

# Ab Initio Machine Learning in Chemical Compound Space

Bing Huang and O. Anatole von Lilienfeld\*



Cite This: *Chem. Rev.* 2021, 121, 10001–10036



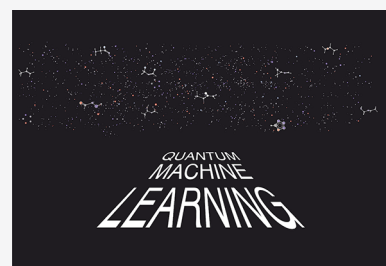
Read Online

ACCESS |

 Metrics & More

 Article Recommendations

**ABSTRACT:** Chemical compound space (CCS), the set of all theoretically conceivable combinations of chemical elements and (meta-)stable geometries that make up matter, is colossal. The first-principles based virtual sampling of this space, for example, in search of novel molecules or materials which exhibit desirable properties, is therefore prohibitive for all but the smallest subsets and simplest properties. We review studies aimed at tackling this challenge using modern machine learning techniques based on (i) synthetic data, typically generated using quantum mechanics based methods, and (ii) model architectures inspired by quantum mechanics. Such Quantum mechanics based Machine Learning (QML) approaches combine the numerical efficiency of statistical surrogate models with an ab initio view on matter. They rigorously reflect the underlying physics in order to reach universality and transferability across CCS. While state-of-the-art approximations to quantum problems impose severe computational bottlenecks, recent QML based developments indicate the possibility of substantial acceleration without sacrificing the predictive power of quantum mechanics.



## CONTENTS

1. Introduction	10001	7.2. Molecular	10019
1.1. Multiscale Nature of CCS	10002	7.3. Intermolecular	10020
1.2. Machine Learning the Potential Energy Surface	10003	8. Data Sets	10021
1.3. Navigating CCS from First Principles	10004	8.1. GDB	10021
2. Heuristic Approaches	10005	8.2. PubChem and ZINC	10022
2.1. Low-dimensional Correlations	10006	8.3. Barriers and Spin	10022
2.2. Stoichiometry	10006	8.4. Transition Metals	10023
2.3. Connectivity Graph	10006	8.5. Solid and Solid Surface	10023
2.4. Coarse-grained	10007	9. Software Packages	10024
2.5. Property Based	10007	10. Compound Discovery	10024
3. QML Methodology	10007	11. Outlook and Conclusion	10024
3.1. Regressor	10008	Author Information	10025
3.2. Learning Curves	10009	Corresponding Author	10025
3.3. Loss Functions	10010	Author	10025
4. Representations	10010	Notes	10025
4.1. Discrete	10011	Biographies	10025
4.2. Continuous	10011	Acknowledgments	10026
5. Regressor	10012	References	10026
5.1. ML Models of Parameters	10012		
5.2. $\Delta$ -ML	10013		
5.3. Multifidelity	10013		
5.4. Multilevel Grid Combination	10014		
5.5. Transfer Learning	10014		
6. Training Set Selection	10015		
6.1. Genetic Algorithm	10015		
6.2. Active Learning	10016		
6.3. AMON Based QML	10017		
7. Properties	10018		
7.1. Atomic	10018		

## 1. INTRODUCTION

Promising applications of machine learning techniques have been rapidly gaining momentum throughout the chemical sciences. Apart from this present special issue in *Chemical*

**Special Issue:** Machine Learning at the Atomic Scale

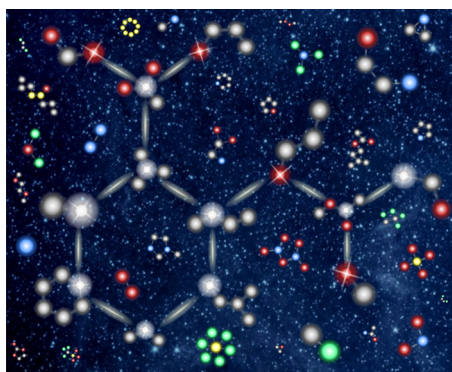
**Received:** December 21, 2020

**Published:** August 13, 2021



Reviews, a number of special issues in common theoretical chemistry community journals have appeared, including *International Journal of Quantum Chemistry* (2015),<sup>1</sup> *Journal of Chemical Physics* (2018),<sup>2</sup> *Journal of Physical Chemistry* (2018),<sup>3</sup> *Journal of Physical Chemistry Letters* (2020),<sup>4</sup> and *Nature Communications* (2020).<sup>5</sup> Books, essays, reviews, and opinion pieces have also been contributed by practitioners in the field.<sup>6–23</sup> Such growth of interest prompted a general discussion in *Angewandte Chemie* within a trilogy of essays by Hoffmann and Malrieu on the seemingly conflicting nature of simulation and understanding in quantum chemistry.<sup>24–26</sup> The overall enthusiasm in the hard sciences for machine learning has even led to the introduction of novel journals, such as Springer's *Nature Machine Intelligence*, IOP's *Machine Learning: Science and Technology*,<sup>27</sup> or Wiley's *Applied Artificial Intelligence Letters*.<sup>28</sup>

In this review, we attempt to provide a comprehensive overview on recent progress made regarding the problem of using machine learning models to train and predict quantum properties throughout chemical compound space (CCS) (Figure 1). In contrast to the current trend of machine



**Figure 1.** A cartoon of similarities among atoms across chemical compound space, not in conflict with quantum mechanics. The exemplary molecule aspirin is highlighted by bonds, and each of its atoms is superimposed with a similar atom in another molecule (hydrogens omitted for clarity). Green, yellow, gray, red, and blue refer to sulfur, phosphor, carbon, oxygen, and nitrogen, respectively. Reproduced with permission from ref 15. Copyright 2020 Springer Nature.

learning in quantum computing, we here refer to the application of statistical learning to quantum properties as “quantum machine learning” (QML). This notation follows a common convention in atomistic simulation, where the quantum nature of the object to be studied corresponds to a prefix, while the actual algorithms are rather classical in nature. Examples include Quantum Monte Carlo or Quantum Molecular Dynamics (also known as *ab initio* or “first-principles” molecular dynamics).

Within this introductory section, we will begin by first providing a qualitative description of chemical compound space (CCS) in terms of fundamental variables, which is consistent with the quantum mechanical picture within the Born–Oppenheimer approximation and neglecting nuclear quantum and relativistic effects. Thereafter, we briefly review related but complementary and system specific QML models which predominantly are not used throughout CCS but rather for training and predicting potential energies and forces in terms of conformational degrees of freedom, e.g., using

molecular dynamics. Quantum mechanics based explorations for the purpose of materials design are mentioned subsequently, followed by a short subsection on studies which establish the quantitative and rigorous quantum chemistry based view on CCS.

### 1.1. Multiscale Nature of CCS

Figuratively speaking, CCS refers to the virtual set of all the theoretically (meta-)stable compounds one could possibly realize in this universe. To paraphrase Buckingham and Utting, a compound “...is a group of atoms...with a binding energy which is large in comparison with the thermal energy  $kT$ .”<sup>29</sup> In other words, with respect to all its spatial degrees of freedom, it is that locally averaged atomic configuration, for which the free energy is in a local minimum surrounded by barriers sufficiently high to prevent spontaneous reactions within some observable lifetime. As such, CCS depends on external conditions. It loses all meaning, for example, when conditions are such that bonding spontaneously emerges and vanishes (e.g., aggregation state of plasma).

The mathematical number of compounds grows explosively with the number of constituting atoms due to the mutual enhancement of combinatorial scaling at three rather distinct but well established energetic scales: First, the number of possible stoichiometries for any given system size (in terms of electrons and total proton number) represents an integer partitioning problem which grows combinatorially, see ref 30, for example. The energetic variance among compounds that differ in stoichiometry is on the scale of chemical bonding due to having different number and different types of atoms. Second, the number of possible connectivity patterns, i.e., incomplete labeled undirected weighted graphs distinguishing constitutional isomers/allotropes (commonly drawn as Lewis structures) is mathematically known to grow combinatorially with number of atoms.<sup>31–33</sup> The energetic variance among constitutional isomers is on the scale of differences in chemical bonding. Third, the number of possible conformational degrees of freedom grows combinatorially with number of atoms in a molecular graph (cf. Levinthal’s paradoxon for polymers), and one could even consider different atomic configurations of disconnected graphs, i.e., macromolecular or molecular condensed systems, to fall into this category of isomers. As such, the energetic variance among conformational isomers is on the scale of noncovalent intra- as well as intermolecular interactions. We note that stereoisomerism typically occurs among constitutional and conformational isomers. Its extension to compositional chirality has been proposed only recently.<sup>34</sup> Given such size and diversity, highly universal, and efficient methods are in dire need in order to meaningfully explore CCS in search of deepened chemical insight and intuition and of new compounds and materials which exhibit desirable properties. While quantum mechanics and statistical mechanics offer the appropriate physical framework for dealing with CCS in an unbiased and universal manner, the computational complexity of the equations involved has hampered their widespread use.

We note that our *ab initio* definition of CCS implies that only those compounds are part of CCS that should, at least in principle, be experimentally accessible as long as sufficiently sophisticated synthetic chemical procedures and reservoirs of the necessary chemical elements are available. While any such synthetic procedure would have to follow the corresponding relevant free energy paths, by navigating the virtual analogue of

CCS we do enjoy more design freedom and can, namely for any property that is a state function, also exploit unrealistic fictitious transformations in line with Hess' law, i.e., without the need for direct correspondence to experimental realization (cf. "alchemical" transmutations).

We conclude this section by noting that our definition generalizes the more commonly made reference to CCS, which typically excludes conformational isomers, reactive intermediates, or minima in electronically excited states. For example, first steps toward an *ab initio* based representative exploration of the latter were also proposed in 2013 for drug-like compounds by Beratan and co-workers.<sup>35</sup> However, for this review, we do not assume the most general view on CCS which would still be consistent with quantum mechanics, namely, that CCS comprises any chemical system, i.e., compounds with *any* chemical composition and *any* atomic configuration (being close to some state's energy minimum or not). Such an encompassing definition would sacrifice the minimal free energy requirement mentioned above, and it would trivially correspond to the entire domain of CCS. Therefore, it would forego the useful link to observable lifetimes of systems as well as the appealing complementarity (not to be confused with orthogonality) to the well established problem of sampling potential energy hyper surfaces to study free energies or competing elementary reaction steps.

## 1.2. Machine Learning the Potential Energy Surface

While QM based studies of CCS are mostly concerned with (meta-)stable compounds, from inspection of the electronic Hamiltonian, it is quite clear that the effect of nuclear charges and nuclear coordinates are intimately linked. The well-known cusp condition due to Kato's theorem<sup>36</sup> explicitly links these two variables through the electron density observable. As such, *ab initio* studies of the PES aimed at calculating geometric distortion, transition states, or statistical mechanical averages are closely related to the topic of this review. More specifically, early attempts of QML have focused on the PES of homonuclear system (e.g., diamond<sup>37</sup> or Si<sub>n</sub> cluster<sup>38</sup>) due to its relative simplicity (cf. compositional degree of freedom), for which many QML methods developed are also applicable to CCS. The distinction between CCS and the PES is somewhat arbitrary. For example, some molecular quantities of significant interest, such as libraries of ensemble properties of protein–ligand binding free energies, require accurate potentials as well as representative sampling of CCS. Also, instead of considering (meta-)stable constitutional or conformational isomers as distinct compounds, they can also equally well be viewed as local minima of the global PES hypersurface.

As mentioned above, within studies of the PES, the focus (at least currently) is typically placed on a single system and on computing energies and forces from scratch, i.e., *ab initio*. As such, one does not exploit correlations, constraints, and relationships, which only emerge through relationships observed among constitutional and compositional isomers, i.e., throughout all dimensions of CCS. The most common use-case of quantum methods for atomistic simulations deals with the problem of sampling the configurational degrees of freedom of the atoms of a given system. To develop a better informed understanding of the field, we now also briefly discuss relevant and select machine learning studies which touch upon the quantum based understanding of CCS but which primarily are concerned with the PES.

The question of how to best model a PES using some (physical or surrogate) function approximator and based on scarce and expensive potential energy surface data sets of specific systems, i.e., not through CCS, obtained from computationally demanding calculations, is long-standing. Potential energy hypersurfaces were traditionally studied for the purpose of molecular spectroscopy or for molecular dynamics applications of a given system. The development of empirical interatomic potentials, particularly the reactive force-field (ReaxFF) approach developed by van Duin and co-workers since 2001,<sup>39,40</sup> amounts essentially to a traditional multidimensional regression problem for fixed functional basis functions and constitutes one of the mainstream efforts in this active field. Unlike traditional force field approaches, ReaxFF requires no predefined connectivity between atoms (topology) and casts the empirical interatomic potential within a formalism of bond order, which depends on the interatomic distances only. This improved adaptation of an atom to its environment allows for accurate descriptions of bond breaking and bond formation and has been applied extensively to model reactive chemistry at heterogeneous interfaces, involving typically very large systems,<sup>40</sup> made up of millions of atoms.

The force-field approach, despite its efficiency, its chemical motivation, and its broad applicability and potential accuracy, suffers from the fixed functional forms imposed when relying on empirical interatomic potentials, implying that the model is hard to improve by adding more training data and could even fail catastrophically in certain regimes and classes. This limitation motivates interest in more flexible data-driven models. For example, already in 1994, Ischtwan and Collins improved the Shepard interpolation scheme for PES approximations.<sup>41</sup> This paper illustrates the close relationship to QML: The authors utilized a formalism very similar to the modern kernel ridge regression, one of the workhorses of QML. The authors also already discussed one of the frequent challenges coming along with any new ML model project: How to best down-select optimal configurations for minimal data acquisition and training costs, and how to obtain systematic model improvements with increasing training set size.<sup>41</sup> Early awareness of the trade-off between accuracy and training cost was also addressed more than a decade earlier in the 1981 paper by Wagner, Schatz, and Bowman: Given a finite compute budget, data for which training instances should be acquired in order to obtain the most accurate model of the potential energy hypersurface?<sup>42</sup> When facing the exploration of CCS with QML models, analogous questions must be addressed. For references to similar studies related to the problem of potential fitting and preceding 1989, we refer the reader to the comprehensive review by Schatz.<sup>43</sup>

Most of the data-driven models in the 1990s favored the neural network regressors for PES fitting. More specifically, in 1992, Sumpter and Noid published a neural network model for macromolecules.<sup>44</sup> Additional neural network potentials were published by Blank et al. in 1995<sup>45</sup> for the CO/Ni(111) system, and Brown et al. in 1996<sup>46</sup> for the study of ground-state vibrational properties of two weakly bound molecular complexes: (FH)<sub>2</sub> and FH–ClH. Neural networks were revisited for systems with increased size in the same year by Lorenz, Gross, and Scheffler<sup>47</sup> for H<sub>2</sub>/K(2×2)/Pd(100) (with substrate fixed), followed by their application to represent high-dimensional potential energy surfaces for H<sub>2</sub>O<sub>2</sub> by Manzhos and Carrington in 2006.<sup>48</sup> Even larger systems, i.e., water clusters (up to 6 units), were dealt with by Handley and



Popelier from 2009 onward,<sup>49</sup> accounting for important electrostatic properties through learning of the atomic multipoles. These early developments used Cartesian/internal coordinates directly as the input of NN models, which is justified for modeling small to medium-sized systems; for large systems, however, this setup proves to be too inefficient. In 2007, Behler and Parinello published much improved deep neural network-based potentials,<sup>50</sup> encoding molecular geometry effectively in terms of rotation, translation, and permutation invariant atom-centered symmetry function (ACSF), followed by molecular dynamics applications using metadynamics to identify Si phases under high pressure.<sup>51</sup> A detailed overview of various neural network-based advances since 2010 was given in 2017.<sup>52</sup> Starting in the same year, multiple, more universal, neural network models were introduced. In particular, Smith et al. advanced the idea of Behler's symmetry functions in neural networks with the aforementioned normal mode displacements in order to generate a powerful neural network trained on millions of configurations of tens of thousands of organic molecules, called ANI.<sup>53</sup> An accurate and transferable neural network exploiting an "on-the-fly" equilibration of atomic charges was introduced that same year by Faraji et al.,<sup>54</sup> and soon thereafter, equally universal neural nets SchNet<sup>55</sup> and PhysNet<sup>56</sup> were published. An extensive review on neural network potentials for modeling the potential energy surfaces of small molecules and reaction is also part of the present issue in *Chemical Reviews*.<sup>57</sup>

Kernel models started to play a noticeable role for PES fitting in the late 1990s. In 1996, Ho and Rabitz presented kernel-based models for the fitting of potential energies<sup>58</sup> for three small systems, He–He, He–CO, and H<sub>3</sub><sup>+</sup>. Similar to early PES fitting works within NN framework, these early kernel-based models also utilized simple Cartesian/internal coordinates as input, and therefore applicability was limited. While the mathematics of kernel-based surrogate models was firmly established many decades ago, only from 2010 and onward, kernel-based models began to flourish, building on the seminal work contributed by Csanyi, Bartok, and co-workers within their "Gaussian-Approximated Potential" (GAP) method, relying on Gaussian process regression (GPR) and an atom index invariant bispectrum representation.<sup>59</sup> In 2012, Henkelmann and co-workers introduced an interesting application of support vector machines (SVM) toward the identification of transition states.<sup>60</sup> One year later, the first flavor of the smooth overlap of atomic positions (SOAP) representation for GPR based potentials was published.<sup>38</sup> The SNAP<sup>61</sup> method popularized the GAP idea using linear kernels in 2015, and other GPR applications with automatically improving forces were published the same year.<sup>62,63</sup> Around the same time, a first stepping stone toward a universal force field, trained on atomic forces throughout the chemical space of molecules displaced along their normal modes, was established.<sup>64</sup> Reproducing kernels were also shown in 2015 to be applicable toward dynamic processes in biomolecular simulations,<sup>65</sup> and ever more accurate GPR based potentials were introduced in 2016<sup>66</sup> and in 2017.<sup>67,68</sup> GPR was also applied to challenging processes in ferromagnetic iron<sup>69</sup> and to the problem of the on-the-fly prediction of parameters in intermolecular force fields.<sup>70</sup> Amorphous carbon was studied using SOAP based GPR/KRR models,<sup>71,72</sup> and GDML, another series of highly robust and accurate GPR/KRR based molecular force fields, was introduced in refs 67 and 73–75 starting in 2017.

