Towards self-driving laboratories in chemistry and materials sciences: The central role of DFT in the era of AI

Bing Huang,[1, *] Guido Falk von Rudorff,[2, 3, †] and O. Anatole von Lilienfeld[4, 5, 6, ‡]

[1] University of Vienna, Faculty of Physics, Kolingasse 1416, AT1090 Wien, Austria
[2] University Kassel, Department of Chemistry, Heinrich-Plett-Str.40, 34132 Kassel, Germany
[3] Center for Interdisciplinary Nanostructure Science and Technology (CINSaT), Heinrich-Plett-Straße 40, 34132 Kassel
[4] Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada
[5] Departments of Chemistry, Materials Science and Engineering, and Physics,
University of Toronto, St. George Campus, Toronto, ON, Canada
[6] Machine Learning Group, Technische Universität Berlin and Berlin
Institute for the Foundations of Learning and Data, 10587 Berlin, Germany

Density functional theory plays a pivotal role for the chemical and materials science due to its relatively high predictive power, applicability, versatility and low computational cost. We review recent progress in machine learning model developments, which has relied heavily on density functional theory for synthetic data generation and model architecture, and provide some broader context for its general relevance to the chemical sciences. Resulting in models with high efficiency, accuracy, scalability, and transferability (EAST), these developments will pave the way for the routine use of successful experimental planning software within self-driving laboratories.

INTRODUCTION

We undoubtedly live in an era of artificial intelligence (AI). By now AI has touched upon and affected almost any branch of human activity, assuming centre stage in many domains of daily life, such as autonomous flight navigation or self-driving cars. While also the humanities and natural sciences have already benefited strongly from machine learning research in many domains, AI based robotic experimentation for chemistry and materials is still in its infancy. Promising first steps with regards to robotic and autonomous experimentation only made most recently[1–6], e.g. delivering self-driving laboratories for thin film discoveries [7]). By comparison, within synthetic biology the dream of an AI based robot scientist (named 'Adam') to assist and accelerate scientific discovery has been introduced already nearly twenty years ago [8–10]. Such ground-breaking progress has led Krenn et al . to survey community members and to fundamentally reconsider the meaning of understanding in the context of the scientific process itself [11]. As already outlined previously by Aspuru-Guzik, Lindh, and Reiher [12], the success of autonomous self-driving labs in chemistry and materials will depend crucially on the availability of control software capable to reliably forecast and rank experimental outcomes with sufficient accuracy in real time. Machine learning models trained across the relevant chemical and materials spaces [13], as well as corresponding reaction processes offer such capability, and will additionally improve (through frequent updating) as increasingly more data is being accumulated through subsequent iterations being performed by the robot. Consequently, it seems self-evident that there is an urgent need for the seamless integration of predictive machine learning based models of materials and chemistry properties.

The computational design and discovery of materials and molecules represents a long-standing challenge to the atomistic simulation community and has motivated decades of research [14–16] Applications are as diverse as the chemical sciences and include improved solutions for batteries, transistors, catalysts, coatings, ligands, or photo-voltaics, among others. All such efforts have in common that they attempt to virtually navigate chemical compound space (CCS) in order to narrow down the search space for subsequent experimental verification and characterization. CCS refers to the tremendously large set that emerges for all conceivable stable combinations of chemical elements and molecular or periodic structures [17]. Thermodynamic and kinetic stability being well defined via the quantum statistical mechanics of electrons and nuclei, reliance on a quantum mechanics based approach towards CCS is as obvious as *alternativlos* [18]. Unfortunately, the relevant equations of quantum and statistical mechanics can only be solved exactly for the simplest of systems, rendering numerical solutions of approximate expressions necessary [19]. With even tiny domains of CCS already containing inaccessibly many members, e.g. 10^{60} pharmacologically active molecules based on Lipinski's rule of five[20] or 10^{14} binary materials of $3 \times 3 \times 3$ hexagonally close-packed cells[21], applying quantum chemistry (QC) methods of any complexity is infeasible. Additionally, with few exceptions such as in Ref. [22, 23], conventional quantum chemistry based compute campaigns in CCS consider every system independently with little transfer of information [24, 25].

The importance of electronic structure information for computational materials identification, characterization and optimization has recently been highlighted by Marzari, Ferretti, and Wolverton [26] The probably most powerful compromise between predictive power and computational burden for calculating properties and behav-
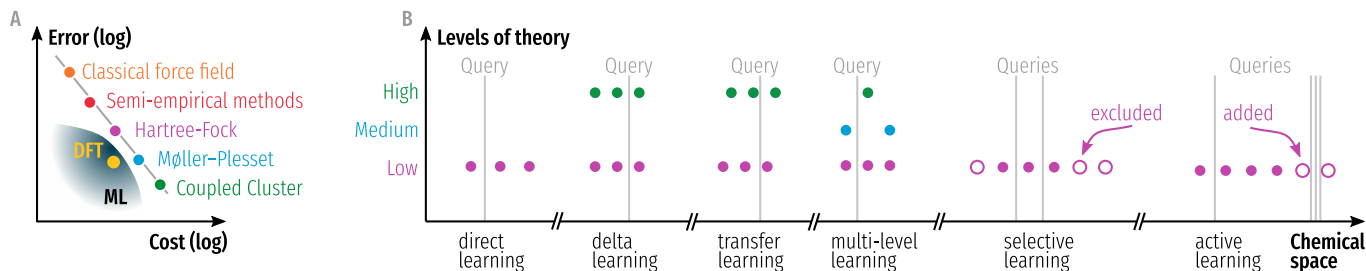
Figure 1. Approaches to sample chemical space that account for finite compute budget. A: Pareto plot of computational simulation approaches. DFT, at the comparably low mean-field cost of Hartree-Fock, often comes close in accuracy to methods with explicit electron correlation. B: ML model strategies for transferability, possibly including reference data (points) from different levels of theory given a set of queries (grey lines). Direct learning refers to all data points from one level, while delta and trasnfer learning requires knowing the same compound on different levels. Multi-level learning integrates data from multiple levels. Selective learning only include data points (filled) close to the queries and ignores all others (empty). Active learning biases the data collection (empty) to improve prediction accuracy.

ior of gaseous and condensed systems from first principles is Density Functional Theory (DFT). In particular, the effective single-particle flavour of DFT, approximating the electronic kinetic energy contribution within the Kohn-Sham framework [27] has proven immensely useful to the entire community. It has been enabled by a sheer countless number of ever improving approximations to the exchange-correlation potential [28–31]. Effectively, these contributions have dramatically pushed down the Pareto front which defines the common trade-off between computational cost and predictive accuracy, see Fig. 1 (A)). With one of the co-founders of DFT, Walter Kohn, having been awarded the Nobel Prize in Chemistry in 1998, expectations for further improvements ran high with the turn of the century [28], and during the widespread adaption of time dependent DFT to also treat electronic excitations [32]. As such, it was not too surprising to find two DFT related contributions featured among the top 10 papers of all times as highlighted in *Nature* in 2014 [33]. Ample contemporary studies have led to further improvements [30, 34], highlighted the importance of numerical reproducibility of properties of solids with respect to implementation details [35], or emphasized the role of using the electron density as a measure of quality, rather than the energy [36] (see Fig. 2 (D)).

