deep QMC shown as offering exceptional compromise between accuracy and cost.

In a seminal paper, Carleo and Troyer [163] made a connection between ML and quantum mechanics by modeling wave functions as NNs and interpreting the energy expectation value as a ML loss function, which is minimized according to the variational principle of the SE. By self-consistently sampling from the square of the wave function with QMC, this approach generates its own data on the fly (figure 22). The initial applications of the novel approach targeted spin systems on lattices. Generalizations to electrons and to real space followed shortly after, starting with electrons on lattices as in the Hubbard model [164], through the first exploratory work for electrons in real space [165].

Recently, PauliNet [166] and FermiNet [167] have been proposed as two highly accurate yet affordable deep-learning architectures that solve the SE using antisymmetric NNs,

$$\psi\left(\mathbf{r}_1,\ldots,\mathbf{r}_N\right) := \sum_k \det\left[\phi_i^k\left(\mathbf{h}_j(\mathbf{r}_1,\ldots,\mathbf{r}_N)\right)\varphi_i^k(\mathbf{r}_j)\right]$$

where $\mathbf{h}_j$ is an output of a permutation-equivariant NN for the jth electron, while ϕ_i^k and φ_i^k are many- and one-electron functions, respectively, for the ith generalized orbital in a kth Slater determinant. PauliNet and FermiNet differ in how $\mathbf{h}_j$, ϕ_i^k, and φ_i^k are constructed. In certain characteristics they both outperform well-established methods, such as the coupled-cluster method, with a computational cost that scales only as N^3 to N^4. While the exponential scaling of the exact solution of the electronic SE is fundamental and will dominate at large N, the aim of deep QMC is to push the onset of this exponential scaling to large enough N so as to offer an electronic structure method for intermediate-sized atomic systems with a few hundreds of electrons with unprecedented compromise between accuracy and computational cost (figure 23).

3.4.2. Current and future challenges

To date, the largest systems to which deep QMC has been applied have around 30 electrons, and the excellent results obtained so far provide no indication that it cannot be scaled up. This leaves an order of magnitude in system size to be bridged in the near future for deep QMC to become highly practical. Besides the sheer number of electrons, the higher nuclear numbers and the associated difficulties that plague all QMC approaches will present another obstacle in going to complex chemical systems such as transition-metal complexes.

The accuracy of any polynomial electronic structure method must inevitably deteriorate with increasing system size, owing to the computational complexity of the electronic many-body problem. This is, however, a theoretical asymptotic consideration—what matters in practice is what is the onset and the rate of this deterioration for relevant system sizes. While full characterization of this accuracy decay remains unresolved even for well-established methods [168], their modes of failure are well-known. Although first results of this kind for deep QMC have been already published [169], most of the work in this area remains to be done.

By far the most applications of standard QMC are not to molecules but to solids, because unlike for molecules, there is essentially no other electronic structure method practical for solids that would match the accuracy of QMC. Being an explicitly many-body method, QMC uses a supercell approach to treat periodic systems, which again translates to the demand on treating larger numbers of electrons, but extending deep QMC to solids also presents other challenges, such as the ability of the NNs to capture long-range electron interactions.

Does deep QMC have anything more to offer than just highly accurate variational QMC? While this achievement would already make deep QMC a worthwhile endeavor, one can naturally ask whether the use of deep NNs in QMC might open entirely new avenues that would be simply impossible without them. For instance, deep autoregressive models for quantum lattice systems avoid the need to run lengthy Markov-chain Monte Carlo simulations by directly generating samples of electron configurations [170], and it remains to be seen whether such an approach can be transferred to electronic real-space systems.

3.4.3. Advances in science and technology to meet challenges

Standard QMC is a well-established electronic structure method that has been implemented in several mature, high-performance software codes, which use numerous advanced techniques to make the calculations more efficient. Pseudopotentials enable efficient treatment of heavier atoms. Diffusion Monte Carlo substantially increases the accuracy of the energy that can be obtained from a given, already optimized ansatz. Such techniques, and many others, can be transferred to deep QMC, while carefully considering how is their cost-benefit ratio changed by the use of NNs in the ansatzes.

On the deep learning side, novel NN architectures should be considered, such as graph networks whose convolutions incorporate not only distance but also angular information. Another important aspect is to explore where the optimum lies in constructing the ansatz between large-scale deep learning architectures (such as FermiNet) and architectures that incorporate more physical knowledge (such as PauliNet). The evaluation of the Laplacian of the wave function in the kinetic energy term is by far the biggest computational bottleneck in deep QMC, but automatic differentiation in popular deep learning frameworks has not been optimized with such terms in mind, so advances in this area could yield high gains in efficiency. Finally, we must develop benchmarks to test the performance of the learning system in a practically relevant manner—while the pioneering works used the variational (absolute) energies as benchmarks, these are less informative for large molecules where energy differences matter more.

The real-space formulation of the electronic SE, as used in deep QMC, is also referred to as first quantization. Second quantization is an alternative formulation that uses one-electron basis sets to transform this differential equation into an algebraic problem, and is the foundation of quantum chemistry. Second quantization has been also subjected to improvement via NNs [171], which opens the question of marrying the two alternative approaches, which might result in a more robust method that would combine the advantages of both.

Using deep NNs for the sampling part of QMC would require entirely new architectures, that would be able to encode antisymmetry in an autoregressive fashion. At present, this seems beyond the reach of existing techniques in deep learning.

3.4.4. Concluding remarks

Standard electronic structure methods have been developed for decades to get to the point where they are now, with numerous methodological and computational techniques and tricks under their belt. Deep QMC has only been recently developed, yet it is already competing with those standard methods. This suggests a promising future for this novel approach once it receives more attention and efficient codes are developed. We believe that deep QMC will provide chemists and materials scientists with a new powerful computational tool.

Acknowledgments

We acknowledge funding and support from the European Research Commission (Grant No. ERC CoG 772230), the Berlin Mathematics Research Center MATH+ (Project Nos. AA2-8, EF1-2, and AA2-22), and the German Ministry of Education and Research (Grant No. 01IS18037J, BIFOLD–BZML).

3.5. Disordered quantum systems

Sebastiano Pilati

University of Camerino and INFN-Sezione di Perugia

3.5.1. Status

In textbooks on condensed-matter physics, solid-state systems are usually described as clean and periodic structures. On the contrary, disorder is ubiquitous in real materials, and it drastically affects the system properties. It induces consequential phenomena such as Anderson localization, which causes insulating behavior even when the band is not full, in contrast to textbook band-structure theory. Disordered systems may also undergo so-called many-body localization, meaning that in isolation they fail to reach thermodynamic equilibrium. As a consequence, even the basic assumptions of (equilibrium) statistical mechanics are not applicable. For these reasons, disorder challenges the conventional top–down approach to condensed-matter theory, whereby material properties are predicted from fundamental equations and basic principles. On the other hand, disordered systems may lend themselves to a data-driven approach, whereby bottom-up inference is performed by identifying regular patterns from large experimental or synthetic (i.e., computer generated) datasets. In fact, when disorder is included in the theoretical modelling, many random realizations of the same underlying model have to be considered. In conventional approaches, observable properties are predicted from averages over these random datasets. In recent years, physicists started exploring the use of ML techniques to exploit such datasets more fruitfully. The goal is twofold: on the one hand, researchers aim to develop computational techniques more adequate for disordered systems than conventional top–down theories. On the other hand, they are investigating whether randomness can be introduced on purpose to enable the utilization of ML techniques in the domain of condensed-matter theory. For example, deep NNs have been trained to predict the ground-state energies of quantum particles in disordered potentials [172]. After being trained on datasets including many random realizations, the networks provided accurate predictions for previously unseen instances, bypassing the direct solution of the SE. Such studies pave the way to different strategies to solve both clean and disordered quantum many-body problems, but they also present researchers with new challenges that need to be addressed.

3.5.2. Current and future challenges

Disordered models are being used as a challenging testbed to evaluate the performance of ML techniques in solving condensed-matter problems. In reference [172], supervised learning was used to map disordered potentials formed by randomly placed Gaussians to the corresponding ground-state energies. The adopted convolutional NN demonstrated capable of automatically extracting the relevant features, avoiding recurse to ad-hoc engineered descriptors. Reference [173] performed an analogous mapping, considering as input a model for ultracold atoms in disordered optical speckle fields. Since these two studies addressed non interacting particles, sufficiently copious training databases (order of 10^5 instances or more) could be generated at an affordable computational cost. Addressing interacting many-body systems is more challenging, since producing so many training instances is impractical. While this problem is encountered in most applications of supervised learning to quantum many-body systems, it is particularly relevant for disordered systems, since the number of descriptors required for their characterization has to scale with the system size and, as a consequence, larger datasets are needed for training.

NNs are also being used to implement compact representations of many-body wave-functions [174] (see section 3.4). A recent application to disordered quantum spin models highlighted the need of further exploring which network connectivity (e.g., all-to-all versus sparse) is optimal for quantum many-body problems [175].

ML algorithms are being applied also to DFT (see sections 3.2 and 3.3). Reference [150] employed KRR to reconstruct the kinetic-energy functional of Hamiltonians with randomly placed Gaussians, opening new paths to build orbital-free theories. The appealing feature of KRR is that training requires only order of 100 instances. However, accurately computing the functional derivatives is a challenging problem [150], still under intense investigation [130]. Also convolutional NNs have been applied to DFT data for disordered models [176], showing that they allow bypassing the KS scheme.

Interacting many-body systems have been addressed within DFT in references [177, 178], considering one-dimensional lattice models with on-site disorder. The training sets were produced via exact-diagonalization and via DMRG techniques. However, addressing realistic higher-dimensional models still represents a challenging task, due to the cost of creating suitable training databases.

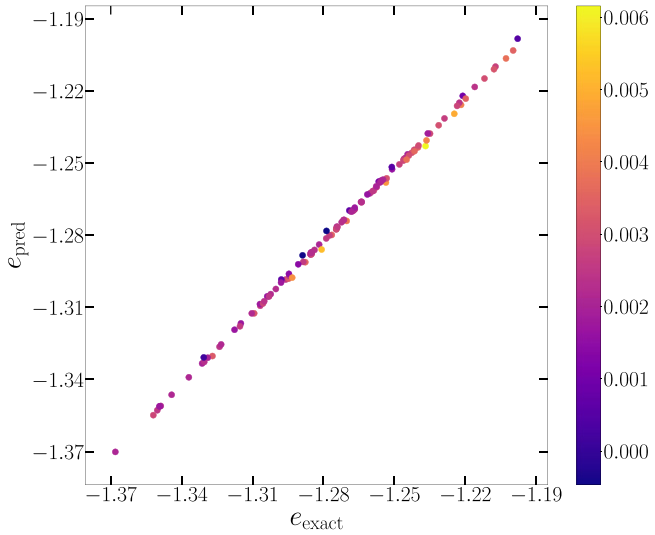

Figure 24. TL for a disordered quantum Ising chain with nearest-neighbour and far-neighbour (10 sites) couplings. The ground-state energy per spin e_{pred} predicted by a scalable NN is shown as a function of the exact value e_{exact}. The network is trained on small chains with $N = 15$ spins, which can be solved via exact diagonalization. It is tested on larger chains with $N = 50$ spins, which are simulated via unbiased QMC simulations. The random couplings are uniformly sampled in the interval $[0, J]$, the transverse-field intensity is $\Gamma \cong 0.36\,792J$, and the energy unit is the maximum coupling J. Reprinted (figure) with permission from [180], Copyright (2020) by the American Physical Society.

3.5.3. Advances in science and technology to meet challenges

The supervised training of the currently available NNs for disordered quantum systems requires massive datasets. This is a critical problem, since this many instances can be generated only for small quantum systems, unless uncontrolled approximations are accepted. Most network architecture adopted so far have been adapted from those already is use in the fields of image analysis and speech recognition. Hopefully, novel architectures specifically designed for quantum matter will reach superior learning speeds.

TL is emerging as an alternative strategy to accelerate the training process. In the field of image analysis, it is standard practice to adopt pre-trained networks, previously optimized on generic databases, to then transfer the learned parameters to more specific classification tasks. An analogous strategy has recently been applied also to quantum matter. Specifically, it has been used to transfer knowledge from small to larger system sizes. For example, reference [179] introduced size-extensivity by combining identical parallel networks, each one addressing a small tile of the whole input system. A small overlap between adjacent tiles was allowed to account for spatial correlations. In reference [180], size scalability has been implemented by including global pooling layers in a convolutional model. Reference [181] considered an ad-hoc descriptor for the particle number, allowing the network operating with heterogeneous datasets including different densities. Notably, these architectures have also been tested in extrapolations tasks, i.e., in making predictions for system sizes larger than those included in the training set. The obtained results for a disordered quantum Ising chain are shown in figure 24. However, the regime of validity of these extrapolation techniques needs to be further investigated, especially in the presence of long-range or frustrated interactions [180].