GPR and NN are currently the two most popular regressors for PES fitting, and each exhibits advantages and disadvantages. Seemingly very different in design, they do resemble each other to some extent in the sense that they take the role of basis functions (to be elaborated in section "Regressor"), although the similarity may be blurred within the framework of deep NN. Numerical comparison of the performance of these two methods is interesting. Most notably, such a comparison was made for modeling the potential energy surface of formaldehyde in 2018 by Manzhos and co-workers.<sup>76</sup> A similar yet independent study on the same system was performed in 2020 by Meuwly and co-workers.<sup>77</sup> Both studies confirm that kernel based QML models reach higher predictive power than neural network based models for same training set sizes. A highly related comparative study on modeling vibrations in formaldehyde was contributed by Käser et al.,<sup>77</sup> also in 2020.

As for the active learning of interatomic potentials, most of the related studies relied on the kernel framework, some of them also detailed below in the section "Training Set Selection". As early as in 2004, De Vita and co-workers proposed updating potential parameters to ab initio results during molecular dynamics runs ("learning-on-the-fly"),<sup>78</sup> for a very large system, i.e., silicon systems composed of up to ~200 000 atoms, although the reference level of theory is quite approximate. Podryabinkin and Shapeev proposed the so-called *D*-optimality criterion for selecting the most representative atomistic configurations for training on-the-fly as early as 2017.<sup>79</sup> Using kernel ridge regression (KRR, a variant of GPR), Hammer and co-workers revisited the on-the-fly learning idea for structural relaxation in 2018,<sup>80</sup> and investigated the exploration vs exploitation trade-off.<sup>81</sup> In 2019, Weinan, Car, and co-workers contributed another active learning procedure for accurate potentials of Al–Mg alloys,<sup>82</sup> and Westermayr et al. extended the use of neural networks for molecular dynamics in the electronic ground state toward photodynamics.<sup>83</sup> Among the many purposes (also challenges) of QML for PES, one particular one is to scale to an extremely large (thus more realistic) system. Numerous efforts have pushed us closer to this goal, and most notably, Weinan, Car, and co-workers made full use of the Summit supercomputer to simulate 100 million atoms with ab initio accuracy using convolutional neural networks,<sup>84</sup> for which they subsequently were awarded the Gordon Bell prize 2020 by the *Association for Computing Machinery*.

### 1.3. Navigating CCS from First Principles

The scientific research question of how properties trend across CCS lies at the core of the chemical sciences. Because of ever-improving hardware performance, improved approximations to Schrödinger's equation, most notably within density functional theory and localized coupled cluster theory, QM data sets of considerable size have emerged, enabling the use of statistical learning to train surrogate QML models which can provide accurate and rapid quantum property estimates for new compounds within their applicability domain.

While quantum mechanics based computational materials design efforts had been undertaken as early as the 1990s,<sup>85–88</sup> with important progress made in the 1980s,<sup>89,90</sup> the first-principles based computational high-throughput design has by now become an important success story.<sup>91</sup> First attempts to employ machine learning and quantum predictions to discover new ternary materials databases date back to seminal work by Hautier and Ceder in 2010.<sup>92,93</sup>

As a promising alternative to *ab initio* high-throughput computations (or solving Schrödinger equation in general), one often assumes locality of atoms in molecules when constructing the mapping from molecular distance/similarity to difference in properties within QML, and the final predictive performance depends on how similar two local (and thus global) entities are, i.e., nuclear types covered by a test set are required to be retained in the training set. The capability of QML to treat species made up of elements not seen in training set is, however, very limited. There exists the so-called “alchemical” methods, being quite different in philosophy, allowing for effective and efficient treatment of the change of nuclear types, with or without the constraint that the number of electron number ( $N_e$ ) being fixed (i.e., isoelectronic). We note in passing that alchemy is typically established within the density functional theory framework, as it would be tremendously simpler to expand molecular property (mostly energy) as a function of four variables ( $x$ ,  $y$ ,  $z$ , and  $Z$ ) than  $4N_i$  ones in the case of wave function-based formulation.

Previous methodological works tackling chemical compound space from first-principles through variable (“alchemical”) nuclear charges were contributed by various pioneers, including Wilson’s formal four-dimensional density functional theory<sup>94</sup> in 1962, which expresses the exact nonrelativistic ground-state energy of an electronic system as a functional of the electron density, which per se is a function of the spatial coordinates, and nuclear charges. Following Wilson’s idea, Politzer and Parr<sup>95</sup> in 1972 made one step further toward practical computation by transforming Wilson’s formula into a functional of the total electrostatic potential  $V(r, Z)$  and derived some useful semiempirical formulas for the total energy of atoms and molecules, through the use of thermodynamic integration.<sup>95</sup> Later in the 1980s, Mezey made some interesting discoveries<sup>96,97</sup> about the global electronic energy bounds for a variety of isoelectronic polyatomic systems, which may be found useful for quantum-chemical synthesis planning, using multidimensional potential surfaces.

The theoretical alchemical research was resurrected in the new millennium. Among the numerous contributions, notable ones include a variational particle number (variable proton and electron number) approach for rational compound design<sup>90</sup> proposed by one of the authors and collaborators, followed by a more detailed description of the underlying theories, in the name of molecular grand-canonical ensembles (GCE).<sup>98</sup> In the same year, a reformulation of GCE in terms of linear combinations of atomic potentials (LCAP)<sup>99</sup> (instead of  $Z$  and  $N_e$  as in GCE) was proposed by Wang et al., but for the optimization of molecular electronic polarizability and hyperpolarizability, with the optimal molecule determined analytically in the space of electron–nuclei attraction potentials. For the isoelectronic case, related works include the development of *ab initio* methods for the computation of higher-order alchemical derivatives<sup>100</sup> by Lesiuk et al. in 2012, as well as the assessment of the predictability of alchemical derivatives<sup>101</sup> by Munoz et al. in 2017. More recently, alchemical normal modes in CCS,<sup>102</sup> alchemical perturbation density functional theory,<sup>103</sup> and even a quantum computing algorithm for alchemical materials optimization<sup>104</sup> were proposed, further enriching the field.

Starting in 1996 with stability of solid solutions,<sup>105</sup> multiple promising applications, based on quantum alchemical changes, have been published over recent years, including thermody-

amic integrations,<sup>106</sup> mixtures in metal clusters,<sup>107,108</sup> optimization of hyperpolarizabilities,<sup>109</sup> reactivity estimates,<sup>110</sup> chemical space exploration,<sup>111</sup> covalent binding,<sup>112</sup> water adsorption on BN-doped graphene,<sup>113</sup> the nearsightedness of electronic matter,<sup>114</sup> BN-doping in fullerenes,<sup>115</sup> energy and density decomposition,<sup>116</sup> catalyst design,<sup>117–119</sup> and protonation energy predictions<sup>120,121</sup> Symmetry relations among perturbing Hamiltonians have also enabled the introduction of “alchemical chirality”.<sup>34</sup>

An extension of computational alchemy toward descriptions which go beyond the Born–Oppenheimer approximation has been introduced within path-integral molecular dynamics, enabling the calculation of kinetic isotope effects, already in 2011,<sup>122</sup> and subsequently by Ceriotti and Markland.<sup>123</sup>

However, also varying the electron number is a longstanding concept within conceptual DFT.<sup>124,125</sup> Actual variations have only more recently been considered, e.g., to estimate redox potentials,<sup>98,126</sup> higher-order derivatives,<sup>100–102</sup> or for the development of improved exchange-correlation potentials.<sup>127</sup>

## 2. HEURISTIC APPROACHES

Modern systematic attempts to establish quantitative structure–property relationships (QSPRs) have led to computationally advanced bio-, chem-, and materials-informatics methodologies. Unfortunately, conventional approaches in QSPR predominantly rely on heuristic assumptions about the nature of the forward problem, and are thus inherently limited to certain applicability domains. The implicit bias, often due to lacking basis in the underlying physics is known, as discussed, e.g. in a 2010 review by G. Schneider,<sup>128</sup> and many improvements have been contributed more recently.<sup>20</sup>

While heuristic in nature, QSPR can still provide useful qualitative trends and insights for relevant applications, and sometimes yield accurate predictions for specific property subdomains and systems. Albeit not directly relying on the laws of quantum mechanics, these early developments are still valuable, in the sense that some just correspond to special variants of the more complicated models, for instance, a linear model can be mapped onto the framework of kernel method, by choosing a linear kernel, instead of say Gaussian kernel for Gaussian process regression (GPR). Other heuristic approaches, exhibiting more quantitative characteristics can be considered important precursors for modern QML. Such examples include Collin’s improved Shepard interpolation scheme<sup>41</sup> for accurate representation of molecular potential energy surfaces, which resembles the form of kernel methods except that the weights are determined in a heuristic way, instead of being regressed as in GPR. One may also argue that Collins’ scheme could be recast into the kernel framework, except that a specific kernel is chosen such that the Shepard interpolation weights in Collins’ scheme are exactly reproduced (with the constraint that these weights sum up to 1). Another highly related concept is Ramon Carbo-Dorca’s quantum similarity (for a comprehensive review, see ref 129), derived based on density matrix, or molecular orbitals, or other related quantum quantities, it is also closely linked to kernel based methods and may be used directly as parameter-free kernel matrix elements (unlike in GPR, kernel matrix element characterizing similarity is typically hyper-parameter dependent).

In the sections below, we focus on relevant literature regarding three distinct perspectives which largely follow

chronological order: (i) low-dimensional correlations or simple models from the early days of chemistry, (ii) coarse representations of molecules and derived quantities, mostly providing an overview of QSPR, and (iii) molecular representations based on properties.

### 2.1. Low-dimensional Correlations

Early practices of fundamental chemical research dealt with spotting correlations between inherent properties of the system and systematic changes of observed quantities. Possibly the most famous example for such work is Mendeleev's discovery of the periodic table.<sup>130</sup> Other important examples correspond to Pauling's electronegativity concept and covalent bond postulate,<sup>131</sup> or Pettifor's Mendeleev numbering scheme.<sup>132,133</sup> Work along such lines has been continued, and recent contributions include revisiting Pettifor scales,<sup>134,135</sup> use of variational autoencoders to "rediscover" the ordering of elements in the periodic table,<sup>136</sup> or the chemplitude model which extends Pauling's concept,<sup>137</sup> among many others. Free-energy relationships are the subject of yet another broad category of early research which is still active today. Relating logarithms of reaction constants (free energy difference) across CCS for related series of reactions<sup>138</sup> has led to the famous Hammett equation, a 2D projection of all degrees of freedom onto composition and reaction conditions.<sup>139–141</sup> Similarly low-dimensional effective degrees of freedom have been identified within Hammond's postulate,<sup>142</sup> or Bell–Evans–Polanyi principle.<sup>143,144</sup>

Most of the aforementioned concepts were proposed to gain a better (or more useful) understanding of molecular behavior in the first place. For extended systems such as metallic surfaces, complexities arise and many of the simplified molecular models are no longer applicable. With the advent of density functional theory (DFT),<sup>145–149</sup> alternative descriptors have been proposed during the past decades, playing an increasingly important role. Notable contributions include the *d*-band center model by Hammer et al.,<sup>150,151</sup> the generalized coordination number,<sup>152,153</sup> and the Fermi softness.<sup>154</sup> Free-energy relationships are more robust against subtle changes in the electronic structure and are being widely applied in analyzing surface elementary reaction steps.<sup>155</sup> Scaling relations between the energetics of adsorbed species on surfaces<sup>156</sup> also enjoy extensive attention and have been proven useful for catalyst design regardless of the surface not being metallic.<sup>157–159</sup> Many of the empirical chemical concepts such as electronegativity, softness/hardness, and electrophilicity/nucleophilicity can be rationalized and quantified within what is known as "conceptual" DFT.<sup>160,161</sup> This specific field, as pioneered by Fukui or Parr and Yang,<sup>160</sup> has been championed and furthered by many including Geerlings, De Proft, Ayers, Cardenas, and co-workers.<sup>161,162</sup>

We note that simple models, involving one or a few variables in general, represent effective coarse-grained schemes applicable to specific subdomains of chemistry. While they lack the desired transferability of quantum mechanics, they often do encode well tempered approximations and therefore are capable of capturing much of the essential physics. As such, they have much to offer, and they could, for example, serve the design of robust and general representations enabling the training and application of improved QML models (see below or refs 6 and 163). Alas, this idea, to connect low-dimensional model, based on well established heuristics, with more recently developed generic ML models, is still largely unexplored,

despite the fact that the latter often bear (magic) black box characteristics allowing for little qualitative insights. Unifying modern ML with low-dimensional model could therefore also help resolve open challenges in QML. For instance, how can we properly represent different electronic (spin-) states of molecules in the molecular representation or different oxidation states? Conceptual DFT derived linear or quadratic energy relationships suggest treating the number of electrons ( $N_e$ ) and/or its powers as independent features might be a reasonable starting point. Another thus-inspired direction of research is to utilize conceptual DFT-based local indicators as properties of composing atoms/bonds/fragments of a target molecule as a starting point (much like the fundamental variables such as *Z* and *R*) for building representations. This might be necessary in order to address hard and outstanding problems such as building QML models of intensive properties or to account for multireference character in the electronic structure.

### 2.2. Stoichiometry

Given a fixed pattern of structure, stoichiometry alone can be used as a unique representation of the system under study. This idea has been demonstrated for an exhaustive QML based scan of the elpasolite ( $ABC_2D_6$  stoichiometry) subspace of CCS, predicting cohesive energies of all the 2 million crystals made up from main-group elements.<sup>164</sup> Elpasolites are the most abundant quaternary crystal form found in the Inorganic Crystal Structure Database. Comparison of the QML results to known competing ternary and binary phases enabled favorable stability predictions for nearly 90 crystals (convex hull) which subsequently have been added to the Materials Project database.<sup>165</sup> A compact stoichiometry based representation in terms of period and group entry for elements A, B, C, and D was shown to reach the accuracy of explicit geometry based many-body potential representations at larger training set size, indicating the dominance of the former in large training data regimes.<sup>166</sup> Similar work was subsequently done by Ye et al.<sup>167</sup> as well as Marques et al. for perovskites on crystal stability,<sup>168</sup> as well as by Legrain et al.,<sup>169</sup> for predicting vibrational free energies and entropies for compounds, drawn from the Inorganic Crystal Structure Database.

A naive but useful derived concept is the so-called "dressed atom" concept,<sup>170</sup> which characterizes the atom in a molecule of a fixed stoichiometry. When using this approach together with a linear regression model to approximate the total energy (or atomization energy), the accuracy turns out to be surprisingly reasonable,<sup>171</sup> at least for common data sets of organic molecules with small variance among constitutional isomers. For instance, the corresponding mean absolute error (MAE) for QM7 data set is *only* 15.1 kcal/mol.<sup>170</sup> Using bond counting, the MAE could be improved further to less than 10.0 kcal/mol, within reach of a conventional DFT GGA functionals.<sup>172</sup> Therefore, it seems advisable to always use the dressed atom approach for centering the data for any fixed stoichiometry (i.e., averaging out constitutional and conformational isomers) before proceeding to the next level of QML training on the complete set of degrees of freedom. This normalization step can also be seen as data preprocessing, enabling the QML model to focus on "minor" deviations from the mean.<sup>164,173</sup>

### 2.3. Connectivity Graph

When the systems under study do not share some common structural skeleton, stoichiometry alone is not enough, and the



covalent bonding connectivity between atoms, as well as conformations, may have to be also examined.

It is worth pointing out that chemists often assume a one-to-one relationship between the molecular graph and its associated global conformational minima (or the second lowest energy minima, or the third, etc.), and therefore it should be possible to build a QML model to predict relevant quantum properties of such ordered minima from graph-input only. In fact, the remarkable performance of extended Hückel theory for some systems could be explained in this way.

Because of the intuitive accessibility and applicability of (incomplete) graph based representations, such as Lewis structures and their extensions, for a wide range of molecular systems, associated ML methods have received broad attention and wide applications in many fields such as cheminformatics or bioinformatics. Examples of such representations include various fingerprint representations,<sup>174</sup> such as the signature methodology.<sup>175–177</sup> Another notable example corresponds to the so-called extended circular fingerprint (ECFP).<sup>178</sup> ECFP and similar representations have been used for drug design<sup>179,180</sup> and qualitative exploration of CCS.<sup>181,182</sup> ECFP has also been used in KRR models for prediction of quantum properties of QM9 molecules. Numerical results for ECFP based QML models indicate a substantially worse performance compared to more complete, geometry derived representations.<sup>183</sup>

Modern molecular graphs, typically in SMILES format, based neural networks models, have gained considerable momentum during the past decade. A vast amount of related literature deal with chemical synthesis and retrosynthesis in such representation spaces (mainly in organic chemistry),<sup>184,185</sup> typically favoring different deep learning architectures, chemical reaction network,<sup>186</sup> as well as molecular design using variational autoencoder (VAE, which maps a molecule represented by SMILES string to some latent space<sup>187</sup>). The absence or presence of relationships between functional groups and binding affinity was also recently explored through use of random matrix theory in drug design.<sup>188</sup> The incorporation of new and improved formats, such as SELFIES,<sup>189</sup> might still lead to further improvements for such research.

In the context of a first-principles view on CCS, we note however that molecular graphs only encode a (biased) statistical average of the many conformational configurations for a molecule near some local minima in the potential energy surface. As such, they are naturally disposed for use of QML models of ensemble properties. Work along such lines still awaits being explored in the future. Albeit popular and justified for certain problems, graph-based approaches are inherently limited when it comes to noncovalent problems, such as supramolecular assembly processes governed by van der Waals interactions, metal cluster/bulk/surface adsorption involving “multivalent” (transition) metal elements controlled by weaker metallic bond (cf., covalent bond), or chemical reactions requiring the transformation of graphs from one into the other. In such situations, the intuitive concept of a graph is ill-defined, and the necessary corrections are not always obvious.

#### 2.4. Coarse-grained

As the system size grows, the cost in training and prediction of QML models increases accordingly, although with more favorable scaling than typical quantum chemistry methods. Therefore, it may become very demanding or even impossible to deal with system sizes which cross certain thresholds. In

such scenarios, one typically represents the systems in a coarse-grained fashion, meaning “superatoms” (groups of atoms in close proximity, or beads) in a molecule are being considered. Coarse-grained approaches can drastically reduce the number of degrees of freedom and are therefore the only feasible way to model systems at macroscopic scale. More importantly, they enable a significant collapse of the size of chemical space due to the transferability of beads by design.<sup>190</sup>

Current practices of coarse-grained ideas comprise mostly coarse-grained force fields (CGFF) for simulation of large systems such as macromolecular systems and soft matter. With the emerging need for systematic control of the accuracy of models of such systems, coarse-grained representation based QML models (CGQML) may be a rather promising alternative to CGFF, much the same as how QML models based on full-atom representation remedy the deficiencies of classical force field approaches for small to medium-sized molecules.<sup>19</sup> Such comparison between QML and FF makes sense, as the most modern implementations of ML hold promise to approach the computational efficiency of FFs. Some of the first studies on coarse-grained representation used together with QML include John and Csányi’s free energy surface modeling of molecular liquids in 2017.<sup>191</sup> Later, efforts to tackle complicated biosystems were reported by Bereau and co-workers in 2019,<sup>190</sup> as well as by Clementi and co-workers.<sup>192</sup> Compared to CGFF, CGQML could be significantly more accurate once the system information is properly encoded in the coarse-grained representation, as the QML part can recover what is missing in the CG part by careful selection of training data (*vide infra*).