Accounting for all of the underpinning physics of materials and chemistry, the electronic structure model even when solved using DFT still represents the computational bottleneck among all the relevant governing equations in quantum and statistical mechanics. In fact, it imposes such severe computational complexity that even when equipped with modern high-performance computing hardware, DFT based computational chemistry and materials protocols have largely failed to become modern industry wide standards to the extend that they could replace experimentation ('digital twins'). Correspondingly, finding effective approximations to further reduce the computational burden has constituted a holy

grail for many decades. Over the last decade, however, physics based supervised machine learning (ML) techniques, amenable to the sampling of quantum observables throughout CCS, have experienced rapid and accelerating growth within the chemical and materials sciences. By now a considerable number of special issues has appeared in the peer-reviewed literature, including *Int. J. Quantum Chem.* (2015) [39], *J. Chem. Phys.* (2018) [40], *J. Phys. Chem.* (2018) [41], *J. Phys. Chem. Lett.* (2020) [42], *Nature Communications* (2020) [43], *J. Chem. Phys.* (2021) [44] *Chem. Rev.* (2021) [45] Various other overview have also been published [13, 25, 46?–60]. The encouraging success of ML in this domain has undoubtedly also played an instrumental role in the decision to create new journals, such as Springer's *Nature Machine Intelligence*, IOP's *Machine Learning: Science and Technology* [61], or Wiley's *Applied Artificial Intelligence Letters* [62].

We believe that it is difficult to overstate the general importance of these developments. In particular, as also argued previously [48], the emergence of useful statistical surrogate models, such as manifested by machine learning, corresponds in fact to the formation of a fourth pillar in the hard sciences. This notion is universally applicable, i.e. going even beyond just the chemical and materials sciences. More specifically, first, second, third, and forth pillar respectively correspond to manual experimentation, conception and derivation of theoretical framework to predict experimental observables, development and implementation of numerical simulation tools that assist the predictions by solving computationally complex equations of the theoretical framework, and statistical learning approaches that exploit experimental or simulated data in order to infer simulated or experimental measurements. These pillars clearly build onto each other, and DFT can be seen as bridging them, going all the way from the experimentally observable electron probability distribution via the Hohenberg-Kohn theo-
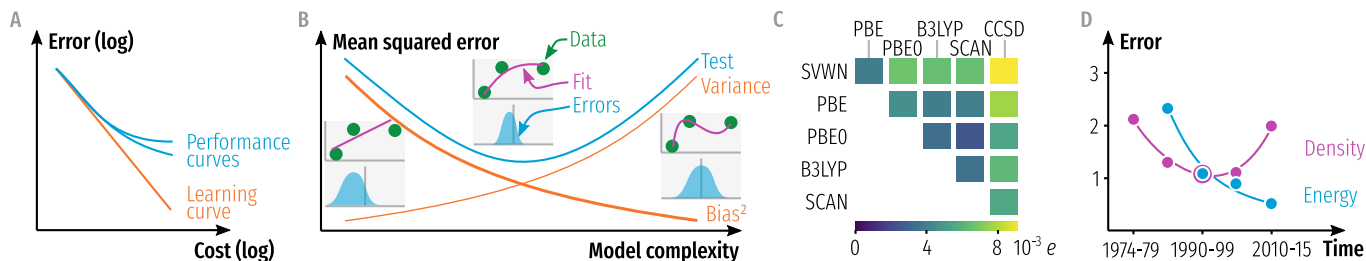
Figure 2. Challenges in multi-objective machine learning models. A: Learning curves are guaranteed to decay with a power law in the limit of large data sets[37], while performance curves, i.e. metrics not included in the loss function, typically do not improve arbitrarily. B: The prediction ("test") error is the combination of a bias from the flexibility of the model to follow the data points and the variance arising from the model flexibility between data points. C: Integrated electron density differences, a common measure for density accuracy, for established quantum chemistry methods averaged over a random subset of QMrxn20 reactants[38]. D: DFT functionals improve on the energy with time but at the same time yield lower accuracy for the electron density (average median-normalised error from literature[36]).
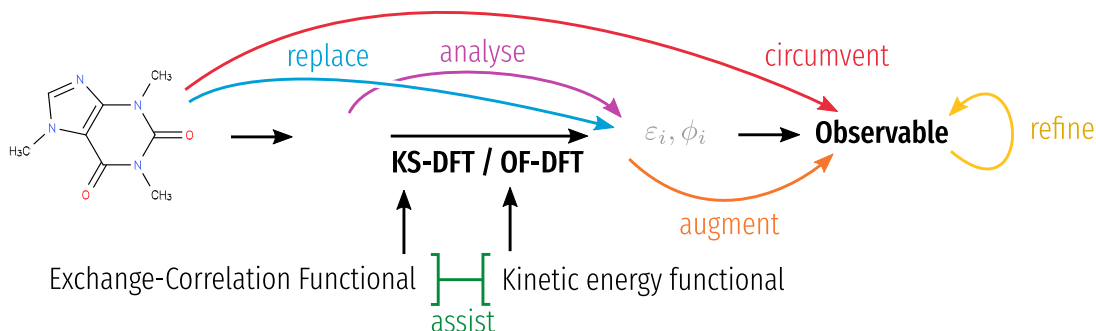

Figure 3. Classification of machine learning approaches going beyond DFT. The traditional workflow in DFT from a molecule via a Hamiltonian to the electron density ρ and, for Kohn-Sham-DFT, the orbitals ϕ_i and their energies ϵ_i is shown in black while the colored elements show routes taken to build upon DFT or to replace it all together.

rems and the Kohn-Sham Ansatz via the many numerical implementations and hardware use cases to its use for generating crucial training data information and informing the design of data-efficient ML model architectures. Tor the sake of brevity we will be taking the importance of DFT for pillars 1-3 as a given, and will focus solely on its role for ML. In particular, Fig. 3 illustrates just some of the possible ways that have been used to make use of DFT within ML models.

In order to further substantiate the claim towards the key role DFT is playing for the fourth pillar of science, we will now review many of the specific ML contributions that have strongly benefited from DFT. To this end, we have opted to structure this review according to four categories, Efficiency, Accuracy, Scalability, and Transferability (EAST), see Fig. 4. EAST components represent an intuitive ordering principle which allows us to meaningful structure, distinguish, and compare some of the most important features that would be necessary for building and using digital twins within the chemical and materials sciences.

EFFICIENCY

Comparing to quantum chemistry (QC) calculations, perhaps the most striking feature of physics based quantum ML models is their prediction speed. Though both approaches start from purely the atomic composition and geometry, a trained ML model makes prediction out of simple linear algebra (i.e., matrix operations) that can be performed rather efficiently (e.g., milli-seconds for a medium-sized organic molecule), in particular when aided by modern GPU architecture. In contrast, QC calculations involve the computation of electron repulsion integrals, and solving complex iterative mathematical equations, both of which can be (highly) non-linear and much more computationally intensive and time-consuming, e.g. DFT calculations at the hybrid level for a medium-sized organic molecule in a decent basis can easily require many CPU minutes.

In order to obtain a complete understanding of the computational efficiency of ML models, it is essential to take into account the associated costs of both, training and testing. In fact, there are several factors that can contribute to the efficiency in training (as quantified through the learning curves shown in Fig. 2 (A)) a)
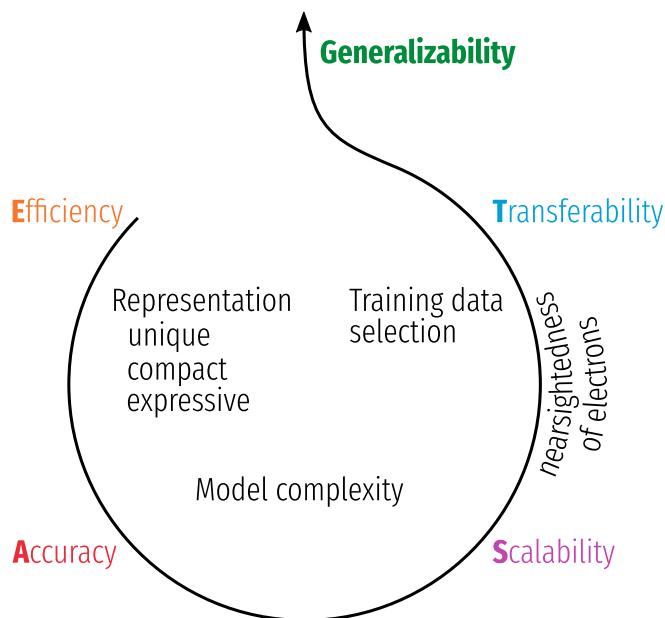
Figure 4. The four key categories of predictive ML models of chemical and materials properties and processes: E (Efficiency), A (Accuracy), S (Scalability) and T (Transferability). A model of EAST generalizes well to unseen systems. Factors that impact EAST include the choice of representation, model complexity and training data selection (from the ML side, shown in the inner circle), as well as the nearsightedness of electronic system, which plays a more fundamental role in determining the S (long-range) and T (short-range) of local atomic environments across different molecular systems.