Since massive training sets will be required to develop novel network architectures, the community will benefit if they will be shared in public repositories. It is also worth mentioning that the training with (slightly) noisy data has recently been tested, showing that the prediction accuracy does not dramatically degrade [173]. This led to the speculation that, in the future, training sets could be produced using (inevitably noisy) cold-atom quantum simulators or other quantum devices.

3.5.4. Concluding remarks

Disordered quantum systems are proving to be particularly suitable for data-driven approaches based on ML algorithms. Various successful studies focusing on paradigmatic testbed models have been performed, and promising applications also to more complex electronic systems have recently been reported [179, 182]. The key enabling factor is the possibility to generate copious datasets of random realizations. Still, various challenges need to be addressed. The learning speed of NNs must be increased by designing architectures specifically tailored to quantum systems. Furthermore, massive databases need to be generated and shared among researchers in the field. Certain TL protocols have already been used to accelerate the learning process, but the applicability of pre-trained models to different kinds of disorder must be further explored. In the long

term, one can envision the use of quantum simulators as generators of suitable training datasets for intractable quantum many-body systems.

Acknowledgments

SP acknowledges fruitful discussions with S Cantori, B Juliá-Díaz, À Martínez Miguel, P Mujal, A Perali, N Saraceni. Financial support from the FAR2018 project titled 'Supervised machine learning for quantum matter and computational docking' of the University of Camerino and from the Italian MIUR under the Project PRIN2017 CEnTraL 20172H2SC4 is acknowledged. SP also acknowledges the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support.

4. Big data for machine learning

4.1. Challenges and perspectives for interoperability and reuse of heterogenous data collections

Claudia Draxl, Martin Kuban, Santiago Rigamonti and Markus Scheidgen

Humboldt-Universität zu Berlin (HU Berlin)

4.1.1. Status

Many or most of our colleagues may now agree that data-centric approaches will complement and change the way how research is currently performed. As a matter of fact, we are experiencing an atmosphere of departure in several aspects. Data-analytics and ML approaches are being developed and applied to various problems, and high-throughput screening is going hand in hand with the establishment of small- and large-scale data collections. The NOMAD Laboratory [183] is quite orthogonal to all of them as it was never dedicated to a particular research topic or material class, but rather aimed at being an open platform for sharing data within the entire community. It allows users to upload computational results from all major electronic-structure codes, now hosting (as by March 2021) more than 100 mio. calculations from individual researchers as well as from other databases (see reference [183] for details).

Such a huge and findable, accessible, interoperable, reusable (FAIR) [184] data repository is a wonderful playground for being explored in view of (i) comparing the performance of different methodologies for one and the same material, (ii) finding trends in the data, e.g., by unsupervised learning, or (iii) using the data pool for developing and applying novel AI tools. As such, we can consider these data as a gold mine of the 21st century. Turning it into gold, however, can only be realized if we fully control and understand this raw material. Among the four characters of FAIR, the *I* (interoperability) is the most critical and largely unresolved issue when bringing together data from different sources. So far, in contrast to quantum chemistry, there exist only a few efforts geared towards reproducibility [185] and benchmarking [186, 187] which hampers the assessment of data quality. Even more critical, on the experimental side, the situation is much worse. Needless to say, a balanced picture, where experimental and theoretical characterization of materials go hand in hand with each other, will be crucial for realizing the 4th paradigm of materials research.

4.1.2. Current and future challenges

As mentioned above, interoperability may be the biggest obstacle, hampering the wider usage of inhomogeneous data. In fact, the most innovative AI method is of little value if data can be mis-interpreted because their quality is either not known or not considered. Thus, our future research not only concerns powerful AI tools but also in-depth analysis and understanding of the data.

How can we control data and assess their quality? Staying within the realm of computational *ab initio* results, the major content of the NOMAD Laboratory, ground-state calculations—in particular the energetics of materials—are likely to be controlled first. Here, first examples of data assessment [188] and error estimates [189, 190] are underway. Excited states, in turn, are a true challenge. As an example of the complexity we mention the *GW* approach of many-body perturbation theory which is sketched in figure 25. Arguably, only experts who have enough insight into the implemented algorithms and approximations, are able to fully judge the quality of the output of such computationally heavy calculations. To address just a few aspects: on the technical side, we may need an auxiliary basis set coming with various parameters. Likewise, there are various ways for doing the analytical continuation of the Green function, as there are various ways for carrying out the required frequency integration; and there are different ways to screen the Coulomb potential, etc. *GW* studies are typically (unless done self-consistently) based on a particular ground-state calculation which also may largely affect the results. Hence, to make *GW* calculations comparable and to be able to distinguish between the accuracy of an approach and the precision of a numerical implementation, an urgent need is to convince code developers to fully document their codes, providing information on used approximations, algorithm and related numerical parameters.

Likewise, particle-based methods (classical molecular-dynamics) not only suffer from huge volumes but also from very many different force-field implementations in a large variety of codes.

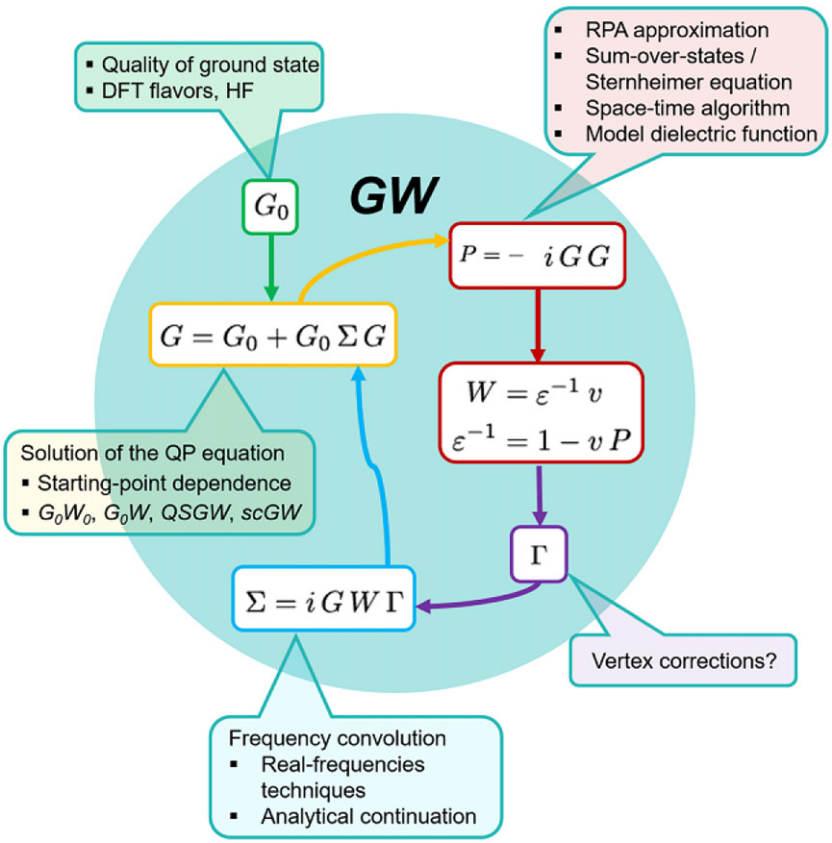

Figure 25. Complexity of excited-states calculations exemplified by the *GW* approach.

Turning to the experimental side, heterogeneity appears to be an even much bigger problem. Nevertheless, the NOMAD Laboratory has made first steps to include results from different probes. A concept based on the NOMAD experience, describing how data from experiment and synthesis shall be processed and incorporated into a federated data infrastructure is described elsewhere [191].

4.1.3. Advances in science and technology to meet challenges

Overall, the most critical step towards FAIR handling of all materials-science data is the establishment of a metadata schema for each synthesis route, experimental probe, and theoretical approach, connected by a materials ontology [190]. Importantly, the metadata must be as *complete* as possible to allow for the assessment of the data quality, i.e., they need to capture all parameters that may influence the results. Here, we introduce a data-analysis tool that is capable of *measuring* the impact of various parameters on *ab initio* calculations. It is based on an implementation that follows the spirit of the density-of-states (DOS) fingerprint by Isayev and coworkers [192]. We take the simple example of bulk silicon to demonstrate the basic idea. In figure 26, we show that among the 2625 single-point calculations hosted by NOMAD, the biggest share has been obtained from **`FHI-aims`** (1525), followed by **`VASP`** (519), **`octopus`** (306), and **`exciting`** (174). Obviously, several XC functionals, basis sets of different quality, etc have been employed to create the data. The top and bottom panel show two examples for how (dis)similar the results are. On the top, we see two calculations by **`exciting`**, one with the local-density approximation (LDA), another one with G_0W_0 on top. We clearly see the well-known effect of G_0W_0, rigidly shifting up the conduction bands in this material. These deviations cause the similarity coefficient to be only moderate (Tc = 0.73). In contrast, using the same functional as is the case in the bottom panel, the results are more similar (Tc = 0.83). Here the differences stem from the usage of different codes (at equal lattice parameter), increasing at lower energies. This example is, of course, only a rough assessment. The method can, however, be refined by including considerations of basis-set quality, *k*-mesh, and various other computational parameters on the one hand or structural differences on the other hand. All this is currently studied in more detail and published elsewhere [190, 193].

Our strategy is to investigate first the origin of discrepancies on the basis of dedicated data sets (e.g., those used in references [185, 189]) before exploring the entire NOMAD data space.

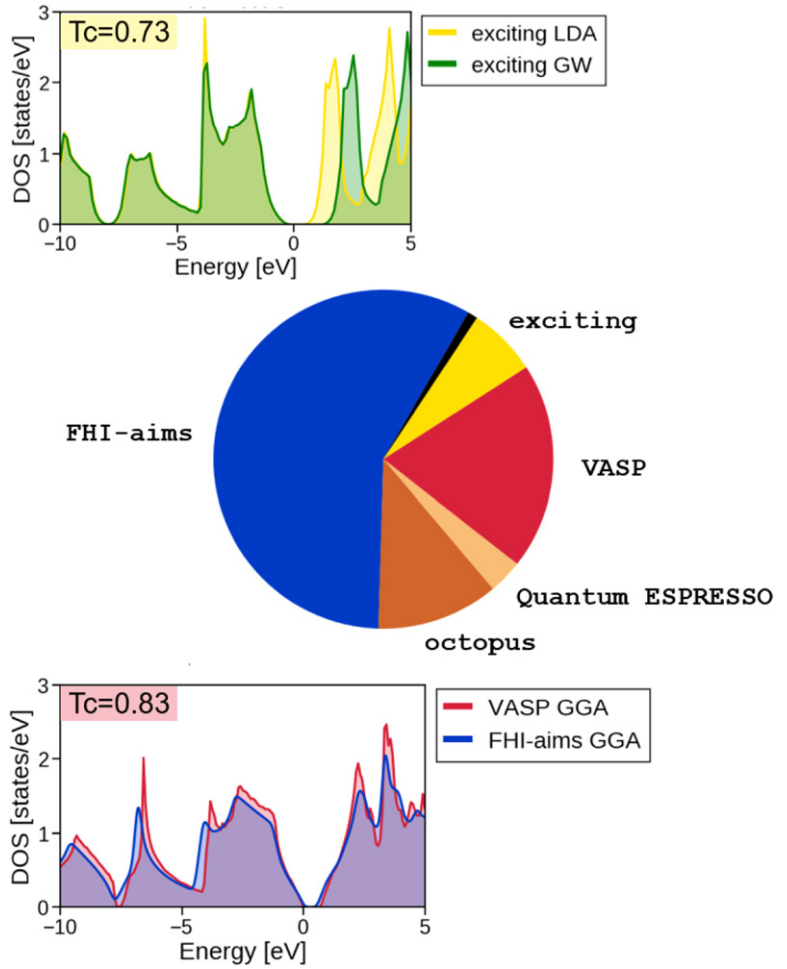

Figure 26. (Center) Pie chart showing the distribution of bulk Si calculations in the NOMAD Laboratory computed by different codes: FHI-aims (1525), VASP (519), octopus (306), exciting (174), Quantum ESPRESSO (86), others (25). (Top) DOS obtained by *GW*@LDA and LDA with **exciting**. (Bottom) DOS obtained by PBE with VASP and FHI-aims. Tc indicates the similarity coefficient [193] between the two respective calculations. It reflects the entire energy range between −10 and +5 eV. For the corresponding data see reference [194].