#### 2.5. Property Based

There exists another type of representation, typically referred to as descriptor and the least “ab initio” in spirit, in which the basic idea is to simply select a set of pertinent atomic/molecular properties as underlying degrees of freedom. The properties can stem from calculation and/or experiment and have to be relatively easy to obtain and are typically supposed to somehow “describe” the property of interest, and hence the name “descriptor”. This representation is often utilized in combination with some nonlinear regressor like a neural network, as the relationship between the chosen properties is commonly highly nonlinear. Although this approach could be universally applicable, no matter the size or composition of target systems, its predictive power is limited by construction due to its potential lack of uniqueness.<sup>30,163</sup> Most of the studies following this direction can be traced back to the early applications of ML in chemistry and related fields, one example being Karthikeyan et al.’s work<sup>193</sup> on melting and boiling points prediction of molecular crystals using the properties of standalone molecules as a feature vector. A more recent and systematic study of this idea has applied optimization algorithms toward the down-selection of descriptor candidates in order to build predictive ML models of formation energies of binary solids.<sup>194</sup> From a first-principles point of view, however, such representations are questionable because relationships between different observables (or other arbitrary mathematical properties), obtained as expectation values of independent operators, are not necessarily well-defined.

### 3. QML METHODOLOGY

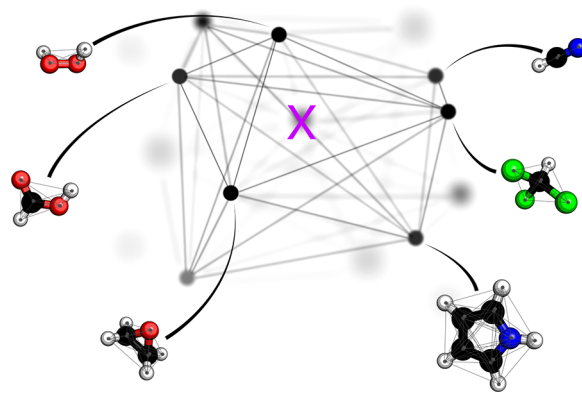
The fundamental idea to employ machine learning models in order to *infer* solutions to Schrödinger’s equation throughout

CCS, rather than solving them numerically, was first introduced in 2011.<sup>195</sup> The authors stated that "...the external potential...uniquely determines the Hamiltonian  $H$  of any system, and thereby the ground state's potential energy by optimizing  $\Psi$ ", and they show that one can use QML instead (encoding the number of electrons implicitly by imposing charge neutrality). As such, the problem of predicting quantum properties throughout CCS belongs to what is commonly known as "supervised learning". One typically distinguishes between unsupervised (compound data only) and supervised (data records including compounds and associated properties) learning. In this review, we focus on the latter, i.e., on the question how, given sufficient exemplary structure–property pairs, properties can be inferred for new, out-of-sample compounds.

The generic procedure for supervised learning requires first defining the model architecture, i.e., the mathematical expression for the statistical surrogate model  $f$ , which estimates some quantum property  $p$  as a function of any query compound  $\mathbf{M}$ ,  $p^{\text{QML}}(\mathbf{M}) \approx f(\mathbf{M}|\{\mathbf{M}_i\};\{p_i^{\text{ref}}\};\{c_i\})$ , where  $f$  corresponds to the regressor, and regression coefficients and hyper parameters  $\{c_i\}$  are obtained via minimization of training loss-function quantifying the deviation of  $p^{\text{QML}}$  from  $\{p_i^{\text{ref}}\}$  for all training compounds  $\{\mathbf{M}_i\}$ . In other words,  $f$  is parametric in regression coefficients and hyperparameter which, in return, are nonlinear functions in the training data. The origin (calculated or measured) as well as the actual existence (some properties, such as energies of atoms in molecules,<sup>116</sup> are not observables but can still be inferred) of  $p^{\text{ref}}$  is secondary. Noise in the data (due to experimental or numerical uncertainty, or due to minor inconsistencies) can be accommodated to a certain degree through well-established regularization procedures. Converged cross-validation protocols help to avoid overfitting and to enable the optimization of hyper-parameters as well as meaningful estimates for any interpolative query. For introductory texts on kernel based regressors, the reader is referred to the book by Rasmussen et al.;<sup>196</sup> as for representation and training sets, several reviews have recently been published.<sup>7,10,19,197</sup>

### 3.1. Regressor

When considering the problem of fitting a generic set of basis functions to precalculated data, some of the most commonly made choices in the field of atomistic simulation include support vector machines (kernel ridge regression), tantamount to Gaussian process regression in their specific model form, neural networks, random forests, or permutationally invariant polynomials (PIPs).<sup>196,198,199</sup> While agnostic about the training labels by construction, the choice of these basis set expansions constitutes a crucial step. Most evidently for support vector machines, nonlinear kernel functions (based on feature representations vide infra) map any nonlinear high-dimensional regression problem into a low-dimensional kernel space within which the regression problem becomes linear and therefore straightforward to solve through a closed-form expression ("kernel-trick"). How kernel space relates to CCS is also quite intuitive to grasp when thinking about it as a graph of compounds. As displayed in Figure 2, each compound, being representable by a molecular graph (or derived matrix such as a Coulomb matrix or Cartesian coordinate and nuclear charge vector) is projected into higher-dimensional feature space (shown are only three principal dimensions from the infinite number of dimensions defined within the framework of



**Figure 2.** 3D projection of high-dimensional kernel representation of chemical compound space. Within kernel ridge regression, chemical compound space corresponds to a complete graph where every compound is represented by a black vertex and black lines correspond to the edges which quantify similarities. Each compound, in return, can be represented by a molecular complete graph (e.g., the Coulomb matrix (CM)<sup>195</sup>) recording the elemental type of each atom and its distances to all other atoms. Given known training data for all compounds shown, a property prediction can be made for any query compound as illustrated by X. Choice of kernel-function, metric, and representation will strongly impact the specific shape of this space and thereby the learning efficiency of the resulting QML model.

KRR/GPR). The complete connection between all compounds in the new space form another type of graph, with each vertex corresponding to a compound and each edge corresponding to a similarity measure of compounds (edge length may indicate a metric distance between two compounds). Inferring the property of a new compound (labeled as pink "X" in Figure 2) may be conceptualized as summing up distance scaled property weights. Within this picture, it becomes intuitively obvious that the interpolating accuracy must improve with increasing compound density.

While deep neural network models are very powerful and possess significant black-box character, their training requires data sets of very large size as well as a substantial calculation effort in order to optimize the regression coefficients and hyper-parameters (no closed-form solution is known). In this sense, kernel methods are rather lightweight and preferable in scarce data scenarios, as they enjoy the potential benefits of being more intuitive and faster to train. The specific architecture of the neural network will affect its performance and data efficiency dramatically. Deep, recurrent, convolutional, message passing, generative, adversarial, geometric neural networks, and other flavors, as well as choices of activation function, number of layers, and neurons, have all shown significant impact on the cost of training and on the predictive power in atomistic simulation.

In the case of GPR/KRR, the architecture is much simpler and hence of a lesser concern (GPR/KRR can be seen as a single-layer neural network model in the limit of infinite width<sup>200</sup>), but the specific kernel space does not only depend on the choice of kernel function but also on the choice of metric. While it is clear that one should avoid similarity measures which do not meet the mathematical criteria of how a metric is defined (identity, symmetry, triangle inequality), the impact of the specific metric choice has not yet been studied much in the field of atomistic simulation. Euclidean, Manhattan, or Frobenius norms are commonly used. Only most recently, the use of the Wasserstein norm has been

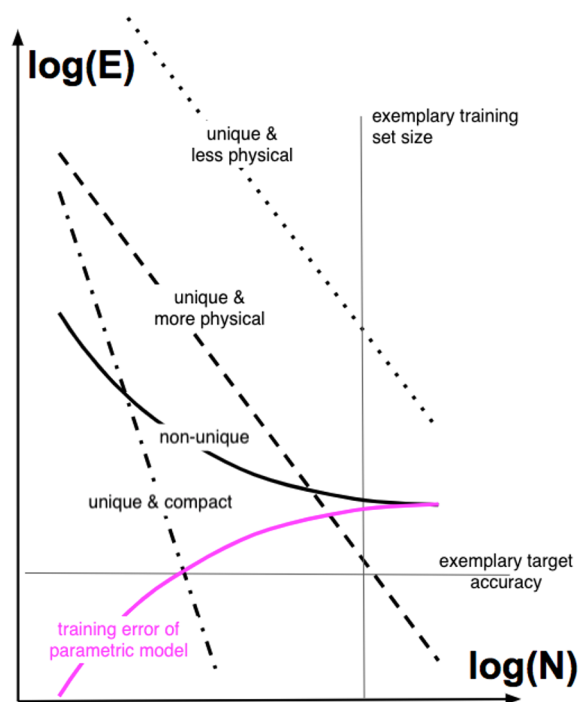


proposed to gain permutational atom-index invariance while using index-dependent matrix representations.<sup>201</sup> From inspection of Figure 2, it should be obvious that any nonlinear change in metric will strongly affect the shape of the kernel regression space and thereby the overall performance.

### 3.2. Learning Curves

Correct implementations of ML algorithms applied to noise-free data sets afford interpolating ML models, which avoid overfitting and enable statistically meaningful predictions of properties of out-of-sample compounds,<sup>198</sup> after proper regularization and hyper-parametrization through converged cross-validation protocols, as discussed in great detail in the literature, for example, in refs 202 and 203.

On the basis of statistical learning theory, the leading order term of the out-of-sample prediction error ( $E$ ) was shown to decay inversely with training set size  $N$ , i.e.,  $E \propto a/N^b$ , for GPR/KRR as well as for neural network models.<sup>204,205</sup> This is not surprising, considering the great similarities shared between NN models and GPR/KRR model, as also mentioned in the preceding subsection. This asymptotic behavior for QML models has been confirmed numerically within numerous and independent studies, many of which are referenced herein. As illustrated in Figure 3, learning curves



**Figure 3.** Illustration of learning curves: Errors ( $E$ ) versus training set size ( $N$ ). Horizontal and vertical thin lines illustrate exemplary target accuracy and available training set size, respectively. For functional ML models, training errors are close to zero (not shown), and prediction errors must decay linearly with  $N$  on log–log scales. Black-solid, dotted, dashed and dotted-dashed lines exemplify prediction errors of ML models with incomplete information (ceases to learn for large  $N$  due to being parametric, using nonunique representations, or training on noisy data), unique and less physical representation, unique and more physical representation, and explicit account of lowered effective dimensionality (i.e., “compact”), respectively. The solid-pink line corresponds to the training error for a parametric model. Training errors for ML models are negligible for noise-free data.

(LC), i.e., prediction error  $E$  vs training set size  $N$ , plotted on log–log scales assume linear form ( $\log E = \log a - b \log N$ ) and serve as a useful standard, facilitating systematic comparison and quality control of the efficiency of differing ML models. For maximal consistency, the QML models should be trained and tested on the exact same cross-validation splits stemming from the exact same data set. When the data contains noise, or when relevant degrees of freedom are neglected (e.g., through use of a nonunique representation, such as the bag of bond (BoB) representation,<sup>170</sup> see section 4.1 for more details), the learning will cease eventually for some training-set size, manifesting itself visually through learning curves which level off, cf. solid-black line in Figure 3. For noise-free data and complete representations, however, a linear correlation between  $\log(E)$  and  $\log(N)$  is to be expected (see the dotted and dashed lines in Figure 3), with some slope  $b$  typically more or less a constant for different unique representations and related to the effective dimensionality of the problem, and some offset  $\log a$ , which typically reflects the capability of the representation to capture the most relevant feature variations in kernel space. More specifically, the offset measures the degree to which the representation encodes the right physics. An illustrative example for this statement can be given by comparison of the learning curves obtained for the CM representation versus derived matrices with off-diagonal entries dependent on alternative interatomic power-laws.<sup>163</sup> For interatomic off-diagonal elements approaching London’s  $R^{-6}$  law, the representation achieved lower offsets than for off-diagonal elements decaying according to Coulomb’s law. Correspondingly, representation matrices with off-diagonal elements linearly or quadratically growing with the interatomic distances resulted in LCs with dramatically increased offsets.<sup>163</sup> At first glance, it might seem that the slope of LC (aka, “learning rate” of QML model) barely changes, when switching from one unique representation to another. It is therefore natural to ask if it is impossible to further speed up the learning process as indicated by the dotted-dashed learning curve in Figure 3, exhibiting a much steeper slope. Through an expert-informed reduction of effective dimensionality (through a priori removal of irrelevant information stored in randomly selected training data), it was shown that this is indeed possible. Such strategies for a more rational sampling of training data will be discussed in section 6. Note that in contrast to conventional curve fitting, training errors for properly trained machine learning models applied to synthetic data are typically orders of magnitude smaller than the variance of the signal. As such, they are negligible and carry little meaning because noise levels are typically close to zero or at least many orders of magnitude smaller than label variance. Consequences of model construction, i.e., choice of regressor, metric, optimizer, loss-function, representation, or computational efficiency, become immediately apparent in the characteristic shape of learning curves. When training a small parametric regressor, e.g., a shallow neural network with few neurons, to estimate a complex and high-dimensional target function, the learning curve will rapidly “saturate” and converge toward a finite optimal residual prediction error that can no longer be lowered by mere increase of training set size. As such, it should come as no surprise that learning curves have emerged as a crucial tool for development, validation, comparison, and demonstration purposes of QML models in the field.

### 3.3. Loss Functions

Imposing differential relationships during training amounts to adaptation of the loss function to better reflect the problem at hand. In particular, inclusion of derivative information (gradients and Hessian) has led to dramatic improvements when tackling the problem of potential energy surface fitting.<sup>66,67,73,74</sup> A generalization of this idea to adapt the loss function for response properties of any QM observable was established for KRR in 2018<sup>206–208</sup> (exemplified for forces, Hessians, dipole moments, and IR spectra) and for deep neural nets in 2020 (FieldSchNet exemplified moreover also for solvent effects and magnetic effects).<sup>209</sup>

While conventional machine learning assumes that train and test loss function are identical, for atomistic simulation (or other application domains for that matter), a mathematically, more “greedy” alternative might exist. In particular, the role of gradients in loss functions differing for training and testing has been studied in ref 210, with results suggesting that for predicting atomization energies throughout a CCS of distorted structures, inclusion of gradients in training improves learning curves negligibly while surely inflating the number of necessary kernel basis functions. However, when it comes to predicting the potential energy surface of a given system, they do improve the energy predictions in the above referenced studies. Conversely, when predicting gradients throughout CCS, the use of energies alone in training offers no advantage over using forces, suggesting that the inclusion of forces (if computationally less demanding than energies) should always be beneficial.

## 4. REPRESENTATIONS

One could consider the choice of functional form of the representation  $\mathbf{M}$  to be part of the machine learning methodology. However, this is a much studied question which is at the heart of how one views CCS. More specifically, what are the truly defining aspects in a compound? And how does one measure similarity? These are old questions which have already been answered for an impressive array of applications and instruct much of the basic and fundamental textbook knowledge. For example, Hammett’s  $\sigma$ -parameter provides a low-dimensional quantitative data-driven measure of similarity between distinct functional groups in terms of their impact on reaction rates or yields.<sup>211,212</sup> Within QML, physically more motivated representations are sought after for subsequent use within high-dimensional nonlinear interpolators which are more universal and transferable. As illustrated for KRR in Figure 2, also the specific form of the representation (as well as the metric used) can dramatically affect the way CCS is represented within the regressor. It should therefore not come as a surprise that the data efficiency of QML models was found to depend dramatically on the specifics of the representation used. Because the importance of the choice of the distance measures has already been mentioned above, in this section, we will focus on research that was done to find improved representations.

The choice of this particular compound representation, aka descriptor or feature, plays a particularly crucial role. Correspondingly, substantial research on the design of descriptors has already been made in the fields of chem-, bio-, or materials informatics where scarce data is typical.<sup>174,213</sup> Often, a large set of prospective features is hypothesized and subsequently reduced within iterative procedures in order to distill the most relevant variables and low-dimensional

projections pertinent to the problem at hand (see above). While it is certainly possible to also pursue this approach within a quantum mechanical description of CCS,<sup>194</sup> its heuristic and speculative character remains as unsatisfactory as its lack of universality and transferability. Fortunately, the quantum nature of CCS allows us to follow more systematic and rigorous procedures in order to address this question.

For example, it is a necessary condition for any successful ML model to rely on uniqueness (or completeness) in the representation, as pointed out, proven and discussed several years ago in refs 214 and 215 and more recently in refs 216 and 217, Uniqueness is essential in order to avoid the introduction of spurious noise due to uncontrolled “coarsening” of that subset of degrees of freedom which is neglected. Molecular graphs based on covalent bond connectivity only, for example, do not account for conformational degrees of freedom. Consequently, their use as a representation will make it impossible to quench prediction errors below the variance of the target property’s conformational distribution, no matter how large the training set.

Other characteristics, desirable for representations to display, include compactness, computational efficiency, symmetries, invariances, and meaning. Representations, in conjunction with the regressor’s functional form, define the basis functions in which properties are being expanded and strongly affect the shape of the learning curves, e.g., accounting for a target property’s invariances through the representation typically leads to an immediate decrease of the learning curve’s offset.

While it is possible to model all QM properties using the same representation and kernel,<sup>218</sup> as also demonstrated for neural nets with multiple outputs already in 2013,<sup>219</sup> it should be stressed, however, that this is a distinct feature of QML which stands in stark contrast to conventional QSAR or QSPR, where the ML model is typically strongly dependent on the target property. If regressor, metric, and representation  $\mathbf{M}$  are independent of the label, i.e., the quantum property, there is a strict analogy to quantum mechanics in the sense of the Hamiltonian (or the wave function) of a system not depending on the operator for which the expectation value of any given observable is calculated.<sup>8</sup> This becomes obvious by considering the training of a KRR model where the regression coefficients are obtained through inversion of the kernel matrix,  $\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{p}^{\text{ref}}$ , where for synthetic calculated data with signals being orders of magnitude smaller than noise, the regularization  $\lambda$  (also known as noise level) is typically close to zero. Using property independent representations, metrics, and kernel functions, it is therefore obvious that the regression coefficients adapt to each property only because of the reference property vector  $\mathbf{p}^{\text{ref}}$ . In ref 218, this has been illustrated numerically by generating learning curves for various properties using always the same inverted kernel matrix for any fixed training set size.