the design of the representation $\mathbf{X}$ that serves as the ML model input. A data-efficient representation should be unique (in the sense of a bijective one-to-one relationship to the molecular geometry), compact (small number of constituent values), and expressive, i.e. being capable of accurate description of the molecules for training and test; b) the ML model complexity, roughly proportional to the number of parameters in the model. Typically, efficiency decreases when the model complexity grows; c) the number of training instances (data set size and atom count). Non-parametric ML models, e.g. based on Gaussian Process Regression, become less efficient as the number and the size of the training compounds increase. When it comes to querying a model, both the representation and the model complexity affect efficiency. Note that typically the number of training points does not impact the prediction efficiency of neural networks of fixed architecture (parametric ML models). Recent kernel ridge regression based developments indicate superior computational efficiency when compared to neural networks [63, 64]

A machine learning model's high efficiency can sometimes result in lower accuracy. For instance, kernel ridge regression based ML models trained on less data are more efficient but less accurate than those trained with more data. To reduce the cost in training without sacrificing accuracy, one can exploit correlations between different quantum-chemistry methods, e.g. by using a multi-level approaches: This method is characterized by the use of few high-fidelity data points (e.g. double hybrid DFT functionals) for small systems and many low-fidelity points (e.g. Local Density Approximation functionals) for large systems (see Fig. 5). While most of the multi-level approaches in literature implicitly only consider the two subspaces training data and fidelity, there is potential in including more dimensions. For example, the combination technique based quantum machine learning (CQML) model [65] can effectively include electron correlation levels, basis set size, and randomly sampled chemical space, as long as a formally strict hierarchy exists along each of those dimensions. Including a hierarchical dimension in chemical space such as atom-in-molecule-on (*amon* i.e., molecular fragments made up of typically ≤ 8 non-hydrogen atoms, obtained through systematic fragmentation [66, 67]) allows to include more data and improves efficiency [68]. Due to the finite number (and therefore enumerable) and hierarchical nature of amons, the curse of dimensionality, which originates from the combinatorial nature of chemical compound space, can be mitigated, and therefore the training efficiency of ML models can be enhanced.

Within the realm of DFT, unifying amons and hierarchical density functionals, as demonstrated for the widely-known Jacob's ladder (see Fig. 5), one can greatly improve model efficiency with a set of extra, low-cost calculations (e.g., LDA or GGA). This concept can also be extended to other QC models, with or without the blending of semi-empirical QC methods [65, 68–71]. Consequently, mixing DFT with other high-level electron correlation models (e.g., CCSD(T), GW, or QMC) within the framework of CQML, is a very promising strategy [68], in particular for construction of large-scale data sets of potential energy surfaces (PES), due to the outstanding trade-off between efficiency and accuracy of DFT methods. In addition to efforts to make DFT property-driven ML more efficient, another promising direction for significant cost reduction is to use ML to help with the improvement of orbital-free DFT (OF-DFT) methods. The core idea of OF-DFT is to remove the explicit dependence on orbitals for the kinetic energy term in KS-DFT by machine-learning the kinetic energy density functional[72] directly from data. 3

Besides computational efficiency, training data efficiency is critical since typically, data is scare in the materials and chemical sciences. There is statistical model uncertainty on the one side but there is also, and independently, the danger of unconscious bias in data leading to severe artefacts and catastrophic forecasting errors. In both cases, predictive accuracy needs to come from being close to reference calculations, which will hardly

ever converge to be fully representative of all chemical space. In fact, sampling representative data is a substantial challenge to global accuracy for small parts of chemical space already[73]. Without breaking down the scaling of CCS, it therefore seems unrealistic that it is possible to obtain one model for all of chemical space. As such, there will always be need for DFT as a pillar of reference and thanks to its universality. In the context of machine learning however, data-driven training avoids human bias in data sets[74]. While DFT is a universally applicable theory and it is often treated as such, empirically fitting functionals will only cover a vanishingly small part of chemical space which often is selected based on heuristics or needs. Active learning (Fig. 1B) selects the training data[67, 75] that resemble the target the most in an attempt to deal with the scaling of compound space[76] and is closely related to on-the-fly training[77], thus steepening the learning curve and thereby increasing data efficiency.

ACCURACY

At the heart of ML is the promise that more training data affords a more accurate model. It is intuitive that, provided the avoidance of overfitting and extrapolation, any sufficiently flexible regressor will improve down to infinitesimally small prediction errors in the limit of infinite training data. According to the central limit theorem, the standard deviation of the error distribution decays with the inverse square root of sample number. It has also been shown for many applications that the leading term of prediction error decays according to an inverse power law for kernel methods as well as neural networks.[37, 78] For these reasons *learning curves* (see Fig 2A) are most conveniently presented on log-log scales resulting in linearly decaying prediction errors as a function of training set size. Such learning curves are helpful tools both to compare models regarding their data efficiency (as manifested in the slope) or their required total training points (then the curves meet an accuracy target). Fitting the performance of just a few trained models to the log-log scale with negative slope and off-set as power law exponent and logarithm of proportionality constant, respectively, is often readily possible and can be used for quantitative comparison among ML models with diverse origins. The systematic improvement of a model with training data however is only guaranteed for the learned label alone. Derived properties might or might not improve coincidentally, as illustrated in *performance curves* (see Fig. 2A). This bears some analogy to DFT functionals where systematic improvements in accuracy with respect to energies does not necessarily imply that also electron densities will improve: They can also be obtained at the expense of a deteriorating quality of the corresponding electron density[36] (see Fig. 2D for an il-

lustration). Note the importance of sufficient flexibility and converged cross-validation in order to achieve systematic without overfitting, see Fig. 2B.

ML models become more accurate (at constant training set size) when hard requirements and boundary conditions are accounted for, e.g. as exploited by DM21 reducing delocalisation errors[79] or by BAML adding interatomic three and four-body interactions to the representation [80]. Similar ideas have been successful in the DFT context where SCAN employs known constraints a DFT functional needs to satisfy.[34] Commonly, this is referred to as "including more physics" (or physics-informed), meaning making models implicitly aware of symmetries, invariances (such as atom permutation and geometry rotation), scaling laws, and other physical constraints. A major challenge is to not only make the model accurate in most cases but rather to detect when the results might be unreliable by quantifying uncertainty[81], especially since the residual errors for common tasks are distinctly non-normal[82]. DFT has been successful despite uncontrolled approximations such as level of theory and basis set discretisation, where bounds on the discretisation error can be obtained in some cases[83]. DFT based ML models proposed so far roughly fall into eihter one of the following three categories (see Fig. 3 for an overview).