4.1.4. Concluding remarks

Reaching full interoperability of data from different sources represents a huge challenge for fully exploiting the enormous data space created by the community. Also benchmark results are largely missing so far. To close this gap, we not only create reference data for prototype materials but also aim at assessing the impact of various approximations and computational parameters. Here, we have shown a tool that enables the assessment of methods and data in terms of the DOS. In the future, our investigations will be expanded to other materials properties of interest, and will be further developed towards inclusion of experimental data. Moreover, our tools can be used to search for materials that exhibit features that are similar to those of other materials but are superior with respect to other criteria. A first version is already implemented in the NOMAD Encyclopedia, providing the most similar materials to a chosen reference.

Acknowledgments

This work received funding from the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 676580), and from the Deutsche Forschungsgemeinschaft (DFG), projects 414984028 (SFB 1404, FONDA) and 460197019 (NFDI consortium FAIRmat).

4.2. The AFLOW framework for computational materials data and design

Marco Esters[1], David Hicks[1] and Cormac Toher[2,*]

[1]Duke University, United States of America

[2]UT Dallas, United States of America

E-mail: cormac.toher@utdallas.edu

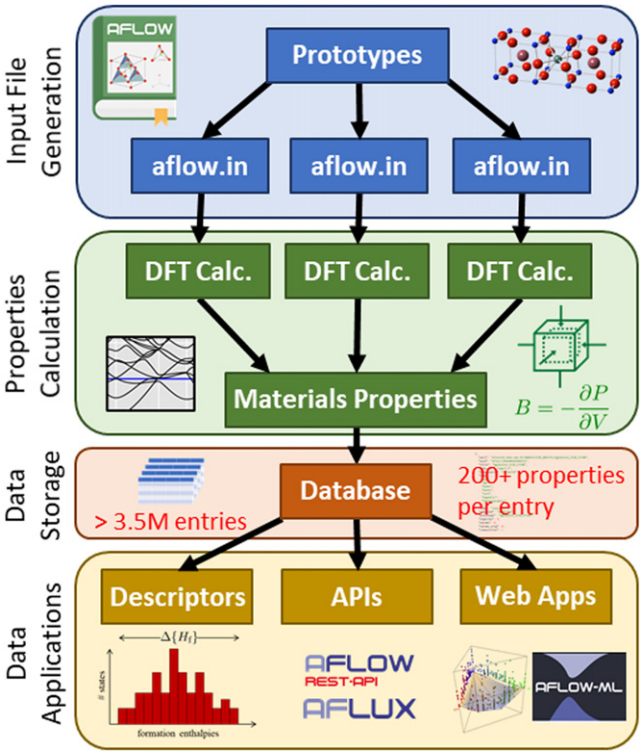

Figure 27. Schematic representation of the AFLOW framework. Starting from a large set of prototypes, properties are calculated from *ab initio* methods and served to the public via web applications and data APIs.

4.2.1. Status

The creation and curation of large, reliable, and standardized data sets to train and validate models poses a significant challenge in applying ML to materials science. Large repositories such as the AFLOW database provide an excellent opportunity to combine automated frameworks with data science and AI [195].

Figure 27 demonstrates the AFLOW data generation workflow. Input structures are based on experimentally observed materials or are generated from the over 1100 crystallographic prototypes spanning all 230 space groups in the prototype encyclopedia [196]. Integrated within the AFLOW software, the encyclopedia automatically decorates these prototypes with different elements to generate new hypothetical compounds. Starting with these structures, AFLOW performs DFT calculations via a standardized set of parameters.

The AFLOW software calculates a variety of structural, thermodynamic, electronic, and thermo-mechanical properties that can be used for training and validating ML models. Structures are characterized with the AFLOW-SYM module, a tool to determine common symmetry descriptors [197]. The routines are self-consistent and adaptive, freeing users from needing to tune tolerance thresholds, and real- and reciprocal-space isometries are guaranteed to be commensurate. Thermodynamic properties are calculated via the AFLOW-CHULL module, which can be used to determine phase stability, phase coexistence, decomposition reactions, and the synthesizability of materials [198]. Thermo-mechanical properties such as bulk and elastic moduli, thermal expansion, and vibrational thermodynamics can be determined using the AFLOW-GIBBS and elasticity libraries [195].

At over 3.5 million entries with over 200 properties each, the resulting database is one of the largest of its kind. It is available to the public through various web applications and data APIs on https://aflow.org. The generated data have been used to develop property descriptors and to train ML models, such as the property-labelled materials fragments (PLMF) model, to predict thermo-mechanical and electronic properties [199]. PLMF and other models are available through the AFLOW-ML web application and Python module [200].

4.2.2. Current and future challenges

Programmatic access to the vast quantity of materials data in the AFLOW database is necessary to maximize its usefulness for training ML models. Materials APIs such as the AFLOW-REST-API simplify the retrieval of properties from a specific entry by querying a uniform resource locator (URL), but only allow access to one database entry at a time and require users to know in advance which materials to request. To generate data for ML, a database needs to be searchable by properties without advance knowledge of its structure and format, while also being code-base agnostic.

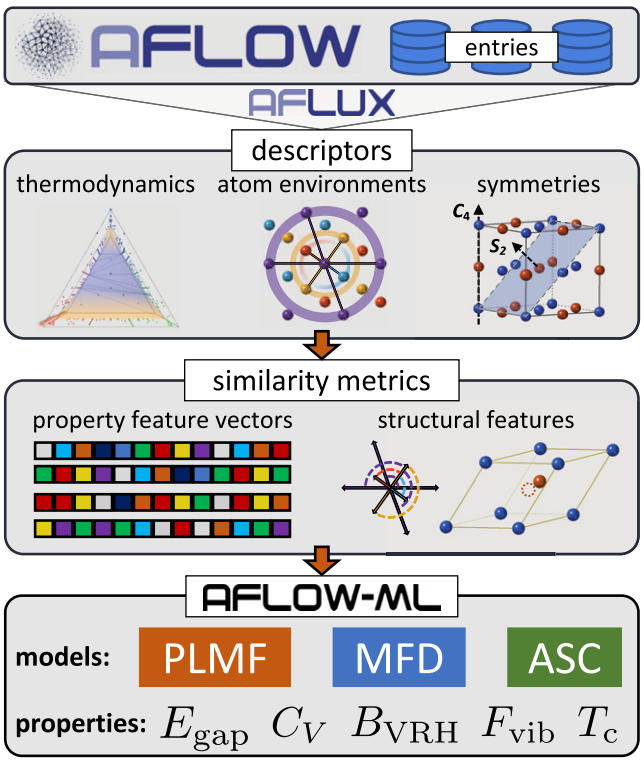

Figure 28. Workflow for training ML models. Materials properties in the AFLOW database can be queried via AFLUX to build descriptors and features. The AFLOW-ML web application and API can be used to predict thermo-mechanical, electronic, and vibrational properties. Reproduced from [202]. CC BY 4.0.

In addition to high-volume data, ML models rely on diverse data sets that are free from duplicates to prevent training bias. Identifying distinct crystalline compounds is a considerable challenge due to varying representations of the structure. Standard conversion techniques are error prone as similar structures may be cast into different representations, and symmetry descriptors alone do not determine structural equivalence. Unique prototype designations are also necessary to distinctly label structures and enable searches for certain structure-types in materials repositories. Since by-hand duplicate removal and structure labelling are intractable given the growth rate of materials data, rigorous structural similarity metrics and classification algorithms are required to remove duplicate compounds to improve model prediction and identify unique structure-types.

Developing reliably predictive ML algorithms requires accurate training data. DFT has shortcomings that can skew ML results: band gaps tend to be underestimated, and formation enthalpies for polar materials such as oxides are unreliable. Meanwhile, high-entropy alloys and ceramics, a newly emerging class of materials, are difficult to model directly, and their synthesizability cannot be sufficiently described using enthalpy alone. They are often represented using large supercells, making it expensive to generate large data sets. New theories and correction schemes are thus required to provide accurate data sets for ML algorithms.

4.2.3. Advances in science and technology to meet challenges

The domain-specific AFLUX language provides programmatic access to the AFLOW database [201]. It combines the accessibility of a data API with the features of a search interface without requiring knowledge of the database structure. Searches can be performed using the query part of the URL with only a minimal set of logical operators. For example,

https://aflow.org/API/aflux/?species(!Pb),Egap(1*,*3),paging(0)

returns all lead-free entries with a band gap of 1 eV to 3 eV. Requests can be arbitrarily complex, giving users control over the data they receive without requiring pruning. AFLUX outputs data in JavaScript Object Notation (JSON) or in plain text for languages without native JSON capabilities. Using AFLUX, AFLOW can be easily integrated into ML workflows as shown in figure 28. Materials properties can be extracted from the AFLOW repositories via the AFLUX API. From the data and through use of the AFLOW modules, descriptors and feature vectors can be fed into ML algorithms to train and validate models.

To enhance the diversity of the database, the AFLOW-XtalFinder module identifies and classifies new prototype structures and maps compounds into their ideal prototype designation [202]. Similarity metrics

distinguish isopointal and isoconfigurational structures regardless of their representation via internal symmetry routines, local geometry analyses, and atom mapping procedures. Unique prototypes are added to the AFLOW prototype encyclopedia, and distinct compounds are prioritized for inclusion in the database. Comparisons can be performed on user data sets to group similar compounds or structures, removing duplicates to save computational resources, and to eliminate training bias for ML models. All AFLOW entries have been mapped into their ideal prototype designation, enabling users to search the database by structure type.

New models have been developed to describe materials that are challenging for DFT. For polar compounds, the coordination corrected enthalpy method corrects formation enthalpies based on bonding environments, significantly improving accuracy [203]. To investigate configurational disorder (see section 1.6 for discussion of disordered materials), the AFLOW partial occupation module generates ensembles of ordered configurations [204]. Calculated properties of the configurations are weighted according to the Boltzmann distribution to model the behaviour and energy spectrum of the disordered material.

4.2.4. Concluding remarks

The AFLOW ecosystem provides an opportunity to combine big data with AI to discover new materials. With over three million entries, its database is the largest of its kind, offering a variety of structural, electronic, thermodynamic, and thermo-mechanical properties. The data can be programmatically accessed and filtered using the AFLUX language. New structures are continuously identified with AFLOW-XtalFinder, and similarity metrics ensure compounds are unique, improving database diversity and reducing training bias for ML models. Methods developed within the *ab-initio* AFLOW-workflow improve formation enthalpy predictions for polar materials and enable modelling of disordered compounds, providing avenues to research new classes of materials. AFLOW can be easily integrated into ML workflows, making it a valuable tool for AI based materials research.

Acknowledgments

The authors thank Frisco Rose and Michael J Mehl for fruitful discussions, and acknowledge support by DOD-ONR (N00014-17-1-2090, N00014-17-1-2876) and by the National Science Foundation under DMREF Grant No. DMR-1921909.

5. Frontier developments of machine learning in materials science

5.1. Adaptive learning strategies for electronic structure calculations

Prasanna V Balachandran

University of Virginia, United States of America

5.1.1. Status

Adaptive learning is an emerging paradigm in materials informatics for rapid and efficient navigation of the vast parameter space [205–208]. The basic idea behind adaptive learning is that a supervised ML algorithm can achieve improved performance with fewer training data points, provided the learning task is carried out by allowing the algorithm to autonomously choose data points from the vast unknown or unexplored space [209]. Any supervised ML method needs training data to build the models. The choice of the training data should be such that it is diverse and representative of the problem of interest. A poor choice will impair the predictive and generalizable capability of the trained ML models. Therefore, it is critical to *optimally* sample the search space. The reason why adaptive learning is particularly well-motivated in electronic structure calculations is due to the expensive nature of the calculations, where a brute-force approach is not efficient or practical. Surrogate models that approximate the predictions of the expensive electronic structure codes with little computational cost has the potential to accelerate the design and discovery of new materials.