The predictive accuracy for specific properties varies wildly as a function of representation and regressor choice.<sup>183</sup> The historic development over years 2012–2018 for a selection of ever-improved machine learning models (due to improved representations and/or regressor architectures) can be exemplified for the prediction errors of atomization energies stored in the QM9 data set<sup>171</sup> and has also recently been summarized in the context of the “QM9-IPAM-challenge” in refs 15, 16, and 18.

The inclusion of increasingly more (less) “physics” in the representation has been demonstrated to systematically

improve (worsen) learning curves<sup>163</sup> and has been followed by a series of developments which have all been benchmarked on the same set of atomization energies of small organic molecules in the QM9 data set<sup>171</sup> and which demonstrate the progress made. While binding energies of “frozen” geometries still constitute an application somewhat remote from most real-world applications in chemistry, from a basic physics point of view they do represent a crucial intermediate step before tackling more complex properties. In other words, if machine learning models failed to predict binding energies, one should not expect them to work for free energies. But also from a practical point of view, the computational cost of single-point energy calculations typically dominates all quantum chemistry compute campaigns and therefore represents one of the most worthwhile targets for surrogate models used for the navigation of CCS.

We note that with the emergence of deep neural networks, the problem of also “learning” the representation can be mitigated to be incorporated in the overall learning problem.<sup>55,220</sup> While many intriguing and sophisticated representations, such as Fourier-series expansions,<sup>215</sup> wavelets,<sup>221</sup> multitensors,<sup>222</sup> or molecular orbitals<sup>223</sup> have been proposed, most representations can be categorized to either correspond to discrete adjacency matrices or to continuous many-body expansions through distribution functions. We therefore limit ourselves to discuss in the following, both predominantly in the context of KRR based QML models. A comprehensive overview on representations for KRR based QML models has also recently been contributed by Rupp and co-workers.<sup>224</sup>

#### 4.1. Discrete

Coordinate-free, bonding neighbors (covalently bonded atom pairs) based graphs, as well as their systematic extensions to arbitrary number of neighboring shells, have formed an important research direction in cheminformatics for many years.<sup>32,174,176,178,213</sup> In 2011, supervised learning was proposed as an alternative to solving Schrödinger’s equation throughout a chemical compound space relying on a representation on a complete undirected labeled graph that encodes the simplex spanned by all atoms.<sup>195</sup> More specifically, this graph was represented by the “Coulomb matrix” (CM), an atom by atom matrix with the nuclear Coulomb repulsion on off-diagonal elements and with approximate energy estimates of free atoms ( $E_I \approx 0.5Z_I^{4.225}$ ) as diagonal elements. Formal requirements such as uniqueness, translational and rotational invariance, as well as basic symmetry relations (symmetric atoms will share the same matrix elements in their respective rows or columns) are all met by the CM. Atom index invariance can be achieved through use of its eigenvalues (thereby sacrificing uniqueness<sup>214,226</sup>), sets of randomly permuted CMs,<sup>219</sup> or sorting by norms of rows,<sup>202</sup> thereby losing differentiability due to sudden switches in ranks.<sup>201</sup> We reiterate once more that the atom indexing dependence can be mitigated through using more sophisticated distance measures such as the Wasserstein metric.<sup>201</sup>

Similarly encouraging findings of KRR based QML models applicable throughout CCS were quickly reproduced for other materials classes such as polymers<sup>227</sup> or crystalline solids.<sup>228</sup> While off-diagonal elements with a London dispersion power law ( $r^{-6}$ ) have subsequently been found to be preferable for QML models of atomization energies,<sup>163</sup> other representations (vide infra) offer lower learning curve offsets. In particular, the

bag of bonds representation (BoB) is worthwhile mentioning.<sup>170</sup> Introduced in 2015, BoB groups the entries of the CM in separate sets for each combination of atomic element pairs within which all entries have been sorted. When calculating the similarity between two molecules, only Coulomb repulsions between atoms with the same nuclear charge are being compared, rendering thereby the similarity measurement more balanced and effectively lowering the learning curve offset. While even more compact than the CM, BoB lacks uniqueness due to being strictly a two-body representation which can not distinguish between homometric configurations.<sup>215</sup> The generalization of BoB toward the explicit incorporation of covalent bond information, angles, as well as dihedrals in terms of a systematic expansion in Bond, Angle, and higher-order interactions (i.e., BAML representation) was accomplished in 2016<sup>163</sup> by using functional forms and parameters from the universal force-field.<sup>229</sup> A similar, but more elaborate, parameter-free, many-body dispersion (MBD) based representation involving two and three body terms<sup>230</sup> was proposed later in 2018.

The CM has been essential as a baseline for the interpretation, analysis, and further development of subsequent QML models. It has also been adapted successfully to account for periodicity in the condensed phase, as evinced by learning curves for formation energies of solids.<sup>231</sup> For other properties, such as forces, electronic eigenvalues, or excited states, the CM (or its inverse distance analogues for QML applications with fixed chemical composition) is still competitive with state of the art representations.<sup>67,73–75,83,232–235</sup> Furthermore, because of its uniqueness, compactness, and obvious meaning, the CM (or its variants) are conveniently used to overcome frequent data analysis problems in atomistic simulations, such as removal of duplicates, quantification of noise levels, and simple learning tests.

Regarding the interatomic distance dependent decay rate of off-diagonal elements, it is also worthwhile to mention exponential functions, rather than  $1/r$ . In particular, the overlap matrix between atomic basis functions of all atoms has been proposed<sup>236</sup> and used with great success for QML models of basis-set effects<sup>237</sup> and excited-state surfaces.<sup>238</sup> The overlap matrix was also included within a recent sensitivity assessment of various state-of-the-art representations and performed in impressive ways.<sup>217</sup> A constant-size descriptor based on a combination of the CM with more common molecular graph fingerprints was also proposed in 2018.<sup>239</sup>

Viewing BoB and CM as first and second rank tensors, to the best of our knowledge, use of a third rank tensor (explicitly encoding the surface of all possible triangles in a compound) has not yet been tested.

#### 4.2. Continuous

Aforementioned discrete and global representations such as BoB enjoy fast computation. One important requirement for this kind of representation to work, however, is to introduce atom indexing invariance by sorting atoms according to the magnitude of entries belonging to each bond or other many-body types. This is artificial and may introduce derivative discontinuities with unfavorable consequences in related applications such as force predictions.

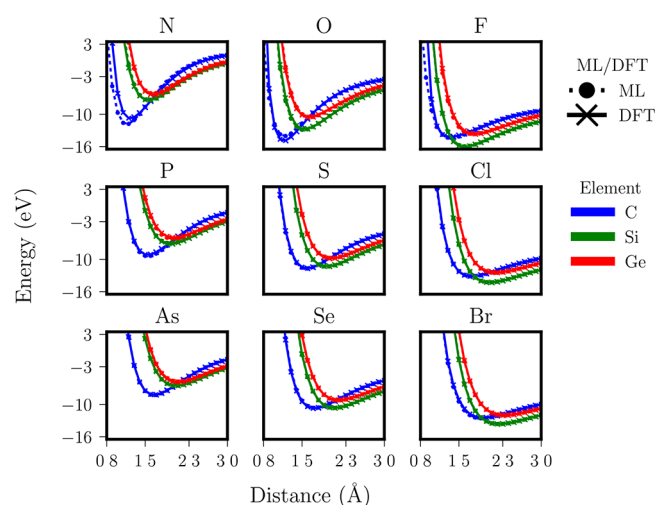
The sorting and associated problems can be naturally overcome by selecting continuous or distribution based representations, which, in essence, integrate out atom index dependent terms such as distance (w/wo angle and dihedral



angle) and/or nuclear charge (i.e., alchemically<sup>166</sup>) through use of smeared out projections (a Gaussian is commonly placed on each degree of freedom). Distribution based representations, also closely related to many-body or cluster expansion,<sup>240</sup> have gained much popularity also within QML models building on Behler's seminal work on atom-centered symmetry functions for training neural networks on potential energy surfaces,<sup>241</sup> or through the subsequent introduction of smooth overlap of atomic potentials (SOAP) for use in GPR by Bartok et al. in 2013.<sup>38</sup> The first variant of linearly independent distribution based representations for QML models, applicable throughout CCS, a Fourier series expansion of nuclear charge weighted radial distribution functions, was also contributed already in 2013,<sup>215</sup> albeit published in its final version only in 2015. Radial distributions were also used for representing crystals in solids in 2014.<sup>228</sup> The atomic spectrum of London Axilrod–Teller–Muto (aSLATM) terms was first presented in 2017 within the “AMON” approach by Huang et al.<sup>242</sup> (vide infra), yielding unprecedentedly low offsets in learning curves for atomization energies in the QM9 data set.<sup>171</sup> In that same year, SOAP based QML models were generalized and shown to be also applicable throughout CCS.<sup>243</sup>

The generic histogram of distances, angles, and dihedrals (HDAD),<sup>233</sup> a continuous but simplified version of BAML, both including many-body terms up to torsions, was contributed in 2017. In the following year, Faber et al. conceived the idea of adding alchemical degrees of freedom in a structural distribution based many-body representation, dubbed FCHL18<sup>166</sup> (FCHL indicating the first letters of the last names of the authors and 18 in the year 2018). The FCHL family of representations encodes a systematic interatomic many-body expansion in terms of Gaussians weighted by power laws due to the insights gained in ref 163. Power law exponents and Gaussian widths were optimized as hyper-parameters through nested cross-validation during training. FCHL18 consists of three parts: The one-body term corresponds to a two-dimensional Gaussian encoding the chemical identity of the atom in terms of groups and periods of the periodic table; the two-body term encodes the interatomic distance distribution scaled down by  $r^{-4}$ , and the three-body term encodes all angular distributions and is scaled down by  $r^{-2}$ . The impact of four-body terms has been tested on QM9 but was found to have negligible impact on learning curves.<sup>166</sup>

Most importantly within the context of CCS, FCHL18 based QML models have been demonstrated to be capable of accurately inferring property estimates of systems containing chemical elements which were *not* part of training. More specifically, consider the family of molecules of formula  $H_n Y \sim X$ , where  $Y$  corresponds to an element from group IV (either C, Si, or Ge), where “ $\sim$ ” represents single, double, or triple bond depending on chemical element  $X$  being from groups VII, VI, or V, respectively, and where  $n$  is the number of H atoms that saturates the total valences. Semiquantitative covalent bond potential binding curves have been predicted for any  $X/Y$ /bond-order combination using QML models after training on corresponding DFT curves for all other molecules that neither contain  $X$  nor  $Y$ . (see the top- and left-most subplot in Figure 4 for an illustration). For example, the ML binding curve of  $HC\#N$  was obtained after training on binding curves of all other molecules that neither contained N nor C, i.e., when predicting the blue curve in the upper left panel of Figure 4, the red and green curves of that panel were not part



**Figure 4.** QML models infer properties for new chemical compositions. DFT and QML (FCHL+KRR) based predictions of covalent triple, double, and single bonding between groups IV and V (left column), VI (mid column), and VII (right column) elements, respectively. Open valencies in the group IV elements have been saturated with hydrogens. QML models were trained on the DFT results for all of those chemical elements that are not present in the query molecule. Reproduced with permission from ref 166. Copyright 2018 licensed under a Creative Commons Attribution (CC BY) license.

of training nor were any other blue curve from the other panels. FCHL19, a recent revision, has been shown to provide a substantial speed-up in training and testing while imposing only a small reduction in predictive accuracy.<sup>208</sup>

We note in passing also the related moment tensor model (MTM) by Shapeev and co-workers, introduced in 2018,<sup>244</sup> as well as the unifying interpretation of many of the popular distribution based representations by Ceriotti and co-workers.<sup>245</sup>

## 5. REGRESSOR

Depending on how regression parameters are being obtained, the incorporation of legacy methods in QML models applicable throughout CCS is typically done either within neural networks or within Gaussian process regression (GPR) (or kernel ridge regression, KRR for short). Here, we mostly focus on kernel methods, mentioning only shortly the idea of transfer learning in neural network models,<sup>246</sup> which is also applicable to QML models as shown in 2018 by Smith et al.<sup>247</sup>

More specifically, five categories of QML models can easily be distinguished, each of which accounting for legacy information in its own way: QML models of parameters of existing models, QML models of corrections to existing models ( $\Delta$ -ML), multifidelity ML (MF-ML), multilevel-grid-combination (MLGC), and transfer learning techniques. We briefly review each of these in the following.

### 5.1. ML Models of Parameters

Existing force-field models can capture nicely the essential physics of a wide range of chemical systems, the main drawback being that force-field parameters (e.g., atom charges, harmonic force constants, etc.) are often rigid and unable to adapt to different atomic environments. Therefore, it would be natural to make these parameters flexible and predicted by ML models. This idea dates back to the 1990s, and the first piece of

related works was done by Hobday et al.,<sup>248</sup> where they proposed a neural network model to predict parameters of the Tersoff potential for C–H systems. In 2009, Handley and Popelier proposed to use machine learning models for multipole moments.<sup>49</sup> This idea was revisited in 2015, when learning curves for atomic QML models of electrostatic properties, such as atomic charges, dipole moments, or atomic polarizabilities were presented.<sup>249</sup> Their use for the construction of universal noncovalent potentials was established in 2018.<sup>70</sup> Neural-network based equilibrated atomic charges were also proposed in 2015 by Goedecker and co-workers<sup>54,250</sup> and in 2018 by Roitberg, Tretiak, Isayev, and co-workers.<sup>251,252</sup>

Similar strategy could also be applied to semiempirical quantum chemistry methods relying on parameters typically fitted by computational/experimental data. In 2015, QML models of nuclear screening parameters were contributed by Pavlo and co-workers.<sup>253</sup> In 2018, unsupervised learning for improved repulsion in tight-binding DFT was introduced by Elstner et al.,<sup>254</sup> followed by substantial further improvements in 2020.<sup>255</sup> Extended Hückel theory was revisited in 2019 by Tretiak and co-workers.<sup>256</sup>

### 5.2. $\Delta$ -ML

The idea to present QML models of label corrections applicable throughout CCS and which systematically improve with training data size was first established in 2015 in terms of  $\Delta$ -machine learning. Numerical results provided overwhelming evidence for the success of this idea as demonstrated for modeling energy and geometry differences between various levels of theory, including PM7, PBE, BLYP, B3LYP, PBE0, G2MP4, HF, MP2, CCSD, and CCSD(T) for QM9<sup>171</sup> and subsets thereof.<sup>257</sup>

$\Delta$ -ML also works for correcting complex and subtle properties, such as van der Waals interactions in extremely data-scarce limits, as illustrated for DFT corrections based on training sets with less than 100 training instances,<sup>258</sup> or to model higher-order corrections to alchemical perturbation density functional theory based estimates of heterogeneous catalyst activity.<sup>259</sup> Among many other applications,  $\Delta$ -ML has enabled corrections to electron densities,<sup>260</sup> electron correlation based on electronic structure representations within Hartree–Fock or MP2 level of theory,<sup>261</sup> or DFT and CCSD(T) based potential energy surface estimates.<sup>262</sup>

For noise-free data and functional QML models (unique representations), numerical results for learning curves indicate a constant lowering of offset, no matter which training set size. Such nonvanishing improvement appears to turn into vanishing improvement when employing  $\Delta$ -ML in order to correct low-quality or coarse-grained baselines, such as a semiempirical PM7<sup>257</sup> or Hammett's relation.<sup>212</sup>

### 5.3. Multifidelity

The success of  $\Delta$ -ML is encouraging, enabling a significant reduction in high-accuracy reference quantum chemical data necessary for training, to reach the same level of predictive accuracy as traditional QML models. However, it still consumes a considerable amount of data calculated at some high level of theory, as its structure in design is too simple to fully exploit the underlying correlation between varied quality of properties. In fact, well-established quantum chemical methods abound in literature, exploiting effectively the underlying correlation, in the name of the so-called composite methods, for example, the famous  $G_n$  series.<sup>263–265</sup> In essence, these methods approximate some specific part of correlation

energy (e.g., energy lowering due to inclusion of diffuse orbital in basis set) from a high level of theory (for instance CCSD(T)) by the same quantity calculated from a relatively low level of theory (say MP2). Because of error cancellation, composite methods have been proven to be extremely effective toward reaching an accuracy of experimental quality and are widely used for calculations of high-quality thermochemical data.<sup>263–265</sup>

To do interpolation and meanwhile exploit error cancellation effectively, multifidelity ML (MF-ML) comes into play. The core idea of MF-ML is hereafter demonstrated by total energy ( $E$ ) prediction. For brevity, we deal with two levels of theory (the low and high level are denoted by 0 and 1, respectively) and focus on one flavor of MF-ML, i.e., recursive KRR (r-KRR for short, or MF-KRR),<sup>266</sup> which is similar to its counterpart, recursive GPR (r-GPR, or MF-GPR),<sup>267,268</sup> and differs to MF-GPR to some extent, in analogy to the difference between KRR and GPR. Unlike  $\Delta$ -ML, MF-ML comprises multiple machines with different labels to learn (two for our exemplified case). The first one is just a traditional QML model trained on a set of data (denoted as  $S_0$ ) associated with the low level of theory, i.e.,  $E_j^0 = \sum_{i \in S_0} c_i^0 k_{(ij)}$ , where  $j \in S_0$  denotes the molecular representation vector and  $c^0$  is the regression coefficient associated with the low level of theory. This machine is also called the baseline model. Then we build a second machine with training set  $S_1$  satisfying  $S_1 \subset S_0$  and energy delta, i.e.,  $E^1 - E^0$ , as label, the same as a  $\Delta$ -ML model. In math,  $\Delta E_n^{0 \rightarrow 1} = E_n^1 - E_n^0 = \sum_{m \in S_1} c_m^1 k_{(m,n)}$ , where  $n \in S_1$ . Once trained separately for each machine, all  $c^l$ 's are obtained and MF-KRR predicts the property of any query  $q$  out-of-sample at the high level of theory by  $E_q^1 = E_q^0 + \Delta E_q^{0 \rightarrow 1}$ .

Extending r-KRR to more than two levels of theory is straightforward: except the baseline model for the lowest level, one needs to build one machine for every two adjacent levels of theory, and the final test energy is just the summation of the inferred energies by all machines, i.e.,  $E_q^L = E_q^0 + \sum_{l=0}^{L-1} \Delta E_q^{l \rightarrow l+1}$ , where  $l$  is the level indicator (starts from 0, the lowest level) and  $L$  corresponds to the largest  $l$ , or the target level. Bear in mind that  $S_0 \subset S_1 \subset \dots \subset S_L$ .