(a) Direct learning of electronic observables (*circumvent/replace* in Fig. 3) rely on input based on the same information as for the model Hamiltonian: atomic coordinates $\{\mathbf{R}\}$, nuclear charges $\{Z\}$, the number of electrons N_e, and the spin σ. Direct learning is most in line with a black-box usage of ab initio calculations, as it is common in first principles based navigation campaigns of CCS. The main draw-back is that they may fall short of desired accuracy due to training data needs incommensurate with typical computational budgets. Various flavours of Delta-ML [69] (*refine* in Fig. 3) offer more flexibility to enhance predictive accuracy by focusing on learning corrections to labels, rather than absolute labels (error cancellation). For instance, one can use the same (different) many-body atomic representation(s) for baseline and target levels[69], or use instead electronic structure features of the baseline [84], or geometry features from physics inspired baseline, as demonstrated by the use of modified Lennard-Jones distances[85]. Multi-level grid combination technique based ML models[86], closely related to traditional composite QC, basis set extrapolation methods, or Jacob's ladder within DFT (Fig. 5) takes a significant step forward towards systematic exploitation of error cancellations: adding very many and very few data from lower and higher level of theory, respectively, can help to drastically reduce the prediction
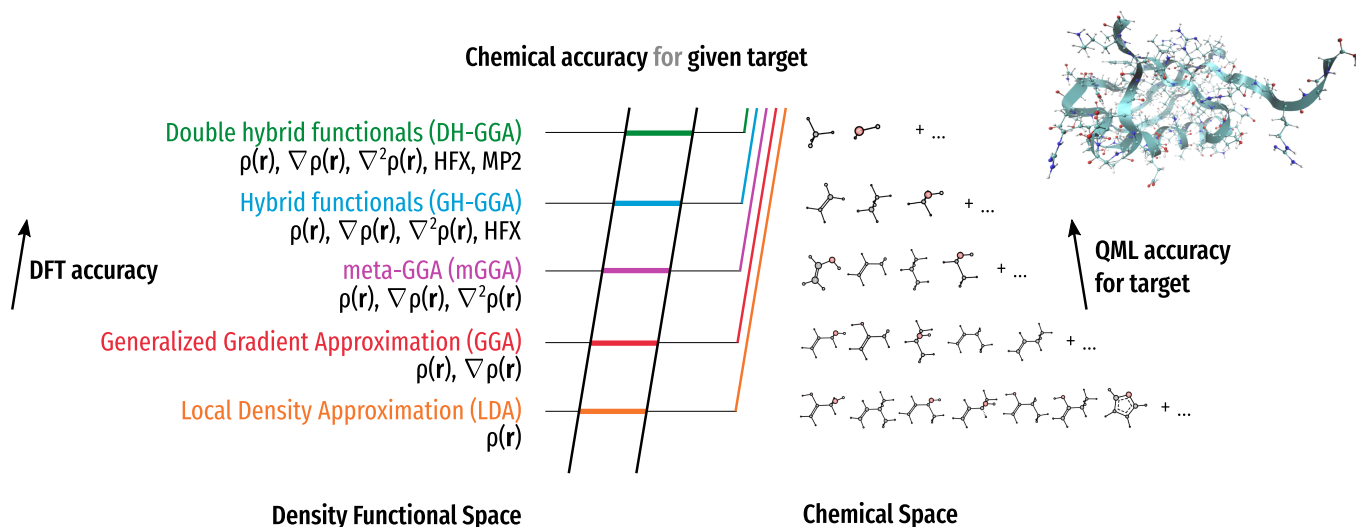
Figure 5. Unifying Jacob's Ladder in the realm of DFT and amons space (target-specific hierarchical chemical space) within ML. The goal is to predict the properties of large targets (shown on the top right is a small protein ubiquitin) with chemical accuracy while minimizing costs. For each pair of chemical space subset (denoted by the number of non-hydrogen atoms) and exchange-correlation functional (DF), the computational burden for training data generation is similar. In the illustration, only the smallest amons are displayed. HFX: Hartree-Fock exchange.

error at constant training data acquisition costs, see Fig. 1B.

(b) Recent contributions introduce hybrid ML/DFT approaches in order to improve the DFT model construction, i.e., learning the effective Hamiltonian as a intermediate quantity (*replace* in Fig. 3), from which target properties follow straightforwardly, as demonstrated e.g. by SchNorb [87] and DeepH [88]. Beside improving accuracy, this hybrid approach can offer better accuracy for intensive properties such as HOMO/LUMO energy. Similar strategies have already been pursued previously to improve semi-empirical quantum chemistry [89] and tight binding DFT [90, 91].

(c) Machine-learned density functionals correspond to the *assist* in Fig. 3. There are two variants, the first within the KS-DFT framework aiming to improve the mapping from the electron density to the exchange-correlation (XC) energy based on higher level reference data (e.g., DFT/CCSD density and energy at the CCSD(T) level). By controlling the quality of density and energy data, it is possible to approximate the exact yet unknown XC functional with high accuracy. The main drawback is that the computational cost of DFT is not alleviated, as most often the explicit dependence on orbitals in the kinetic energy and exchange term is not removed. Well-known contributions include DM21[79] and NeuralXC [92]). The second strategy is more in the spirit of the Hohenberg-Kohn theorems and builds the direct orbital-free map from electron density to energy (*augment* in Fig. 3). Early attempts include the machine-learning kinetic energy density functional[93], the ML-HK map[94] ($\nu(r) \to n(r) \to E$) for the ground state with later extension to excited states in the multi-state Kohn-Sham map [95]. Meanwhile, one may also use ΔE between DFT and CCSD(T)[96] for improved accuracy.

At last, it is worth mentioning transfer learning (TL). The central idea of TL is that the optimized neural network model parameters (of the first few hidden layers) trained on relatively low-fidelity (e.g., DFT) data can be transferred to the model trained on high-fidelity data (e.g., CCSD) [97]. Though effective in practise, TL may suffer from the downsides of low explainability, and difficulty in integrating data from multiple levels. This is to be contrasted to CQML counterparts, which take into account error cancellation in an explicit and systematic way and can take into account arbitrary number of levels of theory. To conclude this section we note that all quantum chemistry approximations are rooted in careful neglect of physical effects. Since most affordable quantum chemistry method include uncontrolled approximations, machine learning models will ultimately require the inclusion of experimental observables to improve substantially over purely computational approaches, e.g. by automated and data-driven approaches such as the Chemputer[98].

7## SCALABILITY

Scalability is critical for enabling the study of larger and more complex electronic systems. Within the ab initio framework, the cost of DFT roughly scales as $\mathcal{O}(N^4)$, where N is a measure of the system size. In combination with a similar pre-factor, this is considerably more favorable than other post-HF methods, such as CCSD(T) ($\sim \mathcal{O}(N^7)$). Nevertheless, the routine use of ab initio molecular dynamics simulations [99] using the most accurate flavours of DFT, i.e. hybrid or range-separated DFT in particular, has remained elusive already for small proteins, e.g. ubiquitin. This is were machine learning can help and, as long as long-range effects are small, affords quasi-linear scaling. Unke et al. have just recently demonstrated this point: Thanks to ML, they were able to investigate the stark differences between smooth and accurate DFT based trajectories of the protein crambin in aqueous solution, and classical force-field based Brownian motion like counterparts with stochastic characteristics [100].

Scalability of ML models can also be assessed in a more chemical sense. Namely with respect to its ability to generalize to larger query systems after training on smaller systems that encode similar local atomic environments as in the larger queries. Such scalability rests upon the locality assumption which is implied when using similarity measurements that are based on atomic environments. This assumption is often justified with reference to Kohn's nearsightedness of electronic matter (NEM) [101], as also recently revisited by Fias et al [102]. As pointed out by Kohn and others [103, 104], the locality of the one-particle electron density matrix (1-PDM) of a system with periodic boundary condition is related to the size of the HOMO-LUMO band gap Δ. The 1-PDM decays asymptotically as $e^{-\sqrt{\Delta}|\mathbf{r}-\mathbf{r}'|}$ and for gapless systems, as $[|x-x'||y-y'||z-z'|]^{-1}$. While there is no rigorous expression for non-periodic systems, similar arguments can be established based on localized occupied orbitals, such as Wannier orbitals [105].