We can broadly classify adaptive learning into two categories: AL and Bayesian optimization. In both AL and Bayesian optimization, we start from a small number of labelled instances to train ML models, but have a massive number of unlabelled instances. In AL, the trained models are programmed to choose an instance from the massive unlabelled dataset that they are least confident in predicting, thus reducing the overall uncertainty or error. Whereas in Bayesian optimization, the interest is in rapidly finding the optimum of a function that are costly to evaluate and lack gradient information. Both methods employ utility (or acquisition) functions that query each unexplored data point in the search space. Promising data points (that satisfy a well-defined constraint) are recommended for the next iteration of validation and feedback. Learning from data is formulated as an iterative process until convergence is reached. We can run these calculations either sequentially (where we select one data point at a time for validation and feedback) or in a batch mode (where several data points are selected at a time for validation and feedback). There are growing examples in electronic structure calculations, where AL and Bayesian optimization approaches are finding increasing use [210–212]. As high-performance

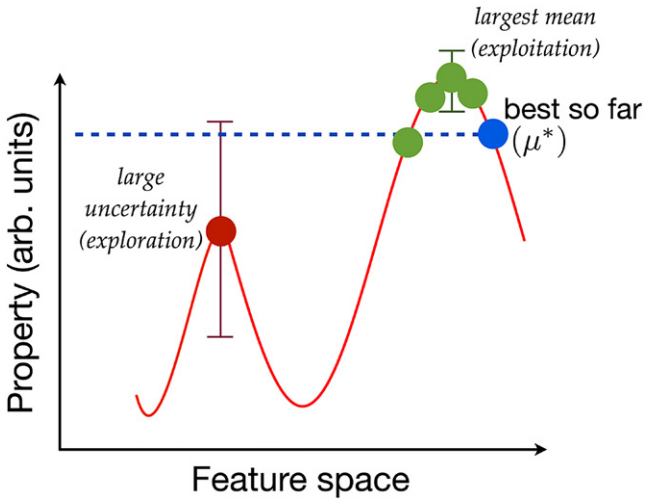

Figure 29. A schematic describing the purpose of a utility function in evaluating the trade-off between exploitation and exploration. The solid red curve is the response surface from a trained ML model. The blue data point is the best data point in the training data. The green and red data points represent unexplored data points in the design space. Reproduced from [208]. CC BY 4.0.

computing capabilities improve and databases grow in number and complexity (e.g., complex interfaces, surfaces and heterostructures), integrating massively parallel electronic structure codes with adaptive learning will be critical for efficient exploration of the vast search space.

5.1.2. Current and future challenges

There are two necessary ingredients for implementing adaptive learning: (1) an ML method that will allow for quantifying uncertainties in every explored and unexplored data point in the search space. It is common to use either posterior probability distributions from Bayes' theorem or parametric confidence intervals for UQ [206, 207]. (2) A utility function that will take the expected value and the associated uncertainties from the trained ML models as input to setup the query, and rank each data point in an order of value, representativeness, and/or diversity.

Off-the-shelf methods such as the random forests and GPR are the workhorses for AL and Bayesian optimization, respectively. These methods provide a probabilistic measure of the output quantity. However, there are many other state-of-the-art ML methods such as the KRR, support vector machines (SVM) and ANNs. But they do not have the intrinsic capability to quantify uncertainties. Application of ensemble learning and Bayesian inference-based approaches to KRR, SVM, and ANN can overcome these limitations. This will equip the community with more tools to build ML models for a given dataset. There are two reasons why we should think beyond GPR and random forests models: (1) *a priori* we do not know which ML algorithm will be better suited for a given data set. No-free-lunch theorems states that there are no universal ML algorithms that will work for every problem [213]. The GPR and random forests are convenient choices, but are not optimal in all settings. (2) Not all ML algorithms have strong scaling performance with the number of training samples and input dimensionality.

Utility functions evaluate the trade-off between exploration and exploitation of the search space on the basis of the current performance of the trained ML models (figure 29). Some of the popular utility functions include uncertainty sampling, expected improvement, knowledge gradient, probability of improvement, upper confidence bound and mean objective cost of uncertainty (to name a few) [207]. If adaptive learning is operated in batch mode, then we need additional strategies to select diverse data points such that the model performance will be improved [209]. Another promising application of Bayesian optimization is in accelerating reinforcement learning (RL) algorithms through efficient hyperparameter optimization.

5.1.3. Advances in science and technology to meet challenges

There is a sufficient body of published research in the literature that demonstrate the efficacies of the adaptive learning methods in computational and experimental materials science. Yet, the field is still in its infancy requiring key advancements to accelerate the pace of scientific discoveries. We list some of the primary challenges, along with the advances needed to meet them. (1) At the ML level, there is a need to develop an improved understanding of the response surface that capture the quantitative input-output relationships. The ML methods work well when the training data points that are located in close proximity to one another in the input space also have similar output properties. Presence of 'property cliffs', where two similar inputs have a large

difference in the responses can have an adverse effect on the performance [205]. Post hoc model interpretability methods (local and global) can provide clues by opening the complex black-box models and making it easier for the domain experts to comprehend why certain predictions are made in order to tackle the property cliff problem. (2) There is also a growing trend to combine classification learning with regression methods to address a common overarching goal [214]. The training data for classification learning and regression may (or may not) be different, but there is a vast unexplored search space that is common to both methods for efficient navigation. It is unclear how the uncertainties will propagate between the two independent models to inform the decision-making process. (3) The role of domain experts-in-the-loop is also vital to advance the adaptive learning paradigm. A vast majority of the current approaches rely on off-the-shelf methods that do not readily incorporate domain knowledge. One of the current trends where domain experts have had an overwhelming influence is via the choice of meaningful descriptors or representations. Advances are needed in the domain knowledge-informed kernel design, UQ, and utility functions. (4) Given that there are many choices for selecting ML methods and utility functions, it is unclear how a particular ML-utility function pair will perform on a given data set. Currently, there are no heuristics that can guide us to select an informed pair for a given problem. We need benchmark datasets (similar to the MNIST dataset in computer science) to reliably test the performances of various adaptive learning strategies [205, 215].

5.1.4. Concluding remarks

The excitement surrounding the adaptive learning research is palpable. The success of adaptive learning will be key to enable autonomous computing of materials properties and on-the-fly closed-loop high-throughput computations and experiments. We envision that many research groups will continue to creatively integrate these strategies into their design scheme, which will positively impact its growth. However, to sustain the excitement, several outstanding research challenges remain to be addressed. Some of the urgent needs are discussed in this article. Future developments will rely on advances in building interpretable ML models, UQ methods, and utility functions that will take advantage of the unique properties of the problems under investigation.

Acknowledgments

PVB thanks Huozhi Zhou, Lav Varshney, and Yangfeng Ji for insightful discussions. Research was sponsored by the Defense Advanced Research Project Agency (DARPA) and The Army Research Office and was accomplished under Grant Number W911NF-20-1-0289. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of DARPA, the Army Research Office, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not with standing any copyright notation herein.

5.2. Reinforcement learning

Isaac Tamblyn[1], Steven Whitelam[2], Colin Bellinger[3]

[1]University of Ottawa, Canada

[2]Lawrence Berkeley National Laboratory, United States of America

[3]National Research Council, Canada

5.2.1. Status

RL is a branch of ML that focuses on an agent repeatedly interacting with an environment for the purpose of maximizing a reward (figure 30). More formally, these are defined as sequential decision making problems. Inspired by animal learning and behaviour, the field of RL dates back to ideas developed at the beginning of the 20th century in the field of human behaviour and learning. Edward Thorndike transformed cognitive psychology with the notion that behaviour could be shaped through the iterative application of reward and punishment. This concept soon became doctrine and was shown to be a crucial part of biological learning. B F Skinners' boxes and animal training experiments in the 1940s made explicit the connection between learning through repetition and reinforcement. These ideas were then codified within the computer science literature by a number of contributors [216].

In recent years, when coupled with the flexible representation and capacity of deep NNs, RL has seen a renaissance, solving problems such as video games, the ancient game of Go, and the navigation and station-keeping of autonomous craft [217]. Dedicated silicon (e.g. GPU, TPU, and FPGA) coupled with modern neural architectures have made it practical for RL agents to learn directly from visual input.

Compared to other ML approaches, RL has not yet seen widespread use within the fields of physics and chemistry. Examples thus far include molecular design tasks [218, 219], navigation of chemical synthesis pathways [220], scientific discovery [221, 222] and experimental control [223]. However, we believe that RL has considerable promise for future applications in the field of material science and in science more generally.

RL algorithms learn, through experience, how to achieve an optimal control policy for a dynamical system. They can also be used to solve problems of optimization, but the same is true of other ML approaches. RL excels at control of dynamics.

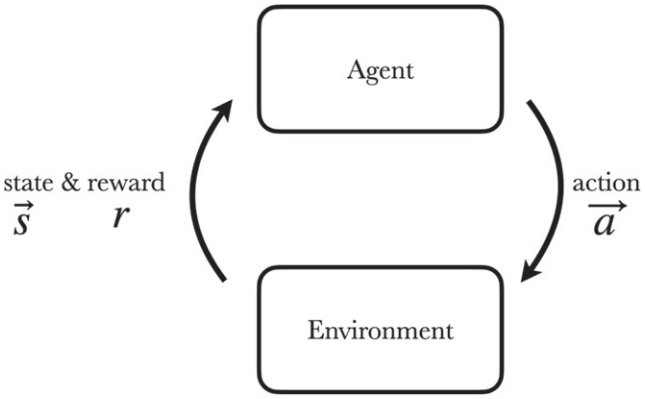

Figure 30. In RL, an agent interacts with an environment through a series of actions which give rise to changes of its state. The agent seeks to maximize the discounted sum of rewards it receives as feedback.

Generally speaking, if a system or environment is one which can respond to external stimuli (e.g. increase the temperature of a reaction vessel, apply an electric field, etc), permits sequential interaction, and there is some way to quantify a notion of success, RL can be applied.

5.2.2. Current and future challenges

Given the broad range of RL algorithms which exist, we highlight here some general issues, rather than those specific to a particular approach. The large number of RL algorithms which exist is in fact one issue in the field. There is general agreement within the community that we have not yet witnessed the 'Imagenet' moment of RL; no single algorithm is broadly applicable and competitive in all cases. Most algorithms are sample inefficient [224], meaning that a large number of training examples (episodes) must be played in order for an algorithm to learn. Like many ML sub-disciplines, interpretable models that scale to real-world problems are still not common. Results can also be quite sensitive to the choice of model hyper-parameters, and often require manual tuning.

Sparse-reward problems are particularly challenging to learn. When success is only defined by the end goal, there is essentially no learning signal for the agent to work with. This can exacerbate the problem of sample inefficiency. Reward shaping is a way to address this, however, if a problem does not have an intrinsic and reliable reward signal, designing one which achieves the desired agent behaviour can be time-consuming and error prone. There are many humorous examples of agents displaying untended behaviour while chasing a poorly designed reward function.

Curriculum learning appears to offer both an effective and intuitive solution to learning, however, determining a curriculum is itself a difficult problem. Indeed, this is something that even our own education systems struggle with regularly. Within the domain of games, self-play has been used as an effective form of curriculum generation, although this approach is not suited to all problems.

5.2.3. Advances in science and technology to meet challenges

Within the current paradigm, RL algorithms require an accurate and efficient mechanism to approximate either the policy, value function or both. Recent progress in the field has been strongly linked to using NNs for this purpose as they have been shown to be good general functional approximators. Unfortunately, deep networks tend to be 'data hungry' and suffer from catastrophic forgetting.

The limitations (and benefits) of NNs directly impact the performance and characteristics of RL implementations which are built with them. A major improvement in RL performance would therefore be achieved simply with algorithmic or hardware acceleration which can produce accurate approximate functions with less training data (and can efficiently incorporate new data). Improvements to function approximators used in RL would have high impact. Meta-learning (algorithms which learn based on the behaviour of other learning algorithms), improvements in off-line training, and better sim-to-real [225] are all areas which offer potential areas for improvement.

More generally, representations are again an area where there is significant room for improvement, particularly in the domain of physics and chemistry. As with supervised learning, when a learning algorithm is provided data without any prior knowledge, a significant amount of signal is required simply to learn the relevant features. This is in contrast to a scientist when they first enter the laboratory; they already have a great deal of experience and expertise for operating in $3d$ environments.

RL algorithms can be broadly categorized as either model-based or model-free (although there are cases where the line becomes blurred). The former category consists of algorithms which use the model for a variety of tasks: planning, obtaining analytic gradients, value-equivalence prediction, and data generation [ref_categories]. An area for improvement could be the incorporation of hierarchies of models into such agents. Currently this is not standard practice, and stands in stark contrast to physics and chemistry which are strongly based on hierarchies of models which describe different effects across a wide-range of length and time scales. Strong physics-based priors (e.g. energy conservation, momentum conservation) are also generally not built into RL models in the way that is typically done in material science.

Intrinsic rewards will likely continue to be a fruitful area of methodological development. Concepts such as novelty and surprise have begun to be applied with promising initial results.