We note in passing that MF-GPR has a rather different formulation compared to MF-KRR and benefits from the stochastic nature of GP, i.e., it is capable of providing the variance estimate of prediction. Like GPR, data at each level of theory in MF-GPR is modeled as a GP,<sup>267,268</sup> and every two adjacent levels are connected by a linear transformation, i.e.,  $E^{l+1} = \gamma E^l + \epsilon$ , where  $\gamma$  is a scaling factor and  $\epsilon$  is a correction term respectively and both of which may depend on the two involved levels of theory (i.e.,  $l$  and  $l+1$ ). Nevertheless, both MF-KRR and MF-GPR could end up with the same set of working equations under certain conditions. For detailed derivation of the equations of MF-GPR, the reader is referred to an early review on QML.<sup>10</sup> Last but not least, one should note that MF-KRR converges toward the conventional KRR model associated with the highest level as the difference between training sets for each machine vanishes.

Albeit well-founded in mathematics decades ago, the power of MF-ML has not been harvested until recently. Applications include quantum collision for the Ar–C<sub>6</sub>H<sub>6</sub> system by Cui et al.,<sup>269</sup> bandgap prediction of solids done by Pilania et al.,<sup>270</sup> dopant formation energy prediction in hafnia by Batra et al.,<sup>271</sup> high-accuracy potential energy surface prediction for small molecules by Wiens et al.,<sup>272</sup> and the recently performed

molecular crystal structure prediction study by Egorova et al.<sup>273</sup>

#### 5.4. Multilevel Grid Combination

In spite of the drastic improvement over  $\Delta$ -ML, MF-ML has its own limitations. For one, the computational cost of the baseline evaluation for every query compound can still be considerable. Furthermore, it must be strictly satisfied that the increasingly more expensive training sets form a nested structure, implying that possible and beneficial correlations between non-nested reference data calculated at different level of theory are not being exploited. To overcome this drawback, Zaspel et al.<sup>266</sup> proposed a multilevel model in 2018, combining successfully ML with sparse grid (SG),<sup>274</sup> a numerical technique widely used to integrate/interpolate high dimensional functions.

The genuine SG approach assumes (quasi-)uniform grids along each dimension, which serves as basis functions (more precisely, centers of basis functions, such as triangular function) and based on tensor products of which any multidimensional function could be represented/expanded.<sup>274</sup> The expansion weight for each tensor product is dependent on only the indices of associated grid and spacing along each dimension and determined by multivariant Boolean algorithm.<sup>275</sup>

Replacing such a grid by abstract variable (or combinations of which) such as electron correlation level ( $x_C$ ), basis set ( $x_B$ ) and expressing system property as a function of these abstract variables represents an appealingly rigorous alternative. For example, the total energy of a system could be expressed as  $E = E(x_C, x_B)$ . Given some sparse grids comprising small  $x_C$  combined with all  $x_B$ 's, and small  $x_B$  combined with all  $x_C$ 's, and intermediate  $x_B$ 's combined with intermediate  $x_C$ 's, we are able to interpolate/extrapolate the  $E$  at some different combination of  $x_C$  and  $x_B$ . Of particular interest is extrapolation to regions unsampled, i.e., regions with large  $x_C$  and  $x_B$ . However, one major issue with such extension is the elusive nature of distance between two abstract variables, which is essential in determining the weight associated with each grid, as mentioned above. More specifically, it is unknown how to quantitatively characterize how distant HF and MP2 are along the dimension  $C$ , although qualitatively it is certain that HF lies closer to MP2 compared to CCSD(T). The  $B$  subspace, is understood much better, as the magnitude of  $x_B$  could be at least characterized by the largest angular channel, or more straightforwardly, although less rigorously, by the number of basis functions. This ill definition of these abstract variables is absent in genuine SG, as grids there reside typically in Euclidean space and therefore distance is well-defined. To rise to this problem, a workaround is to assume uniformity of grids along each dimension (i.e., equidistant) and grids along each dimension is represented simply by indices starting from 0 (now weights depend solely on the indices of grids). This, however, should always be done with great care. In the original MLGC paper,<sup>266</sup> electron correlation levels are reasonably chosen as HF, MP2, and CCSD(T), together with three basis sets, i.e., STO-3G, 6-31G, and cc-pVDZ (the number of basis functions increases by a factor of  $\sim 2$ ).

Note that the aforementioned SG approach deals with typically one system at a time. To incorporate it within ML framework, one extra variable has to be introduced, i.e., training set ( $x_N$ ), the size of which indicates the magnitude of  $x_N$ .<sup>266</sup> Accordingly,  $E = E(x_C, x_B, x_N)$ . Unlike subspace  $C$  or  $B$ ,

$x_N$  is well-defined with explicit value. Nevertheless, it has to be treated in a similar fashion as for  $x_C/x_B$ , i.e., given the minimal  $x_N$  (aka.  $N_0$ ) and a ratio ( $s$ ) between any two adjacent  $x_N$ 's, training sets are to be assigned an array of indices starting from 0 (for  $N_0$ ) and an increment of 1. This assignment is necessary so as to comply to what has been done for subspaces  $C$  and  $B$ . Now each grid in this abstract space is a combination of three variables:  $(x_C, x_B, x_N)$ , with  $x_I \in \{0, 1, \dots, I_{\max}\}$ ,  $I \in \{C, B, N\}$ . For each such combination, an associated ML model is trained (with  $N$  training instances of course). Given a query system, its energy is predicted as a weighted summation of test energies from all ML models, with weights derived from Boolean algorithm.<sup>266,275</sup> Note that in practice, to reduce the cost of generation of reference quantum data, a large  $x_N$  is associated with a low level of correlation and/or small basis set, while only few(er) labeled data are needed for high(er) correlation level and/or large(r) basis set.

With the above setting, Zaspel et al. were able to show<sup>266</sup> that MLGC enables  $\sim 10$ -fold reduction (cf. traditional single level ML model) in the costly highest level quantum data (i.e., CCSD(T)/cc-pVDZ) to reach chemical accuracy in predicting atomization energy of out-of-sample QM7b molecules. Last but not least, it is worth pointing out that MLGC reduces to MF-ML if only one dimension is being considered.

#### 5.5. Transfer Learning

While multilevel methods are most naturally combined with kernel methods, they may have even more far-reaching effects for neural network (NN) models, as training a NN model, deep NN (DNN) in particular, is a nontrivial problem. Furthermore, current ad hoc DNN models are typically specialized, meaning a (D)NN model may need to be retrained for a slightly different task.

Transfer learning (TL)<sup>276</sup> is one popular approach employing multiple levels in machine learning that can greatly alleviate the aforementioned problems, which reuses the knowledge gained through solving one task (base task) as a starting point for a second task (target task), different but highly related. For instance, knowledge obtained from learning to infer DFT energies could be applied to infer CCSD(T) energies.<sup>277</sup> A successful transfer of knowledge can improve the performance of the target DNN model significantly. Speaking the language of learning curve, TL could offer<sup>276</sup> (i) smaller offset as the transferred model per se provides a decent starting point and (ii) steeper learning curve due to the transferred model usually accounting for a parameter space rather close to the optimal one.

On the basis of the type of traditional ML algorithms involved, TL could be categorized into several variants. Here we focus on a variant named inductive transfer learning, in which the labeled source and target domains are the same, yet the source and target tasks are different from each other.

In TL, there are two essential ingredients: (i) a pretrained model, obtained by either training a network from scratch on some data set and a specific task, or simply from published models, and (ii) target network, to be trained on a target data set and task, but utilizing the learned features from (i). This process is likely to work only if the features are general (i.e., generic features) to both base and target tasks, as would be captured by the initial layers of NN models. When retraining the target model, one may choose to either freeze the initial layers in the base network to use them as feature extractors for the target model or fine-tune the last several layers further for



improved performance. A rule of thumb is to freeze when target labels are scarce (to avoid overfitting), while fine-tune otherwise.

To develop a successful TL model, it is vital to choose the proper base model and associated training data set. However, it remains largely an open question how to make the choice. This may require profuse intuition developed via experience. Furthermore, there exists one major potential risk of using TL, i.e., negative transfer, which refers to scenarios where the reuse of base-task knowledge degrades the overall performance of the target task. To avoid negative transfer, one may have to resort to approaches that explicitly model relationships between tasks and include this information in the transfer method.<sup>278</sup>

Applications of TL covers mainly computer sciences, such as image recognition and natural language processing. For chemistry-related problems, TL is emerging as a promising approach. Examples include Smith et al.'s work on predicting CCSD(T)/CBS energies based on transferred knowledge gained through training on DFT energies<sup>277</sup> for the ANI-1x data set (see section 8.1), Iovanac et al.'s work on property prediction of QM9 molecules,<sup>279</sup> as well as Cai et al.'s recent work on drug discovery.<sup>280</sup>

## 6. TRAINING SET SELECTION

Among all factors determining the performance of a QML model, training set selection plays another fundamentally important role in the sense that all knowledge essential for making confident predictions are implicitly encoded in the training data.

Several pertinent fundamental and distinct questions have remained open:

- Q1 How to extract the most representative and least redundant general subset from a given data set?
- Q2 How to quantitatively define the suitability of a given training set for a specific query at hand?
- Q3 How to systematically select the most relevant training set for a specific system?

Because of their highly nonlinear impact of training instances on model parameters, these questions are challenging and have not been studied much. Of course, the problem of training set selection is not a problem unique to chemistry, and it is relevant to most supervised learning problems in other fields. Currently, the aforementioned issues are mostly addressed through random selection. Although universally applicable, random selection suffer inevitably from selection bias inherent in the data itself. More specifically, in the randomly selected training set, many instances can be ignored and their inclusion in training does not improve predictive performance of the QML model (due to redundancy) or could even degrade it (due to being irrelevant for a given query test or due to noise).

Bias could become a very serious issue as the systems under study are increasingly more complicated. The origins of the bias issue could be divided into two components: (i) Curse of dimensionality. This is mainly related to the size of the systems and plagued further by compositional diversity. More specifically, as the system size and/or the encompassing number of types of elements grow, the size of the thus-spanned CCS grows combinatorially (see above). (ii) the inhomogeneity of CCS. The energetics in chemistry typically favors one kind of bonding over another. For instance, hydrogen atom

favors a single  $\sigma$  bond with other atoms, while carbon atom can exhibit several different bonding patterns such as  $sp^3$ ,  $sp^2$ , and  $sp$ . Consequently, random sampling will introduce more subsampling of hydrogen environments, but proportionally fewer C- $sp$  local environments leading to worse model estimates of properties for C than for H.

To tackle the bias issue, previous and ongoing research has been trying to almost exclusively tackle Q1, assuming a pre-existing data set (or a data set that is straightforward to generate, e.g., in molecular dynamics). Examples include the use of genetic algorithms (requiring labeled data to gradually expand the optimal training set),<sup>80,281</sup> or "active learning" approaches,<sup>79,222</sup> which selects the most representative subset "on-the-fly" from a given set of unlabeled configurations, i.e., no quantum chemical data is needed for making decisions about whether or not a query configuration is redundant. The AMONs concept proposed by the authors<sup>242</sup> partially resolves question Q3 (cf. Q1 and Q2), at the same time allowing for significant dimension reduction of CCS as well as the effective removal of statistical redundancy of training sets (see below for details). Other related work shifts the attention to training set reduction instead, primarily in molecular dynamics simulations, for instance, Li et al.<sup>62</sup> proposed a "learning-and-remembering" scheme, in which the decision to recompute QM data for a new configuration was taken every  $n$  steps. Another relevant contribution to active learning in CCS was made in 2018 by Smith et al.,<sup>282</sup> relying on "query by committee", i.e. ensemble information obtained through use of multiple neural networks (of the ANI kind<sup>53</sup>). Potentially promising alternative directions could possibly be inspired by recent developments in computer science, among many others notably the idea of artificial "soft" labels, curated through carefully blending features of training instances.<sup>283,284</sup> In the original paper, this idea was tested on the MNIST data set (a database of handwritten digits) and similar performance was achieved with much fewer but soft labels, as compared to training on almost the entire data set. This idea should in principle also be applicable to CCS requiring the design of some fictitious averaged training molecules, interestingly probable to violate common rules of chemical bonding. In the following, we review the three most promising approaches toward training set selection: genetic algorithm, active learning, and the AMONs approach.

### 6.1. Genetic Algorithm

Genetic algorithms (GA) have been widely used in (global) optimization problems in quantum chemistry, such as first-principles based global structure optimization<sup>285</sup> (for compounds with desired physio-chemical property), a key topic in the inverse-design problem.<sup>286</sup> To the best of our knowledge, the first piece of work about using GA for training set selection within QML was done by Browning et al.<sup>281</sup> for molecules, followed by Jacobsen's work<sup>80</sup> on SnO<sub>2</sub>(110) surface reconstruction.

In the following, we discuss the central idea of GA for the selection of the most representative set of QM9 molecules as done in ref 281. For applications to other properties and systems such as chemisorption systems,<sup>287</sup> only technical details will differ. Given a set ( $S_0$ ) of  $N$  molecules, GA carries out three consecutive steps for optimization: (a) Generate  $M$  random sets of size  $N_1$ . This forms a starting population of training sets (aka. the parent population), labeled as  $\hat{S}^{(1)} = \{S_i\}$ , where  $i \in \{1, 2, \dots, M\}$ . Note that the initial size needs to be

balanced against the diversity of the molecules for optimal performance. (b) Train a QML model on each set in  $\hat{S}^{(1)}$  and then test on some joint preselected out-of-sample molecules (i.e., not part of  $\hat{S}^{(1)}$ ), the resulting test error  $\epsilon_i$  (measured by for instance mean absolute error) serves as a “fitness” indicator, characterizing how fit  $s_i$  is as a training set (smaller  $\epsilon_i$  means better fitness). (c) Evolution of  $\hat{S}^{(1)}$  takes place through three consecutive steps: selection, crossover, and mutation. In the selection step, decisions have to be made on which  $S_i \in \hat{S}^{(1)}$  should be kept in the population to produce a temporary refined smaller set  $\hat{t}^{(1)}$ , and a set with larger fitness value means higher probability to be kept in  $\hat{t}^{(1)}$ . The crossover step involves the update on  $\hat{S}^{(1)}$  from  $\hat{t}^{(1)}$ , and the resulting new population is relabeled as  $\hat{S}^{(2)}$ , each of which is obtained by mixing molecules from two subsets of  $\hat{t}^{(1)}$ . The last step mutation randomly modifies molecules in some subset of  $\hat{S}^{(2)}$  to promote diversity, e.g., replace  $-\text{NH}_2$  group by  $-\text{CH}_3$ . To avoid introduction of chemical environments alien to the whole data set, the replacements in mutation have to be constrained locally in  $S_0$ . (d) Go back to step (b) and repeat b–d until there is no improvement in the population and the fitness value has no significant improvements for over  $n$  iterations. The final converged set corresponds to a “optimal” training set and is labeled as  $\hat{S}$ .

It is not a surprise that selected  $\hat{S}$  should be able to represent all the typical atomic environments in  $S_0$ , and therefore a QML model trained on  $\hat{S}$  warrants significantly improved test results in comparison to randomly drawn training sets. As the fitness value decreases during the GA iterations, the QML models “tried” out the sensitivity with respect to inclusion of certain training instances, and this can serve their systematic inclusion/exclusion. The usefulness of the optimized set  $\hat{S}$  has to be assessed by the generalizability of the QML model trained on  $\hat{S}$  to new molecules absent in  $S_0$ . Indeed, improved generalizability is observed for PubChem molecules compared to random sampling, as was reported in ref 281.

In spite of its power for solving hard optimization problems, such as finding the optimal training set composition, the drawback of most GA implementations is also obvious: It typically relies on the availability of labeled data to evaluate the fitness in each iteration. As such, it only offers computational cost savings in terms of QML model efficiency and not in terms of reducing the total need for available training data. Possible solution to circumvent this is to introduce heuristics in feature space, e.g., accounting for the fitness by some distance metric instead, meanwhile avoiding the costly training-test procedure in each iteration.<sup>288</sup>

## 6.2. Active Learning

Active learning (AL) is more interesting than GA for training set selection, as it can use directly unlabeled data, i.e., *before* the acquisition of costly labels. Intuitively, it makes sense that this should be possible as the quantum properties of any compound are implicit functions of its composition and geometry, which is the only input required for calculating rigorous representations. Among the many categories of AL algorithms used for determining which unlabeled data points should be labeled, below we mainly focus on the variance reduction query strategy, which labels only those points that would minimize output variance (uncertainty in prediction). Note that the task of variance estimation is fundamentally different from mean error estimation, and the variance based selection method differs significantly from the mean error

based selection method (such as GA mentioned above) accordingly. Relevant works on active learning include the  $D$ -optimality approach<sup>79,222</sup> and methods based on variance estimators using Gaussian process regression (GPR),<sup>289–291</sup> as well as neural network (NN) models.

Rooted in linear algebra, the  $D$ -optimality approach<sup>79,222</sup> takes advantage of (i) the dimension of features could in principle be significantly lower than the number of degrees of freedom spanned by the molecules (in particular for molecules that are in or close to their equilibrium states) and (ii) linearly parametrized local atomistic potential. Given a set of  $K$  molecules, the total energy of the  $q$ th molecule could be approximated as  $E^{(q)} = \sum_{i=1}^N V(\mathbf{x}_i^{(q)}) = \sum_{i=1}^N \sum_{j=1}^m \theta_j b_j(\mathbf{x}_i^{(q)}) = \sum_{j=1}^m \theta_j B_j(\mathbf{x}^{(q)})$  (in matrix form,  $E = \theta \mathbf{B}$ ), where  $B_j(\mathbf{x}^{(q)}) = \sum_{i=1}^N b_j(\mathbf{x}_i^{(q)})$  serves as the effective basis function of dimension  $m$  and  $b_j(\mathbf{x}_i^{(q)})$  is some function dependent only on the local representation  $\mathbf{x}_i^{(q)}$  of the  $i$ th atom in  $q$ ,  $N$  is the number of atoms of  $q$ . Then deriving the  $D$ -optimality criteria boils down to finding the “best” submatrix (of size  $m \times m$ ) from the overdetermined matrix  $\mathbf{A}^{K \times m}$  (where  $K > m$  and  $A_{ki} = B_k(\mathbf{x}_i^{(l)})$ ) such that the absolute value of  $\det \mathbf{A}$  reaches its maximum. Well-established algorithms exist to achieve the  $D$ -optimality criteria, e.g., the maxvol algorithm.<sup>292</sup> To obtain an optimal set, one typically has to iterate the procedure, one new query per time. If the corresponding magnitude of  $\det \mathbf{A}$  increases, it would be selected (query strategy) and discarded otherwise. Numerical results<sup>79</sup> have shown much improved performance for long-time MD simulation compared to classical on-the-fly learning.<sup>62</sup> However, the downside of  $D$ -optimality approach is also noticeable, that is, the model has to be updated at each iteration and application of the model could be prohibitive for a data set bearing a large feature space. Furthermore, the linear potential  $B_j$  depends on the proposed representation and the potential form, the latter of which in particular may suffer from lack of expressive power for some systems, i.e., the potential form may lack general applicability for a wide range of molecular systems. And last but not least, this approach relies on the choice of ratio (of  $\det \mathbf{A}$  values of two consecutive iterations) threshold manually chosen, which has to be tailored for a specific data set and may not be applicable to other data sets that only differ ever so slightly.