A series of effects are manifestations of the NEM, such as the extent of charge transfer (which has bigger impact on charged species, as seen in [106]), conjugation, electron correlation (in particular for strongly correlated systems) and London dispersion (in large biomolecules as well as non-covalently interacting molecular complexes) (see also Fig. 4). These effects may propagate along the entire molecule, and are modulated by characteristics of all its local constituents; As such, short- and long-range effects can be intertwined and consequently become non-separable. However, one often refers to these effects as „long-ranged" in order to distinguish them from the purely local covalent bond and hybridization concepts as they are usually seen in semi-empirical quantum as well as force field approaches. Moreover, these effects may not occur in isolation, but collectively impact the overall locality of constituents in molecules, and ultimately the scalability of the ML model trained on these molecules and its ability to generalize to new compounds (see Fig. 4).

In practise, the prerequisite of a scalable ML model is the use of atomic representation for molecular description, where each atom in the molecule is represented by a vector and each vector encodes the many-body interaction between it and its neighbors. However, achieving scalability for large query molecules requires more than just atomic representation. In fact, the lack of scalability ranks perhaps among the most common and conspicuous issues of state-of-the-art ML models. A possible reason for this might be the difficulty of accurately accounting for long-range interactions, which is, by definition, are absent during ML training (since they are much smaller in size compared to the target). Among the many long range effects, interacting Coulombic multipole moments and their polarizabilities may be considered easiest due to their well known classical structure and the availability of many empirical models. As for other effects of quantum mechanical origin, such as conjugation effects and electron correlation, decent scalable ML based approximations are yet to be invented.

TRANSFERABILITY

The physical origins of chemical transferability of atoms, bonds or functional groups remain obscure, but are undoubtedly linked to the aforementioned nearsightedness of electronic matter [102]. Perhaps a more chemically intuitive interpretation of the nearsightedness is through the energy partitioning according to Bader's theory of atom in molecule, which has revealed [107, 108] that the partitioned atomic energy, calculated within the atomic volume enclosed by the zero-flux surface, remains remarkably stable across different molecules. We note however that other partitioning schemes will provide different answers when it comes to the conservation of atomic energies [109]. Nevertheless, transferability and locality (and thus scalability) are so intertwined that it is difficult to rigorously separate them. Here, and for practical purposes, we associate transferability with the capability to generalize short-range effects across CCS. Long-range effects, by contrast, rather affect scalability, as discussed in the previous section. This readily made separation enables us to focus on the well-known force-field concepts such as bond, angle and torsion, as well as the steric hindrance, which, together with their distortions off equilibrium, can be conveniently represented by molecular fragments of increasing size, as illustrated by the amon concept above. [67] The other effects, dominated by long-range characteristics, can be left for approximate treatments with empirical or other relevant

models [106]).

Many QC quantities, being observable or not, are transferable, yet exhibit varying degrees of transferability even if they belong to the same kind. For instance, energy has several variants, such as the total energy (from a specific level of theory), the energy of an atom in molecule (see [110]), the electron correlation energy, relaxation energies, or the difference between two energies (ΔE for short) from two different levels of theory. Recent contributions have shown that training accurate ML models of molecular Hartree-Fock (HF) energies requires more training data than any other type of energy involving some degree of correlation energy (from either post-HF or DFT) since correlation tends to bring electrons closer to nucleus on average [68]. Compared to an absolute energy (e.g. atomization energy), the energy delta demonstrates better transferablity in generally in the delta learning (or more generally, CQML) models, largely due to the cancellation of errors, as mentioned above.

Previous studies have mostly focused on the atomic scale and represent a system through the many-body potential, or interaction between atoms. However, recent research has shown that properties at the finer electronic resolution could be transferable as well. For instance, the electron density of small molecules (e.g., ethene (C_2H_4) and butadiene (C_4H_6)) is found to be transferable to larger ones (e.g., octa-tetraene (C_8H_{10})) [111, 112]. Learning the delta of electron densities, or the defomration density [113], can also help improve the transferability. Further evidences supporting the transferability of electron in molecule include the electronic force field approach (treating electrons instead of atoms as semiclassical particles that interact with each other) [114], and the (occupied) localized molecular orbital feature based approaches [115]. Superior model transferability can also arise from the use of electronic structure (from DFT calculations) based features [116], such as Mulliken charges and bond order. Notably, bond order can also help identify transition metal complexes that do not exhibit multireference nature and can be handled by default DFT approximations. [117]

Another closely related concept at the electronic level, the density functional, is deemed transferable at a more fundamental level. Approximated functionals constructed from heuristics, however, are prone to problems in their accuracy and transferability. For example, torsional potentials of conjugated double bonds have typically been assumed to be properly accounted for by modern density functional approximations, until their qualitative break-down was shown for glyoxal and oxalyl halides, and their thiocarbonyl derivatives [118, 119]. Such short-comings of popular density functional approximations, as also evidenced by Nagai and collaborators' work [120], could be alleviated through systematic construction of density functionals via data-driven ML approaches (*assist* in Fig. 3). By incorporating physical constraints into ML, we can enhance the transferability of density functional. This has also been shown more recently by learning the non-local exchange density functional with a tailored representation that preserves the density distribution under uniform scaling [121]. Beyond DFT, physical constraints (such as Kato's cusp conditions) are also crucial in the correlated framework, as they help enhance the expressiveness of deep neural network for wavefunction approximation. [122]

CONCLUSIONS

In light of the above, we think it self-evident that DFT has been playing an instrumental role for the emergence of machine learning based navigation of chemical compound space. We do not think that it is far-fetched to expect these developments to directly lead to the emergence of autonomous self-driving laboratories. Improving ML models implies access to more and accurate reference data. As such, data availability determines the upper bound of ML accuracy [68]. Among its many advantages, DFT is truly outstanding as a source of, on average, highly accurate calculated properties for arbitrarily chosen molecules or materials — with controllable and reasonable acquisition costs. Outstanding challenges to DFT include surface crossing, open-shell and spin-orbit coupling effects, conductivity and excited states dynamics. But even in these cases, DFT results may still be extremely useful as an intermediate quality method that can be exploited in multi-fidelity ML models.

Availability of large, diverse and high-accuracy (at the experimental level or above) materials and molecular property datasets is a fundamental requirement for developing universal machine learning models that can handle any chemistry and that can conveniently be incorporated within the experimental planning software of future self-driving and closed-loop autonomous experimentation. The general scarcity and lack of data, however, represents in our opinion the most severe current road-block on our path towards that goal. In particular, most of the available databases report only equilibrium structures and properties. Notable exceptions, such as DFT based distortions along normal modes and conformers reported in QM7-X [125], are few, and typically scarcely scattered throughout the chemical compound space. Transition states, defects, charged species, radicals, entire ab initio molecular dynamics trajectories, or excited states, to name just some, are mostly amiss.