5.2.4. Concluding remarks

RL is a powerful way of solving control problems. Thus far, it has seen less use within the materials science literature than other ML methods such as supervised learning (e.g. image classification) and unsupervised learning (e.g. finding patterns within data). We believe that RL is poised to make significant breakthroughs within materials science, judging by its success with game-playing and autonomous control.

Acknowledgements

This work was performed as part of a user project at the Molecular Foundry, Lawrence Berkeley National Laboratory, supported by the Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02–05CH11231. CB performed work at the NRC under the auspices of the AI4D Program. IT acknowledges NSERC

5.3. Interpretability of machine learning models in physical sciences

Luca M Ghiringhelli

The NOMAD Laboratory at the FHI of the Max-Planck-Gesellschaft and IRIS Adlershof of the Humboldt Universität, Berlin, Germany

5.3.1. Status

Training a supervised ML model that yields satisfactory predictions (i.e., that maps input features into values of the target property, with errors below a threshold perceived as tolerable) on test data that are driven from the same distribution as the training data, is a task that is nowadays almost routinely accomplished. However, the crucial interest is in that the trained model can *generalize*, i.e., it can yield trustable predictions also for test data that are significantly different from the training data. As human beings, i.e., users who are asked to judge and/or trust predictions of a ML model, we need to *understand what the model has learned*. Such innate need is related to the notion of *interpretability* of the ML model. The literature on interpretability is vast [226–232], but the field is pre-paradigmatic, i.e., it has not reached a consensus on what are the fundamental questions and what are the quantities to be measured. Two somewhat contrasting aspects are typically associated to interpretability [226–228]: *transparency* of the model and its (*post hoc) explainability*. Transparency connects to scientific practice, where a phenomenon is felt as understood when a predictive mathematical law is formulated, which is expected to work with no exception, at least in a well-defined *domain of applicability*. Such law is expected to be *simple*, so that our brains can process most, if not all, of its consequences. *Explainability* refers to the possibility to inspect a perceived 'black-box', i.e., a model that is in general too complex to be grasped by the human mind, but that can be investigated, in order to reveal, for instance, which parts of the input mostly affected the output. Incidentally, understanding a decision made by a human refers to the *post hoc* explainability of what happens in our brains, whose detailed mechanics are beyond current grasp, while we can provide reasons on how a decision was reached, typically based on 'similar cases' [227]. Understanding interpretability and in particular devising one or a set of consensual *metrics* for assessing the generalizability and trustability of ML model is one crucial next step, or the field might face another 'winter' due to a consequent lack of trust in ML applicability.

5.3.2. Current and future challenges

The tools for addressing the interpretability of ML models vary with the complexity of the models [226–231] (see figure 31). For simpler models, *transparency* is evaluated, i.e., the ability to read and inspect the model. *Sparse* models [233] and in particular *symbolic inference* [234] naturally provide transparent models as they appear as equations (or inequalities) in terms of functions of input features, which are selected out of a possibly large number of candidates. The interpretation is therefore provided by the identification of which input features govern the modeled phenomenon. Here, the notions of *simulatability* and *decomposability* have been introduced. These are the ability to follow step-by-step how the ML model produces an output from the input and the ability to assign a meaning to each part of a model (e.g., the sign and magnitude of regression coefficients), respectively. An outstanding challenge is to define a rigorous *metric of transparency*, so that models

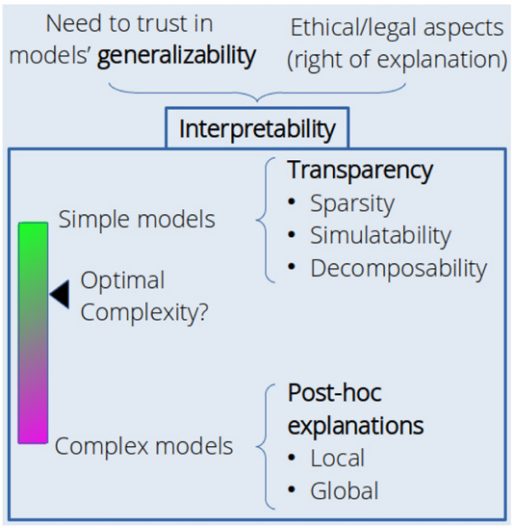

Figure 31. Summarized view of interpretability in ML, elucidating what determines the users' need for interpretability and how the meaning of the word adapts to the complexity of the learned model. Not mentioned in the text, the ethical/legal aspects are felt important when ML-based decisions impact individuals or communities.

can be objectively compared, similarly and complementarily to the routinely performed, but insufficiently informative, comparison in terms of predictive accuracy.

For more complex models, where transparency is lost, a plethora of *post hoc explanation* tools have been developed [229–232], which are commonly divided into *local* (explanation on how a given single output is obtained) and *global* (typically, visual analysis of how the dataset is represented internally by the model). The focus is in general on a statistical analysis on how input features affect the results. The challenge is here to properly account for the (typically nonlinear) relationship among the input features.

It is highly unsatisfactory that two different interpretability concepts exist depending on the complexity of the trained model. In facts, there is a continuum of complexity between sparse, symbolic models and complex ones (e.g., deep NNs); the challenge is to seamlessly adapt the complexity of the learned model, and the related interpretability tools, to the intrinsic complexity of the underlying input-features—target-property relationship.

Finally, the importance of *outliers*, datapoints not conforming to the model being learned, needs to be understood. In physical sciences, a wrongly predicted datapoint may be a signal that a different mechanism from the so-far identified features-property relationship is at work.

5.3.3. Advances in science and technology to meet challenges

ML is urgently requested to undergo a paradigm change. Together with prediction accuracy, strategies for assessing the correct model complexity and interpretability metrics need to be developed. If a simple, symbolic law is the underlying model, a correct ML strategy must be able to recover such exact model. When a more complex, less transparent model is necessary, then the interpretability metric needs to seamlessly adapt to the increased complexity. It should become therefore common practice to compare models in terms not only of their predictive accuracy, but also of their interpretability metric. When applied to the development of scientific (e.g., physical) laws, the purpose of this formidable task is to provide reasons to accept an ML-learned features—property relationship in terms of its consistency with the existing bulk of knowledge, so that the ML model is not felt as a surrogate, until 'something better' is found, but as a new scientific law.

In this respect, it is crucial to be able to treat the nonconforming datapoints. Most current ML approaches are built to neglect such datapoints, a.k.a. outliers, while in physical sciences even one single datapoint not complying with the general law is treated with uttermost care, as it could be the harbinger of 'new physics'. It is therefore desirable that, together with the complexity-aware strategy sketched above, a nonconforming-datapoints strategy is developed (see also section 1.4). For instance, one may wish to detect different *domains of applicability* of more complex, general models, vs specialized but simpler models. A useful analogy could be thinking at general relativity, which is more general and more complex than classical gravitation. The latter is however very accurate in a well-specified and understood *domain of applicability*. In turn, general relativity is expected to be a special, somewhat simpler, case of a (yet to be developed) quantum-gravity theory. Similarly, in ML the level of complexity of the learned models might need to be adapted to well-defined domains of applicability [235], preferably defined by ML algorithms in a data-driven fashion.

5.3.4. Concluding remarks

In conclusion, ML might have reached its maturity in terms of predictive ability, on data that are statistically similar to the training data. However, it is still in its infancy when it comes to (i) generalizability to data significantly different from training data, (ii) treatment of 'outliers', i.e., data do not conform to the model being trained, (iii) having a unified concept of interpretability that seamlessly applies from the obvious transparency of sparse, symbolic models, to the explainability of complex deep NNs, and (iv) adapting the trained model complexity to the intrinsic complexity of the underlying input feature—property relationship. Hopefully, framing the objective in clear terms will stimulate a focused development of ML techniques, which could promote ML tools to become valuable companions of a scientist, in order to foster future scientific discoveries.

Acknowledgments

I acknowledge Jilles Vreeken, Angelo Ziletti, and Matthias Scheffler for insightful discussions. This work received funding from the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 676580 and No. 951786), the NOMAD laboratory CoE, and ERC:TEC1P (No. 740233).

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

H J Kulik https://orcid.org/0000-0001-9342-0191
T Hammerschmidt https://orcid.org/0000-0002-2270-4469
S Botti https://orcid.org/0000-0002-4920-2370
M A L Marques https://orcid.org/0000-0003-0170-8222
M Todorović https://orcid.org/0000-0003-0028-0105
P Rinke https://orcid.org/0000-0003-1898-723X
C Oses https://orcid.org/0000-0002-3790-1377
A Smolyanyuk https://orcid.org/0000-0002-4859-5977
S Curtarolo https://orcid.org/0000-0003-0570-8238
A P Bartók https://orcid.org/0000-0002-4347-8819
S Manzhos https://orcid.org/0000-0001-8172-7903
T Carrington https://orcid.org/0000-0002-5200-2353
J Behler https://orcid.org/0000-0002-1220-1542
O Isayev https://orcid.org/0000-0001-7581-8497
M Veit https://orcid.org/0000-0001-7813-4015
A Grisafi https://orcid.org/0000-0003-1433-125X
J Nigam https://orcid.org/0000-0001-6857-4332
M Ceriotti https://orcid.org/0000-0003-2571-2832
K T Schütt https://orcid.org/0000-0001-8342-0964
J Westermayr https://orcid.org/0000-0002-6531-0742
R J Maurer https://orcid.org/0000-0002-3004-785X
K Burke https://orcid.org/0000-0002-6159-0054
F Noé https://orcid.org/0000-0003-4169-9324
S Pilati https://orcid.org/0000-0002-4845-6299
C Draxl https://orcid.org/0000-0003-3523-6657
M Esters https://orcid.org/0000-0002-8793-2200
D Hicks https://orcid.org/0000-0001-5813-6785
C Toher https://orcid.org/0000-0001-7073-8690
P V Balachandran https://orcid.org/0000-0002-7496-5521
I Tamblyn https://orcid.org/0000-0002-8146-6667
L M Ghiringhelli https://orcid.org/0000-0001-5099-3029

References

[1] Janet J P, Liu F, Nandy A, Duan C, Yang T, Lin S and Kulik H J 2019 Designing in the face of uncertainty: exploiting electronic structure and machine learning models for discovery in inorganic chemistry *Inorg. Chem.* **58** 10592–606