We note in passing that an alternative view<sup>79</sup> of  $D$ -optimality criteria is to assume that the energy has a Gaussian random noise and the best submatrix  $\mathbf{A}$  corresponds to the minimal variance in the solution of  $E = \theta \mathbf{B}$ . Besides, consideration of other properties such as forces could be naturally incorporated into this framework by simply taking derivatives of  $B_j$  with respect to Cartesian coordinates, expanding the feature matrix  $\mathbf{B}$ .<sup>79</sup>

Another variance-based approach relies on the GPR directly. That is, once trained, the model can estimate the variance directly, without referring to other criteria (as in the  $D$ -optimality approach). The estimated variance serves as a natural indicator, telling if any newly added data point would improve the model (if the variance is large with respect to a user-defined tolerance) or not (if the variance is very small). A small variance typically also means that the newly added data lie within or close to the current training space, distant otherwise. Methods like Gaussian process regression (GPR) are stochastic in nature and inherently capable of calculating the variance of prediction. More specifically, GPR aims to estimate the predictive distribution for any test data (unlike the kernel ridge regression model). Related works include that of

Snyder et al.'s,<sup>289</sup> in which the Bayesian predictive variance is shown to correlate with the actual error, and recently Reiher's group<sup>290,291</sup> used GPR to select optimal training sets in an automated fashion to explore chemical reaction network<sup>290</sup> and subsequently adjust for systematic errors in D3-type dispersion corrections, with one (sequential scheme) or multiple systems (batchwise variance-based sampling, BVS) selected each time.

Neural network (NN) based methods also offer a quite distinct perspective on the confidence of predictions. The general finding is that for NN methods estimates tend to be overconfident,<sup>293</sup> possibly due to the lack of principled uncertainty estimates<sup>294</sup> (i.e., NN model typically produces one single value for an input instead of a predictive distribution like GPR) and/or that the tools for mean estimation perhaps do not generalize.<sup>294</sup> In spite of the lack of native variance estimate, variance can still be modeled in practice through consideration of multiple parallel NN models. In analogy to GPR, uncertainty in NN models can be understood by taking a Bayesian view of the uncertainty of weight with some distribution assumed a priori and then updated by training data. There exists several variants of such NN models, including the ensemble neural network models,<sup>282,295–297</sup> where NN models share the same architecture but varied parameters (typically, ensembles are generated by NN submodels training on distinct subsets of data), and the dropout regularized neural network,<sup>298</sup> a lower cost framework for deriving uncertainty estimates (randomly dropout some nodes each time). These NN models are highly dependent on the training data, and therefore the predicted variance may not be reliable if the test data is distinct from training data, as is commonly expected for CCS exploration. Another type of NN model based uncertainty metrics, widely adopted, may alleviate this deficiency, which employs distances in feature space (or some latent space) of the test data point to the current training data to provide an estimate of similarity measure and thus model applicability.<sup>299</sup> This kind of approach enjoys several other advantages, such as easy interpretation, model independence, as well as potentially fast computation, but suffer from high dependence on the representation.<sup>299</sup>

### 6.3. AMON Based QML

Having a closer look at all the selection methods presented above, one notices that there is always some footprint of random sampling, i.e., one prerequisite for all those methods is a pre-existing starting training data, usually randomly selected, and reaching convergence of training data through iterative addition of new feature inputs may be slow if the starting points barely represent the space spanned by test data. The AMONs approach<sup>242</sup> attempts to mitigate these shortcomings through selection of the "optimal" training set on-the-fly, i.e., only after having been provided a given specific query test feature input. In essence, AMON based QML exploits the locality of an atom in molecule which allows to reconstruct extensive properties, such as the ground-state energy, in some analogy to the nearsightedness of electronic systems.<sup>300,301</sup> For the sake of a succinct discussion, we turn our attention to valence saturated system only and we neglect hydrogens. However, extension to other systems (e.g., system involving radicals, charges, conformational changes, vibrations, reactions, or noncovalent interactions) are also possible<sup>206,242</sup>). Note that throughout the whole process, we are only concerned with heavy atoms.

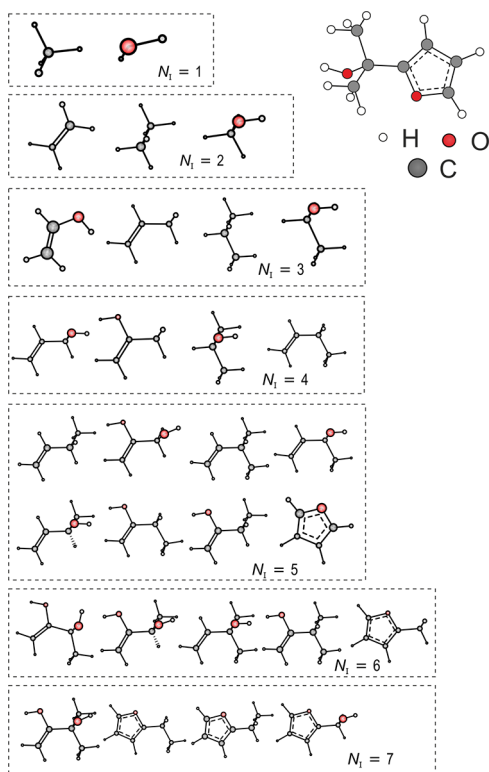
The AMONs selection procedure<sup>242</sup> can be divided into four major steps: (i) The connectivity graph  $G$  of a query molecule is constructed using its 3D geometry. (ii) Next, all subgraphs are enumerated (the  $i$ th subgraph is labeled as  $G_i$ ) of  $G$  with increasing number of heavy atoms (denoted as  $N_i$ ). For a given  $G_i$ , one performs a series of checks to see if it is a representative subgraph: (a) Is it a connected subgraph? (b) does subgraph isomorphism hold true? (c) Are all atoms valency-saturated after rationalization of the subgraph? And (d), is ring structure retained when all associated nodes are present in the current subgraph? If all of these criteria are met, then  $G_i$  is ready for further filtering and discarded otherwise. Criteria b is concise yet informative: subgraph isomorphism ensures that hybridization states of all atoms in the subgraph are retained, implying that bonds in query graph with bond order larger or equal to 2 are not allowed to break for fragmentation. (iii) Perform geometry relaxation for the corresponding fragment (now with valencies saturated with hydrogen atoms) using, for example, Universal Force Field (UFF)<sup>302</sup> or other force-field optimizer with dihedral angles fixed to match the local geometry of the query molecule (to avoid conformational changes in local environments). This step is followed by geometry relaxation using some quantum chemistry program. At this stage, it may happen that the subgraph candidate dissociates (turning into a disconnected graph) or is transformed into a molecule with different connectivity. In the former case, the fragment should be discarded, while in the latter case, the subgraph isomorphism has to be rechecked. (iv) One proceeds if the subgraph candidate has experienced no change in connectivity or if subgraph isomorphism is retained despite there being change in connectivity. The resulting fragment is selected for the AMON database.

As the number of atoms in the subgraph increases, one continues looping through  $\{G_i\}$  until the set has been exhausted. The resulting set of AMONs is considered the query-specific "optimal" set which is representative of all local chemistries in the query molecule.

Figure 5 shows all AMONs for an exemplified QM9 molecule named 2-(furan-2-yl)propan-2-ol with AMON size ( $N_i$ ) being at most 7 by applying the above algorithm. Not surprisingly, there exist only two molecules possessing  $N_i = 1$ , i.e.,  $\text{CH}_4$  and  $\text{H}_2\text{O}$ . For  $N_i = 2$ , a  $\text{C}=\text{C}$  double bond is allowed to be cleaved from the 5-membered ring, forming a valid AMON  $\text{H}_2\text{C}=\text{CH}_2$ , as the resulting AMON retains its original coordination number for C's, meanwhile keeping their valence saturated (i.e., meeting octet rule). While a fragment such as  $\text{H}_2\text{C}-\text{OH}$ , also extracted from the ring, is not a valid AMON as the valence of C atom is not saturated. Repeating similar arguments for increasingly larger  $N_i$ 's, we end up with only 30 AMONs, but which as a whole represent the complete set of local atomic environments present in the target and has the potential to extrapolate accurately the properties of the exemplified target QM9 molecule, as well as infinitely many other molecules that share the same set of AMONs after fragmentation.

AMON based QML models exhibit improved slopes and offsets in learning curves, as evinced for thousands of molecules after reaching respective training set sizes of only  $\sim 50$  on average. By contrast, 20 times larger training set sizes are required using random sampling.<sup>242</sup> One should note that graph based AMONs are not omnipotent. They are best suited for sampling chemical spaces of large systems. To extend





**Figure 5.** All AMONs sizes 1–7 for training system specific QML models of exemplary query molecule 2-(furan-2-yl)propan-2-ol (top right).

AMONs to also handle configurational spaces is possible in principle, but not trivial as it faces challenges similar to modeling large systems without explicit graphs, such as metals, metal surfaces, or molecular crystals or liquids.

## 7. PROPERTIES

As we focus on supervised learning throughout this text, properties (or labels) of molecules have to be always paired with some molecular representation. Starting from regression of experimental properties, e.g., atomization energies, dipole moment, boiling point, in the early practises of machine learning, by now the scope of properties has been expanded significantly.

Because of its determining role for stability and dynamics, energy is among the most important properties, and it is also the primary target property of most studies. As early as in 2011,<sup>303</sup> reorganization energies in a subspace of CCS consisting of polyaromatic hydrocarbons relevant to photovoltaic applications were already predicted using ML models. While the pioneering work on demonstrating the applicability of QML models for navigating CCS was published in 2012 for atomization energies only,<sup>195</sup> a multiproperty neural network was published shortly after,<sup>219</sup> covering not only atomization energies but also polarizabilities, molecular orbital eigenvalues, ionization potentials, electron affinities, and excited-states properties at various levels of theory. The correlations among these properties have confirmed some of the well-established physical principles as well as shown some interesting patterns. As illustrated in Figure 6, the ionization potential (IP) is well correlated with the HOMO energies, as expected from Koopman's theorem; the polarizability is linked to the stability, as often implied by the hard–soft acid–base

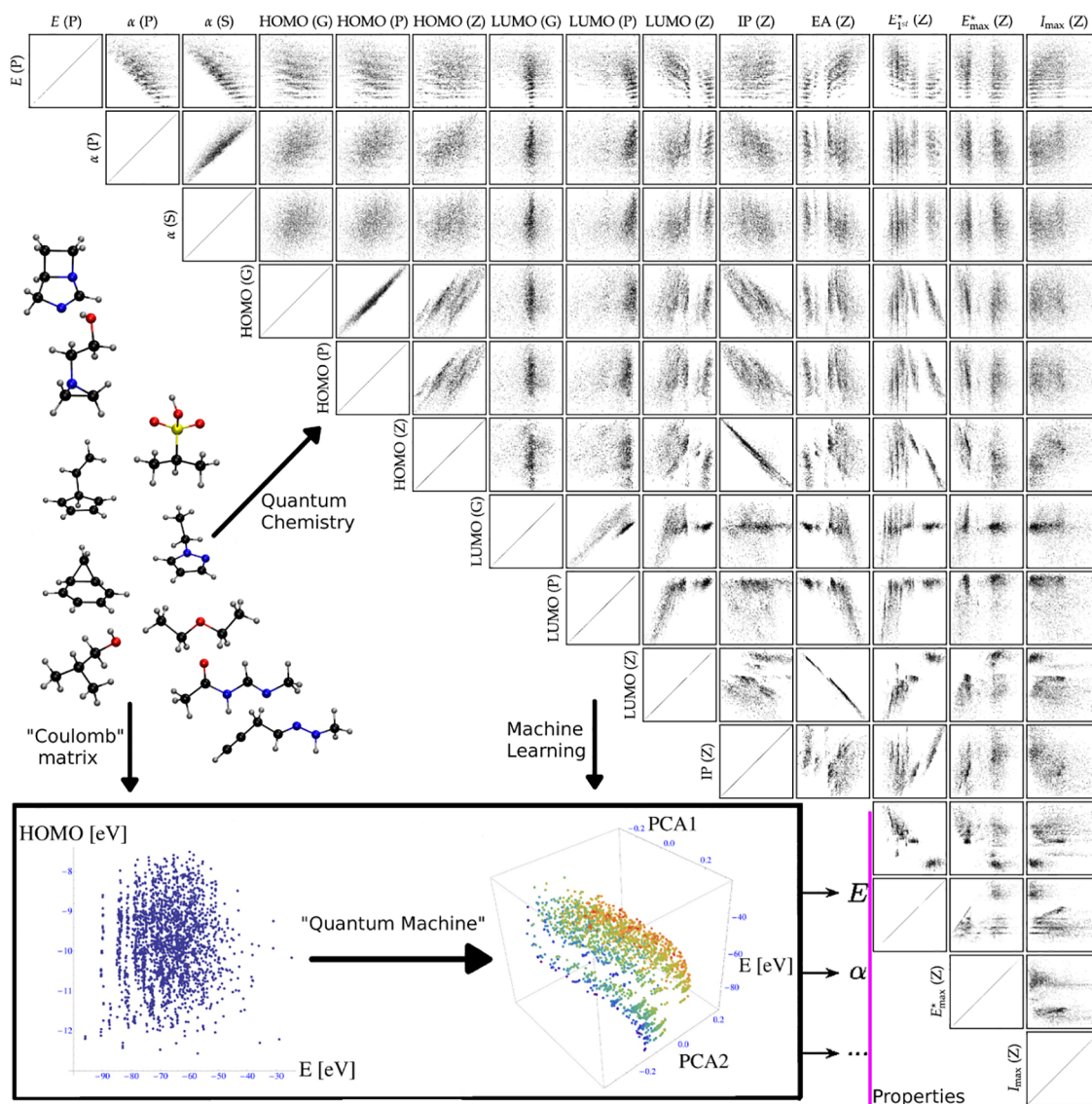
principle. Properties calculated at different levels of theory are strongly correlated, suggesting the possibility to exploit implicit correlations for the training of QML models with superior data efficiency. What is more interesting is that, for properties such as HOMO and atomization energies displaying little correlation, after training the neural network encodes some of the underlying and hidden correlations among these properties (box in Figure 6), indicating already in 2013 that neural network based QML models are amenable to “explainable AI”, as also illustrated subsequently in 2017 for effective atomic chemical potentials.<sup>304</sup>

While QML commonly deals with properties which correspond to observables, other well-defined but more arbitrary labels can also be modeled. Examples include atomic charges or energies which do not have a unique definition. A more exotic application consists of successfully trained QML models of “time-to-solution” in terms of estimates of the number of iterations necessary to reach convergence for given initial conditions: In 2020, QML models of the computational cost of common quantum chemistry calculations have been demonstrated to enable optimal load-balancing and scheduling in ensemble calculations of high-throughput compute campaigns through CCS.<sup>305</sup>

To provide a more comprehensive perspective on the interesting subject of property, below we divide all properties into three main categories depending on the number of atoms/species involved: atomic property (atom/bond/functional group in a molecule), molecular property (the entire molecule), and intermolecular property (at least two molecular species). And within each section, we briefly review models of important properties in a rough chronological order. Note that the boundary between different categories is not clear-cut. For instance, the highest vibrational frequency of a molecule may be attributed to certain functional group but only in an approximate way, the exact value of which may still depend on all the other atoms in the molecule. For this and similar cases, we prefer to classify the relevant properties into atomic rather than molecular properties.

### 7.1. Atomic

Generally speaking, atomic properties are relatively easy to learn as they typically benefit the most from the general assumption of locality of an atom in a molecule. On the basis of the QM9 database, QML models were introduced for atomic properties, such as core level excitations, forces (see previous section), or NMR-shielding constants.<sup>64</sup> Atomic QML models of electrostatic properties, such as atomic charges, dipole moments, or atomic polarizabilities, were introduced in 2015,<sup>249</sup> and their use for the construction of universal noncovalent potentials was established in 2018.<sup>70</sup> Deep neural networks for similar properties were also contributed in 2018 and 2019 by Unke and Meuwly.<sup>56,306</sup> Information from topological atoms has also been used to build dynamic electron correlation QML models in 2017.<sup>307</sup> In 2017 and 2018, atomic energies and potentials were also discussed in refs 166, 173, 308, and 309. QML models of polarizabilities based on tensorial learning were presented in 2020,<sup>310</sup> and most recently, Gastegger and co-workers introduced external field effects within neural networks and demonstrated interesting performance for predictions of IR, Raman, and NMR spectra, as well as for continuum solvent effects on chemical reactions.<sup>209</sup> Multiscale models of atomic properties have also been proposed.<sup>311</sup>



**Figure 6.** Property vs property matrix for  $\sim 7k$  organic molecules at various levels of theory. A multiproperty neural net trained in CCS encodes underlying correlations as evinced by the first principal components of the last layer for 2k molecules not part of training. Reproduced with permission from ref 219. Copyright 2013 licensed under a Creative Commons Attribution 3.0 license.

QML models of NMR shifts in molecules were first studied in 2015<sup>64</sup> and 2017,<sup>242</sup> followed by shifts in solids in 2018,<sup>312,313</sup> and NMR shifts in solvated proteins, coupling, a kaggle challenge, and an in-depth revisit of shifts in molecules were all contributed in 2020.<sup>314–317</sup>

In 2017, self-correcting KRR based models of potential energy surfaces and vibrational states were presented in ref 318 as well as neural network based molecular dynamics for the calculation of infrared spectra.<sup>319</sup> Out of all QML models for properties studied in the 2017 overview study on the CCS of QM9,<sup>233</sup> it was only for the highest vibrational (fundamental) frequency that random forests performed better than KRR or neural networks, the likely reason being that the QML model's task consisted "only" of detecting if an O–H or N–H bond present on top of the C–H bonds, and to assign the typical corresponding bond-frequency, and that typically random forests work well for such classification tasks. Other 2018 studies dealing with infrared spectra include refs 206, 251, and 252.

## 7.2. Molecular

At the molecular level, properties are greatly diversified, ranging from properties for ground state to excited ones, from static to dynamic ones, as well as from single molecule in vacuum to condensed phase.