While notable achievements have been made in successfully applying physics based machine learning to DFT solutions throughout the chemical and materials sciences, there remains a dearth of theoretical research focused on the underlying fundamentals. Some of the few basic questions that still awaiting answers include, (i) can we rigorously define CCS in a mathematical way, maybe
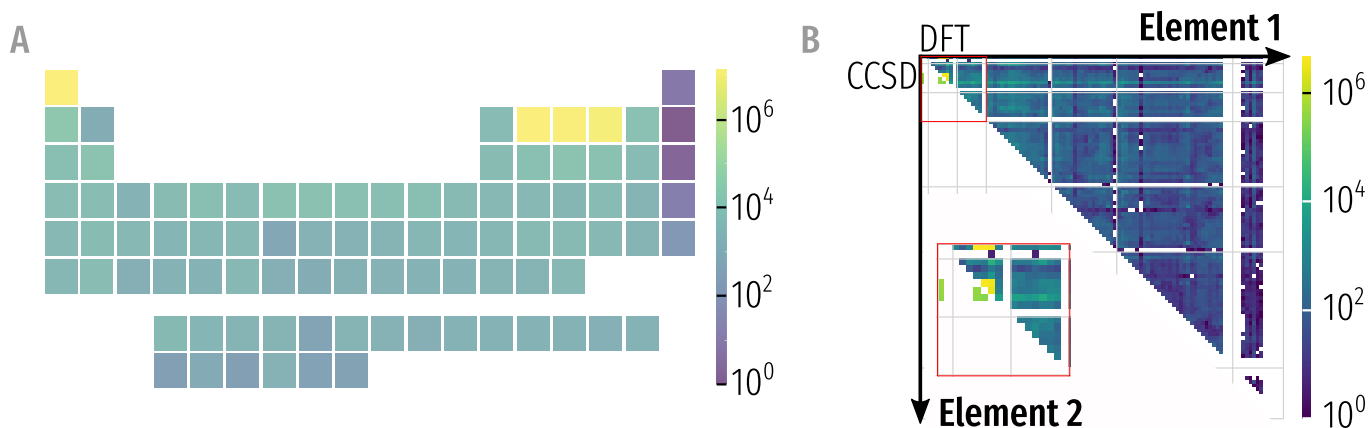
Figure 6. Coverage of chemical space in existing data sets, exemplified by the largest databases of energies for molecules and materials, ANI-1x/ANI-1ccx[123], and Materials Project[124], respectively. A: Histogram of the coverage of the periodic table in both data sets clearly shows a focus on few elements. B: Histogram of the pairwise occurrence of two elements in both data sets (grey lines denote periods) split into DFT reference data (upper triangle) and CCSD reference data (lower triangle). White fields have no data. The inset highlights the region for which CCSD data is available.

akin to the Hilbert space for electronic wave-functions, such that its inherent properties such as density and volume can be quantified [126]? (ii) CCS is discrete in reality, yet which maps enable smooth interpolations into latent spaces that facilitate inverse design [127]? iii) can there ever be one model that allows for a unified yet accurate description of any chemical compound, regardless of its size, composition, aggregation state, and external conditions?

In summary and roughly speaking, DFT has impacted ML in three ways, a) as an *ab initio* solver, with ever-improving performance of the data-driven exchange-correlation and/or kinetic energy density functional over time; b) as a hybrid DFT/ML framework that builds the effective Hamiltonian directly, enhancing further the efficiency and/or accuracy of DFT; c) finally, as a generator of large-scale data. Based on our above review in terms of the four EAST categories: Efficiency, accuracy, scalability, and transferability, it should be clear that DFT has played an outstanding pivotal role for all the chemical sciences in bridging all the pillar of modern science, from experiments via theory and simulation to physics based ML model building. Seeing the impressive progress made by building on DFT, we conclude by phrasing our optimism about the wide-spread emergence of the next, 5th pillar of science in a not too distant future: Self-driving laboratories in the chemical and materials sciences.

ACKNOWLEDGMENTS

* bing.huang@univie.ac.at
† vonrudorff@uni-kassel.de
‡ anatole.vonlilienfeld@utoronto.ca

[1] S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, and R. M. Myers, Organic synthesis: march of the machines, Angewandte Chemie International Edition **54**, 3449 (2015).

[2] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, and L. Cronin, Controlling an organic synthesis robot with machine learning to search for new reactivity, Nature **559**, 377 (2018).

[3] K. Sanderson *et al.*, Automation: Chemistry shoots for the moon, Nature **568**, 577 (2019).

[4] C. W. Coley, D. A. Thomas III, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, *et al.*, A robotic platform for flow synthesis of organic compounds informed by ai planning, Science **365**, eaax1566 (2019).

[5] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, *et al.*, A mobile robotic chemist, Nature **583**, 237 (2020).

[6] A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller, and T. Laino, Automated extraction of chemical synthesis actions from experimental procedures, Nature communications **11**, 3601 (2020).

[7] B. P. MacLeod, F. G. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. Yunker, M. B. Rooney, J. R. Deeth, *et al.*, Self-driving laboratory for accelerated discovery of thin-film materials, Science Advances **6**, eaaz8867 (2020).

[8] R. D. King, K. E. Whelan, F. M. Jones, P. G. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, Functional genomic hypothesis generation and

experimentation by a robot scientist, Nature **427**, 247 (2004).

[9] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, *et al.*, The automation of science, Science **324**, 85 (2009).

[10] R. D. King, Rise of the robo scientists, Scientific American **304**, 72 (2011).

[11] M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam, *et al.*, On scientific understanding with artificial intelligence, Nature Reviews Physics **4**, 761 (2022).

[12] A. Aspuru-Guzik, R. Lindh, and M. Reiher, The matter simulation (r) evolution, ACS central science **4**, 144 (2018).

[13] O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning, Nature Reviews Chemistry , 1 (2020).

[14] G. Ceder, Predicting properties from scratch, Science **280**, 1099 (1998).

[15] J. Hafner, C. Wolverton, G. Ceder, and G. Editors, Toward computational materials design: The impact of density functional theory on materials research, MRS Bulletin **31**, 659 (2006).

[16] N. Marzari, Materials modelling: The frontiers and the challenges, Nature materials **15**, 381 (2016).

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

[18] O. A. von Lilienfeld, First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties, International Journal of Quantum Chemistry **113**, 1676 (2013).

[19] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (John Wiley & Sons, LTD, 2000).

[20] R. S. Bohacek, C. McMartin, and W. C. Guida, The art and practice of structure-based drug design: A molecular modeling perspective, Med. Res. Rev. **16**, 3 (1996).

[21] K. Shinohara, A. Seko, T. Horiyama, M. Ishihata, J. Honda, and I. Tanaka, Enumeration of nonequivalent substitutional structures using advanced data structure of binary decision diagram, The Journal of Chemical Physics **153**, 104109 (2020).

[22] G. F. von Rudorff and O. A. von Lilienfeld, Alchemical perturbation density functional theory, Physical Review Research **2**, 023220 (2020).

[23] M. F. Kasim, S. Lehtola, and S. M. Vinko, DQC: A python program package for differentiable quantum chemistry, The Journal of Chemical Physics **156**, 084801 (2022).

[24] G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Nørskov, Combined electronic structure and evolutionary search approach to materials design, Phys. Rev. Lett. **88**, 255506 (2002).

[25] J. G. Freeze, H. R. Kelly, and V. S. Batista, Search for catalysts by inverse design: Artificial intelligence, mountain climbers, and alchemists, Chemical reviews (2019).

[26] N. Marzari, A. Ferretti, and C. Wolverton, Electronic-structure methods for materials design, Nature materials **20**, 736 (2021).

[27] R. G. Parr and W. Yang, *Density functional theory of atoms and molecules* (Oxford Science Publications, 1989).

[28] A. E. Mattsson, In pursuit of the "divine" functional, Science **298**, 759 (2002).

[29] V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes, J. Chem. Phys. **119**, 12129 (2003).

[30] K. Burke, Perspective on density functional theory, J. Chem. Phys. **136**, 150901 (2012).

[31] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, Density functional theory is straying from the path toward the exact functional, Science **355**, 49 (2017).

[32] R. M. Dreizler and E. Gross, *Density Functional Theory* (Springer Verlag, 1990).

[33] R. V. Noorden, B. Maher, and R. Nuzzo, The top 100 papers, Nature **514**, 550 (2014).

[34] J. Sun, A. Ruzsinszky, and J. Perdew, Strongly constrained and appropriately normed semilocal density functional, Physical Review Letters **115**, 10.1103/physrevlett.115.036402 (2015).

[35] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. Dal Corso, *et al.*, Reproducibility in density functional theory calculations of solids, Science **351**, aad3000 (2016).

[36] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, Density functional theory is straying from the path toward the exact functional, Science **355**, 49 (2017).