[2] Janet J P and Kulik H J 2017 Resolving transition metal chemical space: feature selection for machine learning and structure-property relationships *J. Phys. Chem.* A **121** 8939–54
[3] Meyer B, Sawatlon B, Heinen S, Anatole von Lilienfeld O and Corminboeuf C 2018 Machine learning meets volcano plots: computational discovery of cross-coupling catalysts *Chem. Sci.* **9** 7069–77
[4] Janet J P, Chan L and Kulik H J 2018 Accelerating chemical discovery with machine learning: simulated evolution of spin crossover complexes with an artificial neural network *J. Phys. Chem. Lett.* **9** 1064–71
[5] Janet J P, Ramesh S, Duan C and Kulik H J 2020 Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization *ACS Cent. Sci.* **6** 513–24
[6] Herbol H C, Poloczek M and Clancy P 2020 Cost-effective materials discovery: Bayesian optimization across multiple information sources *Mater. Horiz.* **7** 2113–23
[7] Janet J P, Duan C, Yang T, Nandy A and Kulik H J 2019 A quantitative uncertainty metric controls error in neural network-driven chemical discovery *Chem. Sci.* **10** 7913–22
[8] Liu F, Duan C and Kulik H J 2020 Rapid detection of strong correlation with machine learning for transition-metal complex high-throughput screening *J. Phys. Chem. Lett.* **11** 8067–76
[9] Duan C, Janet J P, Liu F, Nandy A and Kulik H J 2019 Learning from failure: predicting electronic structure calculation outcomes with machine learning models *J. Chem. Theory Comput.* **15** 2331–45
[10] Stein C J and Reiher M 2016 Automated selection of active orbital spaces *J. Chem. Theory Comput.* **12** 1760–71
[11] Pettifor D G 1984 A chemical scale for crystal-structure maps *Solid State Commun.* **51** 31
[12] Glawe H, Sanna A, Gross E K U and Marques M A L 2016 The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining *New J. Phys.* **18** 093011
[13] Bialon A F, Hammerschmidt T and Drautz R 2016 Three-parameter crystal-structure prediction for *sp-d* valent compounds *Chem. Mater.* **28** 2550
[14] Jenke J, Subramanyam A P A, Densow M, Hammerschmidt T, Pettifor D G and Drautz R 2018 Electronic structure based descriptor for characterizing local atomic environments *Phys. Rev.* B **98** 144102
[15] Sutton C *et al* 2019 Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition *npj Comput. Mater.* **5** 111
[16] Rosenbrock C W, Homer E R, Csányi G and Hart G L W 2017 Discovering the building blocks of atomic systems using machine learning: application to grain boundaries *npj Comput. Mater.* **3** 29
[17] Reimann D, Nidadavolu K, ul Hassan H, Vajragupta N, Glasmachers T, Junker P and Hartmaier A 2019 Modeling macroscopic material behavior with machine learning algorithms trained by micromechanical simulations *Front. Mater.* **6** 181
[18] Volz N *et al* 2021 Understanding creep of a single-crystalline Co–Al–W–Ta superalloy by studying the deformation mechanism, segregation tendency and stacking fault energy *Acta Mater.* **214** 117019
[19] Huang B and Anatole von Lilienfeld O 2020 Quantum machine learning using atom-in-molecule-based fragments selected on the fly *Nat. Chem.* **12** 945
[20] Jenke J, Ladines A, Hammerschmidt T, Pettifor D G and Drautz R 2021 Tight-binding bond parameters for dimers across the periodic table from density-functional theory *Phys. Rev. Mater.* **5** 023801
[21] Woodley S M and Catlow R 2008 Crystal structure prediction from first principles *Nat. Mater.* **7** 937–46
[22] Faber F A, Lindmaa A, Anatole von Lilienfeld O and Armiento R 2016 Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals *Phys. Rev. Lett.* **117** 135502
[23] Schmidt J, Shi J, Borlido P, Chen L, Botti S and Marques M A L 2017 Predicting the thermodynamic stability of solids combining density functional theory and machine learning *Chem. Mater.* **29** 5090–103
[24] Faber F, Lindmaa A, Anatole von Lilienfeld O and Armiento R 2015 Crystal structure representations for machine learning models of formation energies *Int. J. Quantum Chem.* **115** 1094–101
[25] Ward L, Liu R, Krishna A, Hegde V I, Agrawal A, Choudhary A and Wolverton C 2017 Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations *Phys. Rev.* B **96** 024104
[26] Xie T and Grossman J C 2018 Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties *Phys. Rev. Lett.* **120** 145301
[27] Goodall R E A and Lee A A 2019 Predicting materials properties without crystal structure: deep representation learning from stoichiometry (arXiv:1910.00617)
[28] Bartel C J, Trewartha A, Wang Q, Dunn A, Jain A and Ceder G 2020 A critical examination of compound stability predictions from machine-learned formation energies (arXiv:2001.10591)
[29] Kim K, Ward L, He J, Krishna A, Agrawal A and Wolverton C 2018 Machine-learning-accelerated high-throughput materials screening: discovery of novel quaternary Heusler compounds *Phys. Rev. Mater.* **2** 123801
[30] Park C W and Wolverton C 2020 Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery *Phys. Rev. Mater.* **4** 063801
[31] Jha D, Ward L, Paul A, Liao W-k, Choudhary A, Wolverton C and Agrawal A 2018 ElemNet: deep learning the chemistry of materials from only elemental composition *Sci. Rep.* **8** 17593
[32] Gilmer J, Schoenholz S S, Riley P F, Vinyals O and Dahl G E 2017 Neural message passing for quantum chemistry *Proc. 34th Int. Conf. Machine Learning* vol 70 pp 1263–72 JMLR.org
[33] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
[34] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8
[35] Sun J, Ruzsinszky A and Perdew J P 2015 Strongly constrained and appropriately normed semilocal density functional *Phys. Rev. Lett.* **115** 036402
[36] Jain A *et al* 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002
[37] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD) *JOM* **65** 1501–9
[38] Curtarolo S *et al* 2012 AFLOW: an automatic framework for high-throughput materials discovery *Comput. Mater. Sci.* **58** 218–26
[39] Wang H-C, Botti S, Marques M A L and Marques L 2021 Predicting stable crystalline compounds using chemical similarity *npj Comput. Mater.* **7** 12
[40] Sutton R 2018 The bitter lesson http://incompleteideas.net/IncIdeas/BitterLesson.html

[41] Draxl C and Scheffler M 2020 Big data-driven materials science and its FAIR data infrastructure *Handbook of Materials Modeling* ed S Yip and W Andreoni (Berlin: Springer) p 49
[42] Foppa L *et al* 2021 Materials genes of heterogeneous catalysis from clean experiments and artificial intelligence *MRS Bull.* **46** 1016–26
[43] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 Big data of materials science: critical role of the descriptor *Phys. Rev. Lett.* **114** 105503
[44] Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M and Ghiringhelli L M 2018 SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates *Phys. Rev. Mater.* **2** 083802
[45] Goldsmith B R, Boley M, Vreeken J, Scheffler M and Ghiringhelli L M 2017 Uncovering structure-property relationships of materials by subgroup discovery *New J. Phys.* **19** 013031
[46] Frazier P I and Wang J 2016 Bayesian optimization for materials design *Information Science for Materials Discovery and Design* ed T Lookman, F J Alexander and K Rajan (Berlin: Springer) pp 45–75
[47] Ghiringhelli L M 2021 An AI-toolkit to develop and share research into new materials *Nat. Rev. Phys.* **3** 724
[48] Scott J G and Berger J O 2006 An exploration of aspects of Bayesian multiple testing *J. Stat. Plan. Inference* **136** 2144–62
[49] Boley M, Goldsmith B R, Ghiringhelli L M and Vreeken J 2017 Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery *Data Min. Knowl. Discov.* **31** 1391–418
[50] Ghosh K, Stuke A, Todorović M, Jørgensen P B, Schmidt M N, Vehtari A and Rinke P 2019 Deep learning spectroscopy: neural networks for molecular excitation spectra *Adv. Sci.* **6** 1801367
[51] Chandrasekaran A, Kamal D, Batra R, Kim C, Chen L and Ramprasad R 2019 Solving the electronic structure problem with machine learning *npj Comput. Mater.* **5** 22
[52] Westermayr J and Marquetand P 2021 Machine learning for electronically excited states of molecules *Chem. Rev.* **121** 9873
[53] Bağcioğlu M, Fricker M, Johler S and Ehling-Schulz M 2019 Detection and identification of *Bacillus cereus*, *Bacillus cytotoxicus*, *Bacillus thuringiensis*, *Bacillus mycoides* and *Bacillus weihenstephanensis* via machine learning based FTIR spectroscopy *Front. Microbiol.* **10** 902
[54] Rehman I U, Khan R S and Rehman S 2020 Role of artificial intelligence and vibrational spectroscopy in cancer diagnostics *Expert Rev. Mol. Diagn.* **20** 749–55
[55] Toyao T, Maeno Z, Takakusagi S, Kamachi T, Takigawa I and Shimizu K-i 2020 Machine learning for catalysis informatics: recent applications and prospects *ACS Catal.* **10** 2260–97
[56] Timoshenko J, Lu D, Lin Y and Frenkel A I 2017 Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles *J. Phys. Chem. Lett.* **8** 5091
[57] Cordova M, Balodis M, Simões de Almeida B, Ceriotti M and Emsley L 2021 Bayesian probabilistic assignment of chemical shifts in organic solids *Sci. Adv.* **7** eabk2341
[58] Ren H, Li H, Zhang Q, Liang L, Guo W, Huang F, Luo Y and Jiang J 2021 A machine learning vibrational spectroscopy protocol for spectrum prediction and spectrum-based structure recognition *Fundam. Res.* **1** 488
[59] Jinadasa M H W N, Kahawalage A C, Halstensen M, Skeie N O and Jens K J 2021 Deep learning approach for Raman spectroscopy *Recent Developments in Atomic Force Microscopy and Raman Spectroscopy for Materials Characterization* (London: IntechOpen)
[60] Zhao Y, Zhan K, Wang Z and Shen W 2021 Deep learning-based automatic detection of multitype defects in photovoltaic modules and application in real production line *Prog. Photovolt., Res. Appl.* **29** 471–84
[61] Zhaochun Z, Ruiwu P and Nianyi C 1998 Artificial neural network prediction of the band gap and melting point of binary and ternary compound semiconductors *Mater. Sci. Eng.* B **54** 149–52
[62] Paul S, Johann G, Henrik T and Reiner S 2000 Rapid access to infrared reference spectra of arbitrary organic compounds: scope and limitations of an approach to the simulation of infrared spectra by neural networks *Chem. Eur. J.* **6** 920–7
[63] Himanen L, Geurts A, Foster A S and Rinke P 2019 Data-driven materials science: status, challenges, and perspectives *Adv. Sci.* **6** 1900808
[64] Stuke A, Kunkel C, Golze D, Todorović M, Margraf J T, Reuter K, Rinke P and Oberhofer H 2020 Atomic structures and orbital energies of 61 489 crystal-forming organic molecules *Sci. Data* **7** 58
[65] Xian R P *et al* 2020 An open-source, end-to-end workflow for multidimensional photoemission spectroscopy *Sci. Data* **7** 442
[66] Pilania G, Gubernatis J E and Lookman T 2017 Multi-fidelity machine learning models for accurate bandgap predictions of solids *Comput. Mater. Sci.* **129** 156
[67] Perim E *et al* 2016 Spectral descriptors for bulk metallic glasses based on the thermodynamics of competing crystalline phases *Nat. Commun.* **7** 12315
[68] Ford D C, Hicks D, Oses C, Toher C and Curtarolo S 2019 Metallic glasses for biodegradable implants *Acta Mater.* **176** 297–305
[69] Ren F, Ward L, Williams T, Laws K J, Wolverton C, Hattrick-Simpers J and Mehta A 2018 Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments *Sci. Adv.* **4** eaaq1566
[70] Kusne A G *et al* 2020 On-the-fly closed-loop materials discovery via Bayesian active learning *Nat. Commun.* **11** 5966
[71] Oses C, Toher C and Curtarolo S 2020 High-entropy ceramics *Nat. Rev. Mater.* **5** 295–309
[72] Dasgupta A, Broderick S R, Mack C, Kota B U, Subramanian R, Setlur S, Govindaraju V and Rajan K 2019 Probabilistic assessment of glass forming ability rules for metallic glasses aided by automated analysis of phase diagrams *Sci. Rep.* **9** 357
[73] Sarker P, Harrington T, Toher C, Oses C, Samiee M, Maria J-P, Brenner D W, Vecchio K S and Curtarolo S 2018 High-entropy high-hardness metal carbides discovered by entropy descriptors *Nat. Commun.* **9** 4980
[74] Lederer Y, Toher C, Vecchio K S and Curtarolo S 2018 The search for high entropy alloys: a high-throughput *ab initio* approach *Acta Mater.* **159** 364–83
[75] Rickman J M, Chan H M, Harmer M P, Smeltzer J A, Marvel C J, Roy A and Balasubramanian G 2019 Materials informatics for the screening of multi-principal elements and high-entropy alloys *Nat. Commun.* **10** 2618
[76] Grabowski B, Ikeda Y, Srinivasan P, Körmann F, Freysoldt C, Duff A I, Shapeev A and Neugebauer J 2019 *Ab initio* vibrational free energies including anharmonicity for multicomponent alloys *npj Comput. Mater.* **5** 80
[77] Noé F, Tkatchenko A, Müller K-R and Clementi C 2020 Machine learning for molecular simulation *Ann. Rev. Phys. Chem.* **71** 361
[78] Anatole von Lilienfeld O, Müller K-R and Tkatchenko A 2020 Exploring chemical compound space with quantum-based machine learning *Nat. Rev. Chem.* **4** 347
[79] Deringer V L, Bernstein N, Bartók A P, Cliffe M J, Kerber R N, Marbella L E, Grey C P, Elliott S R and Csányi G 2018 Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics *J. Phys. Chem. Lett.* **9** 2879
[80] Chmiela S, Sauceda H E, Müller K-R and Tkatchenko A 2018 Towards exact molecular dynamics simulations with machine-learned force fields *Nat. Commun.* **9** 3887