QML models of electronic properties, such as excited states, quantum transport, or correlation, have remained rather sparse over the years. Examples include QML models for electron transmission coefficients for transport across molecular bridges of varying composition,<sup>320</sup> and Anderson impurity models<sup>321</sup> in 2014, and dynamical mean field theory<sup>322</sup> and excitation energies<sup>232</sup> in 2015. Only recently QML has been extended to also study nonadiabatic excited states dynamics for given systems (conformational sampling) by Dral, Barbatti, and Thiel<sup>323</sup> or Westermayr and Marquetand.<sup>83,324</sup> And the recent introduction of SchNarc,<sup>325</sup> a combination of the deep neural net architecture SchNet<sup>55</sup> and the surface hopping ab initio molecular dynamics code SHARC,<sup>326</sup> has led to promising first results for CCS studies involving small sets of small

**Table 1. Overview: Synthetic Quantum Data Sets in Three Data Families of Chemical Compound Space: Generated Data Base (GDB<sup>33,353,360</sup>), Transition Metal Complexes (TMC), and Periodic Systems (Crystalline Solids or Surfaces)<sup>a</sup>**

family	data set	composition	size	method	properties	year	notes
GDB	QM7 <sup>385</sup>	C, O, N, S	7165	PBE0	$E$	2012	
	QM7b <sup>358</sup>	C, O, N, S, Cl	7211	PBE0, ZINDO, GW	$E, \epsilon, \alpha, E^*$ , etc.	2013	
	QM9 <sup>171</sup>	C, O, N, F	134k	B3LYP/6-31G(2df,p)	$E, \mu, \alpha, \epsilon, P_{\text{thermo}}$ , etc.	2014	
	QM8 <sup>363</sup>	C, H, O, N, F	20k	TDDFT, CC2/def2-TZVP	$E^*, f_i, f_2$	2015	excited state
	ANI-1 <sup>366</sup>	C, O, N, F	20M	w97x/6-31G(D)	$E$	2017	off-equilibrium
	QM7bMl <sup>266</sup>	C, O, N, S, Cl	7211	{HF,MP2,CCSD(T)} / {sto-3g, 6-31g, cc-pVDZ}	$E$	2018	multifidelity QML
	Alchemy <sup>362</sup>	C, N, O, F, S, Cl	119k	B3LYP/6-31G(2df,p)	$E, \mu, \alpha, \epsilon, P_{\text{thermo}}$ , etc.	2019	
	QM7-X <sup>359</sup>	C, H, O, N, S, Cl	4.2M	PBE0+MBD	$E, f, \epsilon, \mu, \alpha, q_A, C_6$ , etc.	2020	off-equilibrium
	ANI-1x <sup>367</sup>	C, O, N, F	5M	w97x/def2-TZVPP and CCSD(T)/CBS	$E, f, \mu, q_A$ , etc.	2020	off-equilibrium
	AGZ7 <sup>365</sup>	B, C, N, O, F, Si, P, S, Cl, Br, Sn, I	140k	B3LYP/cc-pVTZ	$E, \mu, \alpha, \epsilon, P_{\text{thermo}}$ , etc.	2020	
TMC	tmQM <sup>382</sup>	3d, 4d and 5d transition metals, B, Si, N, P, As, O, S, Se, halogens	86k	TPSSH-D3BJ/def2-SVP	$E, \mu, q_A, \epsilon$ , etc.	2020	GFN2-xTB geometry
	(MIT) <sup>383,386</sup>	Cr, Fe, Mn, Co, Ni, C, N, O, S, Cl	>2M	B3LYP/LANL2DZ (6-31g*)	$E, \Delta E_{H-L}$ , redox potential	2017, 2020	
periodic	Materials Project <sup>165</sup>	across periodic table	>600k	PBE	$E$ , electronic and response properties	2011	
	AFlow <sup>387</sup>	across periodic table	3M	PBE	$E$ , electronic and response properties	2012	
	OQMD <sup>388</sup>	across periodic table	300k	PBE	$E$ , electronic and response properties	2013	
	OC20 <sup>389</sup>	across periodic table	>1M	RPBE	$E, E_{\text{ads}}$	2020	

<sup>a</sup>Properties covered include  $E$  (total energy (or atomization energy)),  $f$  (atomic forces),  $q_A$  (atomic charges),  $\mu$  (dipole moments),  $\alpha$  (polarizability),  $\epsilon$  (eigenvalues),  $E^*$  (excitation energy),  $f_i$ : oscillation strength for transition from ground state to the  $i$ th excited state ( $i = 1$  or  $2$ ),  $\Delta E_{H-L}$  (high- and low-spin energy difference),  $C_6$  (London dispersion coefficients),  $P_{\text{thermo}}$  (thermochemical properties such as internal energies, enthalpy, free energy, and heat capacity);  $E_{\text{ads}}$  (chemisorption energy).

molecules.<sup>327</sup> For more details we refer to the recently published reviews on this field.<sup>328–330</sup>

QML models of electron affinities and ionization potentials with deep neural networks have also recently been proposed.<sup>331</sup> Symmetry conserving neural networks for efficient calculations of electronic and vibrational spectra have been presented in 2020.<sup>332</sup>

### 7.3. Intermolecular

As the system becomes more complicated, the associated properties also tend to show more interesting, and sometimes surprising patterns. Hereafter, we will focus on energetic properties, unless otherwise stated. Depending on whether or not the system has experienced significant reconstruction in the relative orientation between atoms, intermolecular energetics could be further divided into intermolecular binding energy or reaction energy/barriers. Below, we summarize relevant contributions for each of these two subcategories.

In terms of binding energies within assemblies of atoms, ever since the publication of ref 195 in 2012, a large variety of systems has been addressed, reaching from formation energy predictions of diverse inorganic materials,<sup>164,166,333</sup> over models of chemical bonds in molecules,<sup>334</sup> to models of electronic properties of transition metal complexes.<sup>335</sup> GPR/KRR based QML models represent a unified approach, as demonstrated for applications to surface reconstructions, organic molecules, as well as protein ligands.<sup>243</sup> Symmetry adapted learning of tensorial properties was introduced in 2018,<sup>336</sup> as well as neural networks for atomic energies,<sup>309</sup> on-

the-fly learning for structural relaxation,<sup>80</sup> crystal graph convolution networks for materials properties,<sup>337</sup> solvation and acidity in complex mixtures,<sup>338</sup> and a machine learning based understanding of the chemical diversity in metal–organic-frameworks.<sup>339</sup> An extensive review of big data in metal–organic frameworks was also published in 2020.<sup>340</sup>

Accurate QML prediction of reaction related properties, the reaction barrier in particular, is a difficult task, as typically off-equilibrium configurations are involved, and the training space is undersampled.

The use of QML models to investigate properties relevant for catalysis represents another major domain of research. A GPR model was used in 2016 to estimate free energies of possible adsorbate coverage for surfaces in order to accelerate the construction of Pourbaix diagrams.<sup>341</sup> In 2017, Ulissi et al. introduced a neural network based exhaustive search enabling the identification of active site motifs for CO<sub>2</sub> reduction,<sup>342</sup> as well as a GPR based estimator of adsorption energies for identifying the most important reaction step.<sup>343</sup> QML models of reaction barriers of elementary reactions (using 236 dehydrogenation, 38 N<sub>2</sub> dissociation, and 41 O<sub>2</sub> dissociation examples) on surfaces were proposed by Singh et al. in 2019.<sup>344</sup> Quantum machine learning based design of homogeneous catalyst candidates was presented in 2018.<sup>345</sup> In 2020, QML models of competing reaction barriers and transition state geometries corresponding to S<sub>N</sub>2 and E2 reactions in the gas phase were successfully trained and applied throughout a CCS covering thousands of reactants,<sup>346</sup> relying on the QMrxn data set.<sup>347</sup>



That same year, Bligaard and co-workers employed active learning to identify stable iridium oxide polymorphs and study their usefulness for the acidic oxygen evolution reaction,<sup>348</sup> introduced a Bayesian framework for adsorption energies of bimetallic alloy catalyst candidates,<sup>349</sup> and proposed a bond information based GPR as a means to speed up structural relaxation across different types of atomic systems.<sup>350</sup> In 2020, neural networks have been proposed for the prediction of overpotentials relevant for heterogeneous catalyst candidates,<sup>351</sup> as well as a higher-order correction scheme in alchemical perturbation density functional theory applications to catalytic activity.<sup>259</sup> An overview on machine learning for computational heterogeneous catalysis was also contributed in 2019.<sup>352</sup>

## 8. DATA SETS

As implied already in previous sections, the availability of training sets is vital for any machine learning. Admittedly, it would be ideal to generate training set only when necessary, i.e., to minimize the number of QM computations throughout CCS, or for converging the sampling using molecular dynamics. However, for general applications of QML, a pre-existing data set is indispensable, for instance, to tackle the inverse design problem to identify some compound with unknown composition and exhibiting specified and desirable ground-state physicochemical properties. Currently, this is only feasible with a given labeled data set being as representative as possible for the local chemistries that we know to affect the properties of interest.

Alongside the increasing popularity of QML in chemistry and related sciences, many data sets have emerged in recent years. By now, there are a multitude, built for various purposes. Here we detail all those data sets we know of that encode quantum information throughout CCS, with a coarsened and incomplete overview given in Table 1.

### 8.1. GDB

The synthetic GDB (generated database) data sets created by Raymond and co-workers for the main purpose of exploring the CCS of organic drug-like molecules comprise the probably largest list of systematically generated molecular graphs (constitutional and compositional isomers only) of small to medium sized organic molecules of biochemical relevance.<sup>182,353–355</sup> To date, GDB17<sup>182,355</sup> represents the single largest set of molecules, which contains more than 166 billion molecules made up of H, C, N, O, S, and halogens (up to 17 non-hydrogen atoms), obeying certain chemical rules for stability and synthesizability. GDB17 has two main subsets, GDB11 (26M)<sup>353,356</sup> and GDB13 (970M),<sup>354</sup> together with a variety of smaller subsets featuring specificity of organic chemistry. Because of its systematic enumeration, interesting new structures have been identified and subsequently been synthesized, as exemplified by the synthesis of trinorbornane.<sup>357</sup>

Other than the implicit information that any compound listed corresponds to a stable constitutional isomer, the original GDB data sets are unlabeled in the sense that only molecular composition and connectivity information are detailed, without calculated quantum properties. The first extension of the GDB data set to also include quantum data, QM7<sup>195</sup> consists of 7165 ground-state geometries and energies of molecules with up to 23 atoms (with up to 7 heavy atoms C, N, O, or S) calculated at the PBE0 level. QM7 is also the first quantum

benchmark data set covering the organic subspace CCS for QML. Some extensions exist, such as QM7b,<sup>358</sup> QM7b multilevel data set<sup>266</sup> (QM7bMI for short), and QM7-X.<sup>359</sup>

QM7b<sup>358</sup> extends QM7 by including chlorine-containing molecules (expanding the set size to 7211), and reporting 13 additional calculated electronic properties (e.g., polarizability, HOMO/LUMO energies, excitation energies). QM7bMI<sup>266</sup> was designed for studying QML combinations with legacy quantum chemistry methods such as multilevel, multifidelity, or transfer learning. Starting from the original coordinates at PBE level, geometries of QM7b molecules were refined at the level of B3LYP/6-31G(D), and subsequently single-point energies were calculated at nine levels of theory, corresponding to all possible combinations of electron correlation treatment {HF, MP2, CCSD(T)} and basis sets {STO-3G, 6-31G, cc-pVDZ}. QM7-X, the largest extension of QM7, is a comprehensive data set comprising ~4.2 M equilibrium and nonequilibrium structures of QM7b molecules, accompanied by 42 physicochemical properties computed at the PBE0+MBD level, covering global (molecular) and local (atom-in-a-molecule) properties ranging from ground-state quantities (such as atomization energies and dipole moments) to response quantities (such as polarizability tensors and dispersion coefficients).

Because of the limited molecular size, QM7 and its variants are scarcely scattered across CCS and barely begin to represent its full diversity and complexity. Targeting “big data” Ramakrishnan et al. released the QM9<sup>171</sup> data set in 2014, derived from molecular graphs drawn from GDB17,<sup>360</sup> totalling ~134k organic molecules made up of C, H, O, N, or F, and up to nine non-hydrogen atoms. Except for equilibrium geometries and electronic ground-state properties, QM9 also records a series of thermochemical properties at 298 K and 1 atm pressure estimated based on harmonic frequencies, namely enthalpies, and free energies of atomization at the level of B3LYP/6-31G(2df,p). Alongside, additional QM data is reported for the subset of all of QM9's 6k constitutional isomers with sum formula C<sub>7</sub>H<sub>10</sub>O<sub>2</sub>, i.e., thermochemical properties computed at the G4MP2 level. In 2020, QM9 was augmented by more accurate energies, calculated at multiple levels of theory, including M06-2X, wb97xd, and G4MP2.<sup>361</sup> Another similar data set, dubbed alchemy<sup>362</sup> (sized 119 487) expands the volume and diversity of QMx series and is made up of 9–14 C, N, O, F, S and Cl atoms, sampled from the GDB MedChem subset of GDB17.<sup>355</sup>

The only data set that deals with excited-state properties across CCS is QM8,<sup>363</sup> totalling ~20k structures subsampled from QM9 and comprising up to eight heavy atoms C, O, N, or F. Ground-state energies ( $S_0$ ) and the lowest two vertical electronic singlet–singlet excitation energies ( $S_1$  and  $S_2$ ) are included, calculated at two TDDFT levels employing the density functional theory/basis-set combination PBE0/def2-SVP or CAM-B3LYP/def2-TZVP, as well as post-Hartree–Fock level CC2/def2-TZVP. Corresponding oscillator strengths ( $f_i$ ) for each transition from  $S_0$  to  $S_1$  have also been recorded.

As also evinced for GDB17, when increasing the number of atoms per molecule, the data set quickly grows out of control, and it becomes prohibitive to conduct QM calculations for comprehensive subsets of CCS. The Amon based dictionary of building blocks designed to cover GDB<sup>360</sup> and Zinc<sup>364</sup> and containing no more than seven heavy atoms (AGZ7) has been introduced to alleviate this curse of dimensionality.<sup>365</sup> It was

obtained by systematically fragmenting all larger molecules (from GDB17 and zinc<sup>364</sup>) into smaller entities containing no more than seven non-hydrogen atoms (i.e., atom-in-molecule based fragments, aka, AMONs<sup>242</sup>). To date, AGZ7 is the most compact yet most diverse data set relevant for organic/biochemistry, totalling only 140k molecules but covering up to 13 elements (H, B, C, N, O, F, Si, P, S, Cl, Br, Sn, and I). It also includes a similar set of properties as in QM9 but relying on a slightly different level of theory (B3LYP/cc-pVTZ as well as pseudopotentials for Sn and I).

Apart from QM7-X,<sup>359</sup> all data sets mentioned so far deal with equilibrium geometries only, representing the typical constraint for what defines a stable molecule. To enable the QML based study of dynamics and reactivity of non-equilibrium geometries throughout CCS, however, configurational sampling involving nonstationary geometries has to also be accounted for through the data sets. Similar to QM7-X, ANI-1<sup>366</sup> also explores nonequilibrium geometries but for relatively larger systems drawn from GDB11.<sup>353,356</sup> It consists of more than 20 M off-equilibrium structures (sampling both chemical and conformational degrees of freedom) and wB97x/6-31G(d) energies for 57 462 small organic molecules containing up to 11 CONF atoms. Two follow-up data sets expand ANI-1 considerably, i.e., ANI-1x and ANI-1cc.<sup>367</sup> The former contains multiple QM properties (density-derived properties and forces) from 5 M DFT calculations (wB97x/6-31G\* and wB97x/def2-TZVPP), while the latter contains 500k CCSD(T) energies for estimated CBS limits.

For MD simulations, two main data sets are being frequently benchmarked. One is ISO17,<sup>55,308</sup> containing MD trajectories of 129 molecules randomly drawn from the aforementioned 6k C<sub>7</sub>O<sub>2</sub>H<sub>10</sub> isomers, each comprising 5000 conformational geometries with total energies and atomic forces calculated at PBE level plus van der Waals correction.<sup>368</sup> The other is MD-17,<sup>73,369</sup> which records energies and forces from ab initio molecular dynamics trajectories (133k to 993k frames) at the DFT/PBE+vdW-TS level of theory at 500 K for eight organic molecules: benzene, uracil, naphthalene, aspirin, salicylic acid, malonaldehyde, ethanol, and toluene. More accurate CCSD(T) energies and forces are also available but only for ethanol (with basis cc-pVTZ), toluene and malonaldehyde (cc-pVDZ), and CCSD/cc-pVDZ for aspirin. Recently, a revised MD-17 data set was published,<sup>210</sup> with a lower noise floor in DFT forces thanks to tighter SCF convergence criteria and denser integration grids. In 2020, G4MP2 benchmarks of organic molecules with up to 14 non-hydrogen atoms were contributed by Dandu et al.,<sup>370</sup> and resulting QML models were compared and discussed.

## 8.2. PubChem and ZINC

While the GDB family currently dominates QML campaigns, GDB compounds resulted from virtual exhaustive graph enumeration campaigns and mostly correspond to molecules for which neither thermodynamics stability nor synthesizability has been established. Within practical applications, such aspects matter for the experimental design and fabrication of new chemical compounds. With respect to QML, some theoretically possible local chemical environments may not be viable within the entire molecular framework, and ruling out such possibilities when training could help to further improve data efficiency and transferability. PubChem<sup>371</sup> is an ever-growing open chemistry database hosted at the National Institutes of Health (NIH). As of October 2020, there were

over 111 million unique chemical structures records listed together with many a experimental property, as contributed by hundreds of data sources. To harvest the richness and popularity of this database, Maho Nakata, and co-workers launched the so-called PubChemQC project,<sup>372</sup> consisting of ground-state geometries and properties (at B3LYP/6-31G\* level), as well as low-lying excited states of approximately four million molecules via time-dependent DFT at the level of B3LYP/6-31+G\*. A PubChemQC derived subset, called pc9,<sup>373</sup> covering over 99k molecules made up of CHONF was published afterward and encoded the same set of properties as QM9. The full potential of PubChemQC remains yet to be generally explored.

ZINC,<sup>364</sup> yet another large database, focuses more on biochemistry, in particular drug design. Quantum calculations on this database per se have not taken place, except for its associated fragment set. That is, AZ7, a subset of AGZ7,<sup>365</sup> contains all ZINC AMONs of up to seven non-hydrogen atoms (with optimized geometries and electronic properties, as for AGZ7 described above). AGZ7 could be considered as an effective set covering all local chemistries of ZINC and may serve as a scaffold for building larger drugs through a theoretical approach.