[37] C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker, Learning curves: Asymptotic values and rate of convergence, in *Advances in Neural Information Processing Systems* (1994) pp. 327–334.

[38] G. F. von Rudorff, S. N. Heinen, M. Bragato, and O. A. von Lilienfeld, Thousands of reactants and transition states for competing e2 and s n 2 reactions, Machine Learning: Science and Technology **1**, 045026 (2020).

[39] M. Rupp, Special issue on machine learning and quantum mechanics, International Journal of Quantum Chemistry **115**, 1003 (2015).

[40] M. Rupp, O. A. von Lilienfeld, and K. Burke, Guest editorial: Special topic on data-enabled theoretical chemistry, J. Chem. Phys. **148**, 241401 (2018).

[41] W. F. Schneider and H. Guo, Machine learning, J Phys Chem C **122**, 879 (2018).

[42] O. V. Prezhdo, Advancing physical chemistry with machine learning, J. Phys. Chem. Lett. (2020).

[43] A. Tkatchenko, Machine learning for chemical discovery, Nature Communications **11**, 1 (2020).

[44] M. Ceriotti, C. Clementi, and O. Anatole von Lilienfeld, Machine learning meets chemical physics (2021).

[45] M. Ceriotti, C. Clementi, and O. Anatole von Lilienfeld, Introduction: Machine learning at the atomic scale, Chem. Rev. **121**, 9719 (2021).

[46] K. Schütt, S. Chmiela, O. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K. Müller, *Machine Learning Meets Quantum Physics*, Lecture Notes in Physics (Springer International Publishing, 2020).

[47] R. Ramakrishnan and O. A. von Lilienfeld, Machine learning, quantum chemistry, and chemical space, in *Reviews in Computational Chemistry*, Vol. 30 (John Wiley

& Sons, Inc., 2017) pp. 225–256.

[48] O. A. von Lilienfeld, Quantum machine learning in chemical compound space, Angewandte Chemie International Edition **57**, 4164 (2018), http://dx.doi.org/10.1002/anie.201709686.

[49] J. R. Kitchin, Machine learning in catalysis, Nature Catalysis **1**, 230 (2018).

[50] B. Huang, N. O. Symonds, and O. A. v. Lilienfeld, Quantum machine learning in chemistry and materials, Handbook of Materials Modeling: Methods: Theory and Modeling , 1 (2018).

[51] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, Nature **559**, 547 (2018).

[52] A. Aspuru-Guzik, R. Lindh, and M. Reiher, The matter simulation (r) evolution, ACS central science **4**, 144 (2018).

[53] F. A. Faber and O. Anatole von Lilienfeld, Modeling materials quantum properties with machine learning, Materials Informatics: Methods, Tools and Applications , 171 (2019).

[54] O. A. von Lilienfeld and K. Burke, Retrospective on a decade of machine learning for chemical discovery, Nature Communications **11**, 1 (2020).

[55] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, Machine learning for molecular simulation, Annual review of physical chemistry **71**, 361 (2020).

[56] F. A. Faber, A. S. Christensen, and O. A. von Lilienfeld, Quantum machine learning with response operators in chemical compound space, in *Machine Learning Meets Quantum Physics* (Springer, 2020) pp. 155–169.

[57] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, Machine learning force fields, arXiv preprint arXiv:2010.07067 (2020).

[58] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, *et al.*, Qsar without borders, Chemical Society Reviews (2020).

[59] S. Chibani and F.-X. Coudert, Machine learning approaches for the prediction of materials properties, APL Materials **8**, 080701 (2020).

[60] P. O. Dral, Quantum chemistry in the age of machine learning, The Journal of Physical Chemistry Letters **11**, 2336 (2020).

[61] O. A. von Lilienfeld, Introducing machine learning: Science and technology, Machine Learning: Science and Technology **1**, 010201 (2020).

[62] E. O. Pyzer-Knapp, J. Cuff, J. Patterson, O. Isayev, and S. Maskell, Welcome to the first issue of applied ai letters, Applied AI Letters (2020).

[63] N. J. Browning, F. A. Faber, and O. Anatole von Lilienfeld, Gpu-accelerated approximate kernel method for quantum machine learning, The Journal of Chemical Physics **157**, 214801 (2022).

[64] D. Khan, S. Heinen, and O. A. von Lilienfeld, Quantum machine learning at record speed: Many-body distribution functionals as compact representations, arXiv preprint arXiv:2303.16312 (2023).

[65] P. Zaspel, B. Huang, H. Harbrecht, and O. A. von Lilienfeld, Boosting quantum machine learning models with multi-level combination technique: Pople diagrams revisited, Journal of chemical theory and computation (2018).

[66] B. Huang and O. A. von Lilienfeld, Quantum machine learning using atom-in-molecule-based fragments selected on the fly, Nature Chemistry (2020).

[67] B. Huang and O. A. von Lilienfeld, Quantum machine learning using atom-in-molecule-based fragments selected on the fly, Nature Chemistry **12**, 945 (2020).

[68] B. Huang, O. A. von Lilienfeld, J. T. Krogel, and A. Benali, Toward dmc accuracy across chemical space with scalable δ-qml, Journal of Chemical Theory and Computation (2023), pMID: 36857531, https://doi.org/10.1021/acs.jctc.2c01058.

[69] R. Ramakrishnan, P. Dral, M. Rupp, and O. A. von Lilienfeld, Big Data meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach, J. Chem. Theory Comput. **11**, 2087 (2015).

[70] R. Batra, G. Pilania, B. P. Uberuaga, and R. Ramprasad, Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia, ACS applied materials & interfaces (2019).

[71] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, Nature communications **10**, 1 (2019).

[72] K. Ryczko, S. J. Wetzel, R. G. Melko, and I. Tamblyn, Toward orbital-free density functional theory with small data sets and deep learning, Journal of Chemical Theory and Computation **18**, 1122 (2022).

[73] P. Rowe, V. L. Deringer, P. Gasparotto, G. Csányi, and A. Michaelides, An accurate and transferable machine learning potential for carbon, The Journal of Chemical Physics **153**, 034702 (2020).

[74] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong, Performance and cost assessment of machine learning interatomic potentials, The Journal of Physical Chemistry A **124**, 731 (2020).

[75] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, Less is more: Sampling chemical space with active learning, The Journal of Chemical Physics **148**, 241733 (2018).

[76] T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design, npj Computational Materials **5**, 10.1038/s41524-019-0153-8 (2019).

[77] G. Csányi, T. Albaret, M. C. Payne, and A. D. Vita, "learn on the fly": A hybrid classical and quantum-mechanical molecular dynamics simulation, Physical Review Letters **93**, 175503 (2004).

[78] K. R. Müller, M. Finke, N. Murata, K. Schulten, and S. Amari, A numerical study on learning curves in stochastic multilayer feedforward networks, Neural Comp. **8**, 1085 (1996).

[79] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen, Pushing the frontiers of density functionals by solving the fractional electron problem, Science **374**, 1385 (2021).

[80] B. Huang and O. A. von Lilienfeld, Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity,

J. Chem. Phys. **145**, 161102 (2016).

[81] M. Reiher, Molecule-specific uncertainty quantification in quantum chemical studies, Israel Journal of Chemistry **62**, 10.1002/ijch.202100101 (2021).

[82] P. Pernot, B. Huang, and A. Savin, Impact of non-normal error distributions on the benchmarking and ranking of quantum machine learning models, Machine Learning: Science and Technology **1**, 035011 (2020).

[83] E. Cancès, G. Dusson, G. Kemlin, and A. Levitt, Practical error bounds for properties in plane-wave electronic structure calculations (2021).

[84] M. Welborn, L. Cheng, and T. F. Miller, Transferability in machine learning for electronic structure via the molecular orbital basis, Journal of Chemical Theory and Computation **14**, 4772 (2018).