[81] Schütt K T, Arbabzadah F, Chmiela S, Müller K-R and Tkatchenko A 2017 Quantum-chemical insights from deep tensor neural networks *Nat. Commun.* **8** 13890
[82] Zubatyuk R, Smith J S, Leszczynski J and Isayev O 2019 Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network *Sci. Adv.* **5** eaav6490
[83] Schütt K T, Gastegger M, Tkatchenko A, Müller K-R and Maurer R J 2019 Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions *Nat. Commun.* **10** 5024
[84] Segler M H S, Preuss M and Waller M P 2018 Planning chemical syntheses with deep neural networks and symbolic AI *Nature* **555** 604
[85] Wilkins D M, Grisafi A, Yang Y, Lao K U, DiStasio R A Jr and Ceriotti M 2019 Accurate molecular polarizabilities with coupled cluster theory and machine learning *Proc. Natl Acad. Sci. USA* **116** 3401
[86] Ercolessi F and Adams J B 1994 Interatomic potentials from first-principles calculations: the force-matching method *Europhys. Lett.* **26** 583
[87] Senftle T P *et al* 2016 The ReaxFF reactive force-field: development, applications and future directions *npj Comput. Mater.* **2** 15011
[88] Mackay D J C 2003 *Information Theory, Inference, and Learning Algorithms* (Cambridge: Cambridge University Press)
[89] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104** 136403
[90] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115
[91] Bartók A P, Kermode J, Bernstein N and Csányi G 2018 Machine learning a general-purpose interatomic potential for silicon *Phys. Rev.* X **8** 041048
[92] Vandermause J, Torrisi S B, Batzner S, Xie Y, Sun L, Kolpak A M and Kozinsky B 2020 On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events *npj Comput. Mater.* **6** 20
[93] Bauer M, van der Wilk M and Rasmussen C E 2016 Understanding probabilistic sparse Gaussian process approximations *Advances in Neural Information Processing Systems*
[94] Deringer V L, Caro M A and Csányi G 2020 A general-purpose machine-learning force field for bulk and nanostructured phosphorus *Nat. Commun.* **11** 5461
[95] Caro M A, Deringer V L, Koskinen J, Laurila T and Csányi G 2018 Growth mechanism and origin of high sp^3 content in tetrahedral amorphous carbon *Phys. Rev. Lett.* **120** 166101
[96] Manzhos S and Carrington T 2020 Neural network potential energy surfaces for small molecules and reactions *Chem. Rev.* **121** 10187–217
[97] Beck M, Jäckle A, Worth G A and Meyer H D 2000 The multiconfiguration time-dependent Hartree (MCTDH) method: a highly efficient algorithm for propagating wavepackets *Phys. Rep.* **324** 1–105
[98] Leclerc A and Carrington T 2014 Calculating vibrational spectra with sum of product basis functions without storing full-dimensional vectors or matrices *J. Chem. Phys.* **140** 174111
[99] Manzhos S and Carrington T 2006 Using neural networks to represent potential surfaces as sums of products *J. Chem. Phys.* **125** 194105
[100] Pradhan E and Brown A 2016 Vibrational energies for HFCO using a neural network sum of exponentials potential energy surface *J. Chem. Phys.* **144** 174305
[101] Majumder M, Hegger S E, Dawes R, Manzhos S, Wang X-G, Tucker C, Li J and Guo H 2015 Explicitly correlated MRCI-F12 potential energy surfaces for methane fit with several permutation invariant schemes and full-dimensional vibrational calculations *Mol. Phys.* **113** 1823–33
[102] Castro E, Avila G, Manzhos S, Agarwal J, Schaefer H F III and Carrington T 2017 Applying a Smolyak collocation method to Cl_2CO *Mol. Phys.* **115** 1775–85
[103] Kamath A, Vargas-Hernández R A, Krems R V, Carrington T and Manzhos S 2018 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** 241702
[104] Boussaidi M A, Ren O, Voytsekhovsky D and Manzhos S 2020 Random sampling high dimensional model representation Gaussian process regression (RS-HDMR-GPR) for multivariate function representation: application to molecular potential energy surfaces *J. Phys. Chem.* A **124** 7598–607
[105] Manzhos S, Wang X and Carrington T 2018 A multimode-like scheme for selecting the centers of Gaussian basis functions when computing vibrational spectra *Chem. Phys.* **509** 139–44
[106] Ku J, Kamath A, Carrington T and Manzhos S 2019 Machine learning optimization of the collocation point set for solving the Kohn–Sham equation *J. Phys. Chem.* A **123** 10631–42
[107] Schütt K T, Sauceda H E, Kindermans P-J, Tkatchenko A and Müller K-R 2018 SchNet—a deep learning architecture for molecules and materials *J. Chem. Phys.* **148** 241722
[108] Manzhos S, Chan M and Carrington T 2013 Communication: favorable dimensionality scaling of rectangular collocation with adaptable basis functions up to 7 dimensions *J. Chem. Phys.* **139** 051101
[109] Blank T B, Brown S D, Calhoun A W and Doren D J 1995 Neural network models of potential energy surfaces *J. Chem. Phys.* **103** 4129
[110] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces *Phys. Rev. Lett.* **98** 146401
[111] Artrith N, Morawietz T and Behler J 2011 High-dimensional neural-network potentials for multicomponent systems: applications to zinc oxide *Phys. Rev.* B **83** 153101
[112] Ghasemi S A, Hofstetter A, Saha S and Goedecker S 2015 Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network *Phys. Rev.* B **92** 045131
[113] Ko T W, Finkler J A, Goedecker S and Behler J 2021 A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer *Nat. Commun.* **12** 398
[114] Artrith N and Behler J 2012 High-dimensional neural network potentials for metal surfaces: a prototype study for copper *Phys. Rev.* B **85** 045439
[115] Podryabinkin E V and Shapeev A V 2017 Active learning of linearly parametrized interatomic potentials *Comput. Mater. Sci.* **140** 171
[116] Behler J 2011 Atom-centered symmetry functions for constructing high-dimensional neural network potentials *J. Chem. Phys.* **134** 074106

[117] Dral P O 2020 Quantum chemistry in the age of machine learning *J. Phys. Chem. Lett.* **11** 2336–47
[118] Smith J S, Isayev O and Roitberg A E 2017 ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost *Chem. Sci.* **8** 3192–203
[119] Smith J S, Nebgen B, Lubbers N, Isayev O and Roitberg A E 2018 Less is more: sampling chemical space with active learning *J. Chem. Phys.* **148** 24
[120] Devereux C, Smith J S, Huddleston K K, Barros K, Zubatyuk R, Isayev O and Roitberg A E 2020 Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens *J. Chem. Theory Comput.* **16** 4192–202
[121] Schütt K T, Sauceda H E, Kindermans P J, Tkatchenko A and Müller K R 2018 SchNet—a deep learning architecture for molecules and materials *J. Chem. Phys.* **148** 241722
[122] Lubbers N, Smith J S and Barros K 2018 Hierarchical modeling of molecular energies using a deep neural network *J. Chem. Phys.* **148** 241715
[123] Zubatyuk R, Smith J, Nebgen B T, Tretiak S and Isayev O 2021 Teaching a neural network to attach and detach electrons from molecules *Nat. Commun.* **12** 4870
[124] Zubatiuk T and Isayev O 2021 Development of multimodal machine learning potentials: toward a physics-aware artificial intelligence *Acc. Chem. Res.* **54** 1575–85
[125] Behler J 2016 Perspective: machine learning potentials for atomistic simulations *J. Chem. Phys.* **145** 170901
[126] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti C 2021 Physics-inspired structural representations for molecules and materials chemical reviews *Chem. Rev.* **121** 9759–815
[127] Anderson B, Hy T-S and Kondor R 2019 Cormorant: covariant molecular neural networks *Advances in Neural Information Processing Systems 32* ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox and R Garnett pp 14537–46 Curran Associates, Inc.
[128] Nigam J, Pozdnyakov S and Ceriotti M 2020 Recursive evaluation and iterative contraction of *N*-body equivariant features *J. Chem. Phys.* **153** 121101
[129] Fabrizio A, Grisafi A, Meyer B, Ceriotti M and Corminboeuf C 2019 Electron density learning of non-covalent systems *Chem. Sci.* **10** 9424–32
[130] Meyer R, Weichselbaum M and Hauser A W 2020 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.* **16** 5685–94
[131] Gastegger M, McSloy A, Luya M, Schütt K T and Maurer R J 2020 A deep neural network for molecular wave functions in quasi-atomic minimal basis representation *J. Chem. Phys.* **153** 044123
[132] Hermann J, Schätzle Z and Noé F 2020 Deep-neural-network solution of the electronic Schrödinger equation *Nat. Chem.* **12** 891–7
[133] Lopanitsyna N, Mahmoud C B and Ceriotti M 2021 Finite-temperature materials modeling from the quantum nuclei to the hot electrons regime *Phys. Rev. Mater.* **5** 043802
[134] Bereau T, DiStasio R A Jr, Tkatchenko A and Anatole von Lilienfeld O 2018 Non-covalent interactions across organic and biological subsets of chemical space: physics-based potentials parametrized from machine learning *J. Chem. Phys.* **148** 241706
[135] Veit M, Wilkins D M, Yang Y, DiStasio R A and Ceriotti M 2020 Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles *J. Chem. Phys.* **153** 024113
[136] Grisafi A, Nigam J and Ceriotti M 2021 Multi-scale approach for the prediction of atomic scale properties *Chem. Sci.* **12** 2078–90
[137] Westermayr J, Gastegger M, Schütt K T and Maurer R J 2021 Deep integration of machine learning into computational chemistry and materials science (arXiv:2102.08435)
[138] Welborn M, Cheng L and Miller T F 2018 Transferability in machine learning for electronic structure via the molecular orbital basis *J. Chem. Theory Comput.* **14** 4772–9
[139] Townsend J and Vogiatzis K D 2019 Data-driven acceleration of the coupled-cluster singles and doubles iterative solver *J. Phys. Chem. Lett.* **10** 4129–35
[140] Li H, Collins C, Tanha M, Gordon G J and Yaron D J 2018 A density functional tight binding layer for deep learning of chemical Hamiltonians *J. Chem. Theory Comput.* **14** 5764–76
[141] Wang Z, Ye S, Wang H, He J, Huang Q and Chang S 2021 Machine learning method for tight-binding Hamiltonian parameterization from *ab initio* band structure *npj Comput. Mater.* **7** 11
[142] Stöhr M, Medrano Sandonas L and Tkatchenko A 2020 Accurate many-body repulsive potentials for density-functional tightbinding from deep tensor neural networks *J. Phys. Chem. Lett.* **11** 6835–43
[143] Westermayr J and Marquetand P 2020 Machine learning for electronically excited states *Chem. Rev.* (accepted) https://doi.org/10.1021/acs.chemrev.0c00749
[144] Thomas L H 1927 The calculation of atomic fields *Math. Proc. Camb. Phil. Soc.* **23** 542
[145] Kohn W and Sham L J 1965 Self-Consistent equations including exchange and correlation effects *Phys. Rev.* **140** A1133
[146] Perdew J P, Ruzsinszky A, Jianwei S, Nepal N K and Kaplan A D 2021 Interpretations of ground-state symmetry breaking and strong correlation in wavefunction and density functional theories *Proc. Natl Acad. Sci. USA* **118** e2017850118
[147] Stoudenmire E M, Wagner L O, White S R and Burke K 2012 One-dimensional continuum electronic structure with the density-matrix renormalization group and its implications for density-functional theory *Phys. Rev. Lett.* **109** 056402
[148] Tozer D J, Ingamells V E and Handy N C 1996 Exchange–correlation potentials *J. Chem. Phys.* **105** 9200
[149] Mortensen J J, Kaasbjerg K, Frederiksen S L, Nørskov J K, Sethna J P and Jacobsen K W 2005 Bayesian error estimation in density-functional theory *Phys. Rev. Lett.* **95** 216401
[150] Snyder J C, Rupp M, Hansen K, Müller K-R and Burke K 2012 Finding density functionals with machine learning *Phys. Rev. Lett.* **108** 253002
[151] Nagai R, Akashi R and Sugino O 2020 Completing density functional theory by machine learning hidden messages from molecules *npj Comput. Mater.* **6** 43
[152] Li L, Hoyer S, Pederson R, Sun R, Cubuk E D, Riley P and Burke K 2021 Kohn–Sham equations as regularizer: building prior knowledge into machine-learned physics *Phys. Rev. Lett.* **126** 036401
[153] Brockherde F, Vogt L, Li L, Tuckerman M E, Burke K and Müller K-R 2017 Bypassing the Kohn–Sham equations with machine learning *Nat. Commun.* **8** 872
[154] Dick S and Fernandez-Serra M 2020 Machine learning accurate exchange and correlation functionals of the electronic density *Nat. Commun.* **11** 3509
[155] Kasim M F and Vinko S M 2021 Learning the exchange–correlation functional from nature with fully differentiable density functional theory *Phys. Rev. Lett.* **127** 126403