[85] S. Bag, M. Konrad, T. Schlöder, P. Friederich, and W. Wenzel, Fast generation of machine learning-based force fields for adsorption energies, Journal of Chemical Theory and Computation **17**, 7195 (2021).

[86] P. Zaspel, B. Huang, H. Harbrecht, and O. A. von Lilienfeld, Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited, Journal of Chemical Theory and Computation **15**, 1546 (2018).

[87] K. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions, Nature communications **10**, 1 (2019).

[88] H. Li, Z. Wang, N. Zou, M. Ye, R. Xu, X. Gong, W. Duan, and Y. Xu, Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation, Nature Computational Science **2**, 367 (2022).

[89] P. O. Dral, O. A. von Lilienfeld, and W. Thiel, Machine learning of parameters for accurate semiempirical quantum chemical calculations, Journal of Chemical Theory and Computation **11**, 2120 (2015), pMID: 26146493, http://dx.doi.org/10.1021/acs.jctc.5b00141.

[90] J. J. Kranz, M. Kubillus, R. Ramakrishnan, O. A. von Lilienfeld, and M. Elstner, Generalized density-functional tight-binding repulsive potentials from unsupervised machine learning, Journal of chemical theory and computation **14**, 2341 (2018).

[91] M. Stöhr, L. Medrano Sandonas, and A. Tkatchenko, Accurate many-body repulsive potentials for density-functional tight binding from deep tensor neural networks, The Journal of Physical Chemistry Letters **11**, 6835 (2020).

[92] S. Dick and M. Fernandez-Serra, Machine learning accurate exchange and correlation functionals of the electronic density, Nature communications **11**, 3509 (2020).

[93] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Finding density functionals with machine learning, Phys. Rev. Lett. **108**, 253002 (2012).

[94] Y. Kiat, Y. Vortman, and N. Sapir, Feather moult and bird appearance are correlated with global warming over the last 200 years, Nature Communications **10**, 2540 (2019).

[95] Y. Bai, L. Vogt-Maranto, M. E. Tuckerman, and W. J. Glover, Machine learning the hohenberg-kohn map for molecular excited states, Nature communications **13**, 7044 (2022).

[96] M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, and K. Burke, Quantum chemical accuracy

from density functional approximations via machine learning, Nature Communications **11**, 5223 (2020), number: 1 Publisher: Nature Publishing Group.

[97] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, Nature Communications **10**, 10.1038/s41467-019-10827-4 (2019).

[98] S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone, and L. Cronin, Organic synthesis in a modular robotic system driven by a chemical programming language, Science **363**, 10.1126/science.aav2211 (2019).

[99] R. Car and M. Parrinello, A combined approach to DFT and molecular dynamics, Phys. Rev. Lett. **55**, 2471 (1985).

[100] O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, H. Maennel, S. Kashubin, D. Ahlin, M. Gastegger, L. M. Sandonas, A. Tkatchenko, *et al.*, Accurate machine learned quantum-mechanical force fields for biomolecular simulations, arXiv preprint arXiv:2205.08306 (2022).

[101] E. Prodan and W. Kohn, Nearsightedness of electronic matter, Proceedings of the National Academy of Sciences of the United States of America **102**, 11635 (2005), http://www.pnas.org/content/102/33/11635.full.pdf+html.

[102] S. Fias, F. Heidar-Zadeh, P. Geerlings, and P. W. Ayers, Chemical transferability of functional groups follows from the nearsightedness of electronic matter, Proceedings of the National Academy of Sciences **114**, 11633 (2017), publisher: National Academy of Sciences Section: Physical Sciences.

[103] R. Baer and M. Head-Gordon, Sparsity of the Density Matrix in Kohn-Sham Density Functional Theory and an Assessment of Linear System-Size Scaling Methods, Physical Review Letters **79**, 3962 (1997), publisher: American Physical Society.

[104] W. Kohn, Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms, Physical Review Letters **76**, 3168 (1996), publisher: American Physical Society.

[105] G. Wannier, Phys. Rev. **52**, 191 (1937).

[106] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer, Nature Communications **12**, 10.1038/s41467-020-20427-2 (2021).

[107] R. F. Bader, A quantum theory of molecular structure and its applications, Chemical Reviews **91**, 893 (1991).

[108] R. F. W. Bader, Nearsightedness of electronic matter as seen by a physicist and a chemist, The Journal of Physical Chemistry A **112**, 13717 (2008), pMID: 19032142, https://doi.org/10.1021/jp806282j.

[109] G. F. von Rudorff and O. A. von Lilienfeld, Atoms in molecules from alchemical perturbation density functional theory, The Journal of Physical Chemistry B **123**, 10073 (2019).

[110] M. J. Burn and P. L. A. Popelier, Gaussian process regression models for predicting atomic energies and multipole moments, Journal of Chemical Theory and Computation **19**, 1370 (2023), pMID: 36757024,

https://doi.org/10.1021/acs.jctc.2c00731.

[111] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf, Electron density learning of non-covalent systems, Chemical Science **10**, 9424 (2019).

[112] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, Transferable machine-learning model of the electron density, ACS Central Science **5**, 57 (2018).

[113] K. Low, M. L. Coote, and E. I. Izgorodina, Inclusion of more physics leads to less data: Learning the interaction energy as a function of electron deformation density with limited training data, Journal of Chemical Theory and Computation **18**, 1607 (2022).

[114] M. Cools-Ceuppens, J. Dambre, and T. Verstraelen, Modeling electronic response properties with an explicit-electron machine learning potential, Journal of Chemical Theory and Computation **18**, 1672 (2022).

[115] T. Husch, J. Sun, L. Cheng, S. J. R. Lee, and T. F. Miller, Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states, The Journal of Chemical Physics **154**, 064108 (2021).

[116] C. Duan, A. Nandy, H. Adamji, Y. Roman-Leshkov, and H. J. Kulik, Machine learning models predict calculation outcomes with the transferability necessary for computational catalysis, Journal of Chemical Theory and Computation **18**, 4282 (2022).

[117] C. Duan, A. J. Ladera, J. C.-L. Liu, M. G. Taylor, I. R. Ariyarathna, and H. J. Kulik, Exploiting ligand additivity for transferable machine learning of multireference character across known transition metal complex ligands, Journal of Chemical Theory and Computation **18**, 4836 (2022).

[118] D. N. Tahchieva, D. Bakowies, R. Ramakrishnan, and O. A. von Lilienfeld, Torsional potentials of glyoxal, oxalyl halides, and their thiocarbonyl derivatives: Challenges for popular density functional approximations, Journal of Chemical Theory and Computation **14**, 4806 (2018).

[119] S. Nam, E. Cho, E. Sim, and K. Burke, Explaining and fixing dft failures for torsional barriers, The journal of physical chemistry letters **12**, 2796 (2021).

[120] R. Nagai, R. Akashi, and O. Sugino, Completing density functional theory by machine learning hidden messages from molecules, npj Computational Materials **6**, 10.1038/s41524-020-0310-0 (2020).

[121] K. Bystrom and B. Kozinsky, CIDER: An expressive, nonlocal feature set for machine learning density functionals with exact constraints, Journal of Chemical Theory and Computation **18**, 2180 (2022).

[122] J. Hermann, Z. Schätzle, and F. Noé, Deep-neural-network solution of the electronic schrödinger equation, Nature Chemistry **12**, 891 (2020).

[123] J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev, and S. Tretiak, The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules, Scientific Data **7**, 10.1038/s41597-020-0473-z (2020).

[124] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Materials **1**, 011002 (2013).

[125] J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr, and A. Tkatchenko, Qm7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules, Scientific data **8**, 43 (2021).

[126] D. Lemm, G. F. von Rudorff, and O. A. von Lilienfeld, Improved decision making with similarity based machine learning, arXiv preprint arXiv:2205.05633 (2022).

[127] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS central science **4**, 268 (2018).