[156] Lu D, Wang H, Chen M, Lin L, Car R, E W, Jia W, Zhang L and Zhang L 2021 86 PFLOPS deep potential molecular dynamics simulation of 100 million atoms with *ab initio* accuracy *Comput. Phys. Commun.* **259** 107624
[157] Mardirossian N and Head-Gordon M 2017 *Mol. Phys.* **115** 2315–72
[158] Gillan M J, Alfè D and Michaelides A 2016 *J. Chem. Phys.* **144** 130901
[159] Nagai R, Akashi R, Sasaki S and Tsuneyuki S 2018 *J. Chem. Phys.* **148** 241737
[160] Schmidt J, Benavides-Riveros C L and Marques M A L 2019 *J. Phys. Chem. Lett.* **10** 6425–31
[161] Kanungo B, Zimmerman P M and Gavini V 2019 *Nat. Commun.* **10** 4497
[162] Lin S-C, Martius G and Oettel M 2020 *J. Chem. Phys.* **152** 021102
[163] Carleo G and Troyer M 2017 *Science* **355** 602–6
[164] Luo D and Clark B K 2019 *Phys. Rev. Lett.* **122** 226401
[165] Han J, Zhang L and E W 2019 *J. Comput. Phys.* **399** 108929
[166] Hermann J, Schätzle Z and Noé F 2020 *Nat. Chem.* **12** 891–7
[167] Pfau D, Spencer J S, Matthews A G D G and Foulkes W M C 2020 *Phys. Rev. Res.* **2** 033429
[168] Al-Hamdani Y S, Nagy P R, Zen A, Barton D, Kállay M, Brandenburg J G and Tkatchenko A 2021 *Nat. Commun.* **12** 3927
[169] Schätzle Z, Hermann J and Noé F 2021 *J. Chem. Phys.* **154** 124108
[170] Hutter M 2020 On representing (anti)symmetric functions (arXiv:2007.15298)
[171] Choo K, Mezzacapo A and Carleo G 2020 *Nat. Commun.* **11** 2368
[172] Mills K, Spanner M and Tamblyn I 2017 Deep learning and the Schrödinger equation *Phys. Rev.* A **96** 042113
[173] Pilati S and Pieri P 2019 Supervised machine learning of ultracold atoms with speckle disorder *Sci. Rep.* **9** 5613
[174] Carleo G and Troyer M 2017 Solving the quantum many-body problem with artificial neural networks *Science* **355** 602
[175] Pilati S and Pieri P 2020 Simulating disordered quantum Ising chains via dense and sparse restricted Boltzmann machines *Phys. Rev.* E **101** 063308
[176] Ryczko K, Strubbe D A and Tamblyn I 2019 Deep learning and density-functional theory *Phys. Rev.* A **100** 022512
[177] Denner M M, Fischer M H and Neupert T 2020 Efficient learning of a one-dimensional density functional theory *Phys. Rev. Res.* **2** 033388
[178] Nelson J, Tiwari R and Sanvito S 2019 Machine learning density functional theory for the Hubbard model *Phys. Rev.* B **99** 075132
[179] Mills K, Ryczko K, Luchak I, Domurad A, Beeler C and Tamblyn I 2019 Extensive deep neural networks for transferring small scale learning to large scale systems *Chem. Sci.* **10** 4129
[180] Saraceni N, Cantori S and Pilati S 2020 Scalable neural networks for the efficient learning of disordered quantum systems *Phys. Rev.* E **102** 033301
[181] Mujal P, Marques M A L, Polls A, Juliá-Díaz B and Pilati S 2021 Supervised learning of few dirty bosons with variable particle number *SciPost Phys.* **10** 073
[182] Ryczko K, Wetzel S J, Melko R G and Tamblyn I 2021 Orbital-free density functional theory with small datasets and deep learning *J. Chem. Theory Comput.* **18** 1122–8
[183] Draxl C and Scheffler M 2019 The NOMAD laboratory: from data sharing to artificial intelligence *J. Phys. Mater.* **2** 036001
[184] Wilkinson M D *et al* 2016 The FAIR guiding principles for scientific data management and stewardship *Sci. Data* **3** 160018
[185] Lejaeghere K *et al* 2016 Reproducibility in density-functional theory calculations of solids *Science* **351** aad3000
[186] Gulans A, Kozhevnikov A and Draxl C 2018 Microhartree precision in density functional theory calculations *Phys. Rev.* B **97** 161105(R)
Jensen S R, Saha S, Flores-Livas J A, Huhn W, Blum V, Goedecker S and Frediani L 2017 The elephant in the room of density functional theory calculations *J. Phys. Chem. Lett.* **8** 1449
[187] Nabok D, Gulans A and Draxl C 2016 Accurate all-electron G_0W_0 quasiparticle energies employing the full-potential augmented planewave method *Phys. Rev.* B **94** 035418
Rangel T *et al* 2020 Reproducibility in G_0W_0 calculations for solids *Comput. Phys. Commun.* **255** 107242
[188] Gulans A and Draxl C 2022 Influence of spin–orbit coupling on chemical bonding (arXiv:2204.02751)
[189] Carbogno C *et al* 2020 Numerical quality control for DFT-based materials databases *Npj Comput. Mater.* (arXiv:2008.10402)
[190] Kuban M, Gabaj S, Aggoune W, Vona C, Rigamonti S and Draxl C 2022 Similarity of materials and data-quality assessment by fingerprinting (arXiv:2204.04056)
[191] Scheffler M *et al* 2022 FAIR data enabling new horizons for materials research *Nature* **604** 635–42
[192] Isayev O, Fourches D, Muratov E N, Oses C, Rasch K, Tropsha A and Curtarolo S 2015 Materials cartography: representing and mining materials space using structural and electronic fingerprints *Chem. Mater.* **27** 735
[193] Kuban M, Rigamonti S, Scheidgen M and Draxl C 2022 Density-of-states similarity descriptor for unsupervised learning from materials data (arXiv:2201.02187)
[194] Data from the NOMAD Laboratory [187]; exciting data: 10.17172/NOMAD/2018.08.22, PID 553302; VASP data: PID 69356; FHI-aims: 10.17172/NOMAD/2018.10.05-1, PID: 5842781.
[195] Toher C *et al* 2018 The AFLOW fleet for materials discovery *Handbook of Materials Modeling* ed W Andreoni and S Yip (Berlin: Springer) pp 1–28
[196] Hicks D, Mehl M J, Gossett E, Toher C, Levy O, Hanson R M, Hart G and Curtarolo S 2019 The AFLOW library of crystallographic prototypes: part 2 *Comput. Mater. Sci.* **161** S1–1011
[197] Hicks D, Oses C, Gossett E, Gomez G, Taylor R H, Toher C, Mehl M J, Levy O and Curtarolo S 2018 AFLOW-SYM: platform for the complete, automatic and self-consistent symmetry analysis of crystals *Acta Crystallogr.* A **74** 184–203
[198] Oses C *et al* 2018 AFLOW-CHULL: cloud-oriented platform for autonomous phase stability analysis *J. Chem. Inf. Model.* **58** 2477–90
[199] Isayev O, Oses C, Toher C, Gossett E, Curtarolo S and Tropsha A 2017 Universal fragment descriptors for predicting properties of inorganic crystals *Nat. Commun.* **8** 15679
[200] Gossett E *et al* 2018 AFLOW-ML: a RESTful API for machine-learning predictions of materials properties *Comput. Mater. Sci.* **152** 134–45
[201] Rose F, Toher C, Gossett E, Oses C, Buongiorno Nardelli M, Fornari M and Curtarolo S 2017 AFLUX: the LUX materials search API for the AFLOW data repositories *Comput. Mater. Sci.* **137** 362–70
[202] Hicks D, Toher C, Ford D C, Rose F, De Santo C, Levy O, Mehl M J and Curtarolo S 2021 AFLOW-XtalFinder: a reliable choice to identify crystalline prototypes *npj Comput. Mater.* **7** 30
[203] Friedrich R, Usanmaz D, Oses C, Supka A, Fornari M, Buongiorno Nardelli M, Toher C and Curtarolo S 2019 Coordination corrected *ab initio* formation enthalpies *npj Comput. Mater.* **5** 59

[204] Yang K, Oses C and Curtarolo S 2016 Modeling off-stoichiometry materials with a high-throughput *ab initio* approach *Chem. Mater.* **28** 6484–92
[205] Balachandran P V 2020 Adaptive machine learning for efficient materials design *MRS Bull.* **45** 579–86
[206] Saal J E, Oliynyk A O and Meredig B 2020 Machine learning in materials discovery: confirmed predictions and their underlying approaches *Annu. Rev. Mater. Res.* **50** 49–69
[207] Lookman T, Balachandran P V, Xue D and Yuan R 2019 Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design *npj Comput. Mater.* **5** 180
[208] Balachandran P V, Xue D, Theiler J, Hogden J and Lookman T 2016 Adaptive strategies for materials design using uncertainties *Sci. Rep.* **6** 19660
[209] Settles B 2012 Active learning *Synthesis Lectures on Artificial Intelligence and Machine Learning* vol 6 pp 1–114
[210] Yu M, Yang S, Wu C and Marom N 2020 Machine learning the Hubbard U parameter in DFT + U using Bayesian optimization *npj Comput. Mater.* **6** 180
[211] Herbol H C, Hu W, Frazier P, Clancy P and Poloczek M 2018 Efficient search of compositional space for hybrid organic-inorganic perovskites via Bayesian optimization *npj Comput. Mater.* **4** 51
[212] Ju S, Shiga T, Feng L, Hou Z, Tsuda K and Shiomi J 2017 Designing nanostructures for phonon transport via Bayesian optimization *Phys. Rev.* X **7** 021024
[213] Wolpert D H 2002 The supervised learning no-free-lunch theorems *Soft Computing and Industry* ed R Roy, M Köppen, S Ovaska, T Furuhashi and F Hoffmann (Berlin: Springer) pp 25–42
[214] Balachandran P V, Kowalski B, Sehirlioglu A and Lookman T 2018 Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning *Nat. Commun.* **9** 1668
[215] Henderson A N, Kauwe S K and Sparks T D 2021 Benchmark datasets incorporating diverse tasks, sample sizes, material systems, and data heterogeneity for materials informatics *Data Brief* **37** 107262
[216] Sutton R S and Barto A G 2018 *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press)
[217] Bellemare M G, Candido S, Castro P S, Gong J, Machado M C, Moitra S, Ponda S S and Wang Z 2020 *Nature* **588** 77–82
[218] Gottipati S K *et al* 2020 (arXiv:2010.03744)
[219] Simm G N C, Pinsler R and Miguel Hernández-Lobato J 2020 *Proc. 37th Int. Conf. Machine Learning* (Vienna, Austria)
[220] Gottipati S K *et al* 2020 (arXiv:2004.12485)
[221] Thiede L A, Krenn M, Nigam A and Aspuru-Guzik A 2020 (arXiv:2012.11293)
[222] Gaudin T, Nigam A and Aspuru-Guzik A 2019 *2nd Workshop on Machine Learning and the Physical Sciences NeurIPS* (Vancouver, Canada)
[223] Kober J, Bagnell J A and Peters J 2013 *Int. J. Robot. Res.* **32** 1238–74
[224] Yang Y 2018 *Proc. 27th Int. Joint Conf. Artificial Intelligence (IJCAI-18)* pp 5739–43
[225] Zhao W, Peña Queralta J and Westerlund T 2020 arXiv:2009.13303
[226] Lipton Z C 2018 The mythos of model interpretability *Queue* **16** 31–57
[227] Murdoch W J, Singh C, Kumbier K, Abbasi-Asl R and Yu B 2019 Definitions, methods, and applications in interpretable machine learning *Proc. Natl Acad. Sci. USA* **116** 22071–80
[228] Ribana R, Bohn B, Duarte M F and Garcke J 2020 Explainable machine learning for scientific insights and discoveries *IEEE Access* **8** 42200–16
[229] Doshi-Velez F and Kim B 2017 Towards a rigorous science of interpretable machine learning (arXiv:1702.08608)
[230] Gilpin L H, Bau D, Yuan B Z, Bajwa A, Specter M and Kagal L 2018 Explaining explanations: an overview of interpretability of machine learning *2018 IEEE 5th Int. Conf. Data Science and Advanced Analytics (DSAA)* (IEEE) pp 80–9
[231] Carvalho D V, Pereira E M and Cardoso. J S 2019 Machine learning interpretability: a survey on methods and metrics *Electronics* **8** 832
[232] Nori H, Jenkins S, Koch P and Caruana R 2019 Interpretml: a unified framework for machine learning interpretability (arXiv:1909.09223)
[233] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. Roy. Stat. Soc.* B **58** 267–88
[234] Wang Y, Wagner N and Rondinelli J M 2019 Symbolic regression in materials science *MRS Commun.* **9** 793–805
[235] Sutton C, Boley M, Ghiringhelli L M, Rupp M, Vreeken J and Scheffler M 2020 Identifying domains of applicability of machine learning models for materials science *Nat. Commun.* **11** 4428