Beside PubChem and ZINC, there are several other public big databases being exploited within QML. One of them is the Cambridge Structural Database (CSD),<sup>374</sup> on the basis of which Stuke et al.<sup>375</sup> reported a diverse benchmark spectroscopy data set of 61 489 molecules, denoted OE62. Using geometries optimized by PBE plus vdW correction, OE62 provides total energies and orbital eigenvalues at PBE and PBE0 levels for all molecules in vacuum and at the PBE0 level for a subset of 30 876 molecules in (implicit) water. Also based on CSD, Schober et al.<sup>376</sup> extracted 95 445 molecular crystals thereof and carried out computations on electronic couplings (at the level BLYP and fragment molecular orbital-based DFT) and intramolecular reorganization energies (by QM/MM with an ONIOM-scheme) as two main descriptors for charge mobility, hoping to facilitate the theoretical design and discovery of high mobility organic semiconductors.

## 8.3. Barriers and Spin

Quantum data sets on chemical reaction profiles are rather scarce. The QMrxn<sup>347</sup> reports calculated quantum properties for S<sub>N</sub>2 and E2 reactions amounting to 4466 transition state and 143 200 reactant complex geometries and energies at MP2/6-311G(d) and single-point DF-LCCSD/cc-pVTZ level of theory, respectively. QMrxn covers the subset of CCS that is spanned by the substituents -NO<sub>2</sub>, -CN, -CH<sub>3</sub>, -NH<sub>2</sub>, and with -H, -F, -Cl, and -Br as nucleophiles and leaving groups. A different data set featuring elementary reactions comes from Grambow et al.,<sup>377</sup> totalling 12k organic reactions that involve H, C, N, and O atoms, calculated at the wB97X-D3/def2-TZVP level, with optimized geometries and thermochemical properties for reactants, products, and transition states.

Going beyond mostly singlet-state chemistry, Schwilk et al. introduced QMspin,<sup>378</sup> consisting of ~5k (~8k) singlet (triplet) state carbenes derived from 4k randomly selected QM9 molecules. QMspin also contains optimized geometries (B3LYP/def2-TZVP for triplet state and CASSCF(2e,2o)/cc-pVDZ-F12 for singlet state), as well as the singlet-triplet vertical spin gap computed at MRCISD+Q-F12/cc-pVDZ-F12 level of theory.

For the QML models of the computational cost of typical quantum chemistry computations (measured by the CPU wall time), Heinen et al. reported the QMt data set,<sup>379</sup> consisting of timings of various tasks (single point energy, geometry optimization, and transition state search) for thousands of QM9 molecules at several levels of theory including B3LYP/def2-TZVP, MP2/6-311G(d), LCCSD(T)/VTZ-F12, CASSCF/VDZ-F12, and MRCISD+Q-F12/VDZ-F12.

Treating noncovalent interaction (NCI) within QML is an interesting and important research subject, with relevant large data sets emerging only as of recently. Most notably, several collections of NCI data sets have by now become publicly available,<sup>380</sup> covering 3700 distinct types of interacting molecule pairs: (i) DES370 K, contains interaction energies for more than 370k dimer geometries with NCI energy calculated at the level of CCSD(T)/CBS (MP2(aVTZ, aVQZ) correlation energy is used for extrapolation, and (ii) DESSM, comprising NCI energies calculated using SNS-MP2, for nearly 5 M dimer geometries. The monomers involved include typical organic species, made up of common p-block elements as well as alkali metal ions, most of which containing no more than seven heavy atoms.

Data sets including artificial molecules which violate basic principles of chemical bonding may also of great interest for QML, i.e., they may serve the use of “soft” labels, where relatively few compounds might more effectively represent CCS than selected many. MB08-165,<sup>381</sup> proposed by Grimme, exemplifies that idea, relying on systematic constraints rather than uncontrolled chemical biases. Originally, this data set was designed for benchmarking DFT methods. The potential of such “unbiased” artificial molecules as soft labels (training set) in QML has yet to be unraveled.

#### 8.4. Transition Metals

Transition metal complexes (d-block atom/ion center plus ligands, TMC for short) are pervasive in chemistry and have been widely used and studied. Because of their complicated electronic structure and the resulting higher computational cost (in comparison to typical organic molecules), the effective exploration of the chemical space spanned by TMCs remains a challenge and current efforts into this subspace are constrained to relatively low level of theory, primarily DFTB or DFT method with some small basis. Examples include tmQM<sup>382</sup> and the TMC data sets<sup>288,335,383</sup> from Kulik and co-workers, as described below.

tmQM<sup>382</sup> contains geometries and common electronic properties (as for QM9) of 86 665 mononuclear complexes extracted from the Cambridge Structural Database (CSD). tmQM includes Werner, bioinorganic and organometallic complexes based on a large variety of organic ligands and 30 transition metals. On the basis of the DFTB(GFN2-xTB) geometry, common quantum electronic properties (orbital energies, dipole moment and atomic charges) were computed at the TPSSh-D3BJ/def2-SVP level.

The largest and most comprehensive TMC data sets are from Kulik's group at MIT and have been contributed across multiple publications.<sup>288,335,383</sup> Overall, they correspond to combinations of several metal centers (Cr, Mn, Fe, or Co, Ni) and a wide range of ligands, ranging from weak-field chloride (Cl<sup>-</sup>) to strong-field carbonyl (CO) along with representative intermediate-field ligands and connecting atoms, including S (SCN<sup>-</sup>), N (e.g., NH<sub>3</sub>), and O (e.g., acetylacetonate). Calculated properties are primarily energetic, including total

energy, high and low spin-state energy difference ( $\Delta E_{H-L}$ ), and redox potential and solubility in candidate M(II)/M(III) redox couples, at the level of theory B3LYP/LANL2DZ (6-31G\* for ligands) with or without polarizable continuum model (PCM) for solvents. The total size could reach up to several millions.

Recently introduced metal–organic frameworks (MOF) data set by Rosen and co-workers,<sup>384</sup> called Quantum MOF (QMOF), represent another broad category of metal complexes. QMOF consists of computed properties (energy, band gap, charge density, and density of states) at the PBE-D3(BJ) level of theory, for more than 14 000 experimentally synthesized MOFs, which are made up chemical elements that span nearly the entire periodic table.

#### 8.5. Solid and Solid Surface

Compared to TMCs, solid and solid surfaces present a challenge on their own due to the diversity in composition and spatial arrangements, as well as the resulting complexity of electronic structure. Typically DFT based methods are used for generating large-scale (or high-throughput) data sets for these systems. The most frequently used method is GGA (PBE) or GGA+*U* with PAW (projected augmented wave) potentials. On the basis of relaxed geometry, associated calculated properties fall into either electronic properties, e.g., cohesive energy, band structure (and derived properties including density of states and band gap), or response properties such as elastic tensor, bulk modulus, and thermodynamic properties (vibrational spectra, free energy, specific heat, and entropy) within harmonic approximations.

Relevant well-known solid databases and compute platforms include (i) AFlow,<sup>387</sup> an open data set of more than 3 M material compounds (including alloys, intermetallics, and inorganic compounds) with over 596 M calculated properties. (ii) The Open Quantum Materials database<sup>388</sup> (OQMD), a high-throughput database currently consisting of nearly 300k total energy calculations of compounds from the Inorganic Crystal Structure Database (ICSD). (iii) The Materials Project<sup>165</sup> ([www.materialsproject.org](http://www.materialsproject.org)) covers the properties of almost all known inorganic materials, currently containing over 131k inorganic compounds and more than 530k nonporous materials. (iv) The Materials Cloud ([www.materialscloud.org](http://www.materialscloud.org)),<sup>390</sup> a platform designed to enable open and seamless sharing of resources for computational science, driven by applications in materials modeling. (v) The Novel Materials Discovery (NoMaD, <http://nomad-repository.eu>), led by Scheffler, Draxl et al. (vi) The Open Materials Database (<http://openmaterialsdb.se>, currently under development) spearheaded by Armiento. The latter both are public archives for hosting, sharing, and reusing material data in their raw form. Apart from comprehensive public repositories for solid data sets, there are also select contributions for select materials classes, including the aforementioned data set of ~10k AB<sub>2</sub>C<sub>2</sub> elpasolites covering all main-group elements up to Bi from Faber et al.<sup>164</sup>

Regarding solid surfaces, the new Open Catalyst Project<sup>389</sup> aims to help discover and design new catalysts for renewable energy storage using ML (<https://opencatalystproject.org>), currently including mainly the OC20 data set,<sup>389</sup> consisting of >1 M relaxations (over 26 M single point evaluations) at RPBE level for a wide range of adsorbates (C-, N- and O-containing species) and surfaces.



## 9. SOFTWARE PACKAGES

To perform and supplement the aforementioned studies with methods and data sets, numerous software packages have been developed over recent years. We briefly mention the available codes and categorize them into three main types, the first of which being those related to the acceleration of legacy quantum codes, such as ab initio molecular dynamics (MD) runs in VASP,<sup>391</sup> Gaussian process based geometry optimization in ASE,<sup>392</sup> machine learning adaptive basis sets within CP2K,<sup>237</sup> as well as SNAP<sup>393</sup> in LAMMPS, a machine-learning interatomic potential using bispectrum components to characterize the local neighborhood of each atom of the system.

Codes which fall into the second category are standalone packages, some of which having also been interfaced to other atomistic simulation software. QMLcode<sup>208,394</sup> which is an open-source python-based package featuring the Coulomb-matrix,<sup>195</sup> BoB,<sup>170</sup> SLATM/aSLATM,<sup>242</sup> FCHL18, and FCHL19<sup>166,208</sup> and other representations. AQML code<sup>395</sup> is a variant of QMLcode featuring the BAML<sup>163</sup> representation, and on-the-fly selection of AMONs for training.<sup>242</sup> PLUMED<sup>396</sup> is an open-source, community-developed library that provides a wide range of methods including enhanced-sampling algorithms, free-energy methods, and MD data analysis capabilities. It also interfaces with some of the most popular MD engines. TensorMol<sup>397</sup> is a package of neural networks for chemistry, capable of running many common tasks in quantum chemistry such as geometry optimizations, molecular dynamics, Monte Carlo, nudged elastic band calculations, etc. It can also take into account screened long-range electrostatic and van der Waals interactions. TorchANI<sup>398</sup> is a PyTorch implementation of ANI. It can compute molecular energies, gradients, Hessian and derived properties from the 3D coordinates of molecules. It also include tools to work with ANI data sets (e.g., ANI-1, ANI-1x, etc.).

The third category of software packages deals predominantly with data set construction, management and analysis. In particular, specific platforms include AFlow<sup>387</sup> which has been mentioned above in section 8.5, and AiiDA,<sup>399</sup> an open-source infrastructure for automation, management, sharing and reproduction of the workflows associated with big data in computational sciences.

## 10. COMPOUND DISCOVERY

The computational design and discovery of new compounds can be generally conducted following either one of two distinct approaches. The Edisonian and more basic one is straightforward, within a brute-force high-throughput screening, through solving Schrödinger equations sequentially or in parallel for potential materials candidates one by one, followed by subsequent ranking and selection. Given sufficient coverage and having used the data for training, the ab initio solver could successively be replaced by QML models, capable of making faster and equally accurate predictions of target properties of interest. It is obvious that such an approach suffers from limited domains of compounds conceived in the first place, no matter what solver is used for computation of properties. Also, as the intended search domain expands in CCS, the number of possible potential candidates will grow combinatorially. Therefore, when adopting this strategy, one should refrain from generally expanding the search domain and rather focus on a constrained subdomain of compounds, sharing one or

more common features, e.g., the same stoichiometry and space group, as was exemplified for the elpasolite family  $ABC_2D_6$  by Faber et al.,<sup>164</sup> where compounds with exotic atomic oxidation states were identified.

The second more sophisticated approach attempts to solve the problem in an inverse fashion; more specifically, given a specific (range of) value(s) for the target property, how to best locate the corresponding optimal (set of) compound(s) from CCS. One particularly promising variant is the gradient-based inverse design,<sup>14</sup> which can be reformulated as a global optimization problem and has the potential to search chemical subspace for substantial domains, due to its analytical nature. Strictly speaking, almost all current ML-guided studies (mostly neural network based) on gradient-based inverse design (e.g., ref 187, for a review, see ref 286) fall into the QSPR regime, as the input is seldomly 3D geometry, but rather SMILES or other molecular graph derived features (therefore, the mapping from representation to property is not unique). This strategy is however the only attainable way by now, as otherwise (i) the search subspace (when optimizing for the “optimal” compound) would become overwhelmingly large due to the explosion of conformational degrees of freedom (Levinthal’s paradox). (ii) There exists, to the best of our knowledge, no 3D geometry-based representation that is compact enough for decoding, i.e., restoring the original geometry from its representation vector/matrix/tensor (or simply  $\mathbf{x}$ ), even with the help of a neural network model like variational encoder (VAE), as the entries in  $\mathbf{x}$  are highly intertwined (significantly more so than the SMILES string). The fact that many representations are still being haunted by the uniqueness issue, further plagues these efforts, as often only two- and three-body terms are included in distribution-based representation. While inclusion of four-body terms are mandatory for reconstructing geometry, as evinced by the Z-matrix representation of geometry, the resulting  $\mathbf{x}$  would become very expensive for generation, and more importantly, this could further perplex the feature vector decoder. However, representing a molecule in its most native form in terms of nuclear charges and coordinates,  $\{Z; R\}$ , i.e., by the variables employed in the electronic Hamiltonian, or some transformed form, such as an external potential, one is free from such problems. This strategy would be consistent with the aforementioned GCE and LCAP approach detailed in section 1.3.

## 11. OUTLOOK AND CONCLUSION

While QML is still in its infancy, very encouraging progress has already been achieved. It is still a long way, however, before we will reach the goal of routinely designing and discovering novel molecules and materials on a computer. Some of the most fundamental problems, also among the most common tasks in quantum chemistry calculations, such as correctly predicting ground-state energy and forces of novel molecules or materials with high efficiency and accuracy, still remain unresolved at large. Such seemingly simple tasks are particularly challenging when it comes to systems that are highly distorted, charged, or multireference in nature or that involve long-range nonbonded interactions. Successful QML models could easily demonstrate their applicability by energy ranking of competing structures of real materials. We believe that such tasks will be crucial for subsequent more challenging QML applications.

Another interesting path to pursue might be the integration of alchemical perturbation theory into QML. Because the alchemical problem could be essentially reformulated as a ML

problem that involves both energy and energy gradient with respect to nuclear charges. A corresponding extension would exploit similarities between alchemical interpolations in pseudopotential parameter space and compositional representations that explicitly account for group and period in the periodic table, on top of all the structural degrees of freedom. Within the FCHL representation,<sup>166</sup> preliminary results for inferring properties of chemical elements absent in training have already been obtained (see Figure 4).

Besides the curse of dimensionality, imposed by the compositional, constitutional, and conformational diversity of CCS, the lack of a more theoretical underpinning of the genesis of data sets is maybe among the most severe shortcomings. Little, if nothing, is known about fundamental questions such as: (i) Are there any basic quantities characterizing the completeness of a molecular data set, for instance in terms of diversity and/or sparsity? (ii) On the basis of the inherent properties of the data set, representation and regressor, can we infer the performance of a model without actual training/test runs, as translated into the slope and offset of resulting learning curve. (iii) What, if any, characterizes the “correct” distribution in CCS. Answering such questions rigorously, i.e., based on the laws of physical chemistry, is not only of conceptual importance but would also benefit the practical design of more efficient/accurate QML models.

Other unresolved issues include (i) the lack of appropriate QML models that deal with intensive properties such as HOMO/LUMO energy, or dipole moments, which may require careful consideration of both local and long-ranged features of a molecule. (ii) The lack of high-accuracy data sets at the level of experimental quality (e.g., CCSD(T)-F12/CVQZ-F12 or multireference) for medium-sized molecules: published data sets of such quality are still limited to very few or small molecules, containing typically no more than three heavy atoms.


As the field has been growing massively and rapidly,<sup>16</sup> we can unfortunately not guarantee completeness of our outlook. Furthermore, several related important new research directions, i.e., going beyond the mere supervised learning problem of the electronic Schrödinger equation, possibly being out of the scope of “conventional” QML, have not been mentioned. They include, for example, variational autoencoders which can be used to help solving the inverse design challenge in CCS (e.g., applied to the design of improved molecular electronics<sup>400</sup>), the reconstruction of quantum states,<sup>401</sup> or the generation of molecular structures.<sup>402</sup> Other intriguing efforts deal with tackling the problem of reaction planning,<sup>403–408</sup> phase diagrams,<sup>409–414</sup> studying the electronic structure in more depth and detail,<sup>260,415–418</sup> or the systematic incorporation of experimental information in order to improve experimental design.<sup>419</sup>

To recap, we have provided succinct explanations and pointers to three major ingredients of QML: representation, regressor, and training set. We have briefly discussed select relevant studies which deal with the development and use of surrogate machine learning models of quantum properties throughout CCS. One of the primary goals of QML, i.e., rational computational discovery and design of compounds with desired properties, however, has not yet been achieved in general, and most of the relevant studies are either conducted in a high-throughput fashion, merely accelerated by QML, or rely on coarsening the problem through neglect of relevant degrees of freedom. We have pointed out several open

questions and challenges that must be overcome to reach this general goal, as well as potential solutions, and suggestions about interesting new research directions. Given the overall rapid growth and the multiple success cases already achieved in this young field, we are optimistic about its future and strongly believe that QML will develop into a helpful component for solving some of the long-standing problems in the atomistic sciences.

## AUTHOR INFORMATION

### Corresponding Author

**O. Anatole von Lilienfeld** – Faculty of Physics, University of Vienna, 1090 Vienna, Austria; Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, 4056 Basel, Switzerland;  [orcid.org/0000-0001-7419-0466](https://orcid.org/0000-0001-7419-0466); Email: [anatole.vonlilienfeld@univie.ac.at](mailto:anatole.vonlilienfeld@univie.ac.at)

### Author

**Bing Huang** – Faculty of Physics, University of Vienna, 1090 Vienna, Austria

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.chemrev.0c01303>

### Notes

The authors declare no competing financial interest.

### Biographies

Bing Huang (Hubei, China, 1987) was initially trained in physical chemistry under the supervision of Prof. Lin Zhuang in Wuhan University and completed his Ph.D. there in 2015, investigating and developing reactivity theory concerning solid surface. Afterwards, he moved to Basel, Switzerland, to work as a postdoc with Anatole von Lilienfeld at the Department of Chemistry, University of Basel, shifting research interests to the development of machine learning models and methods in quantum chemistry to explore chemical compound space. As of 2020, he has relocated to Vienna, Austria, to continue his postdoctoral research with Anatole von Lilienfeld at the Faculty of Physics, University of Vienna. His research interests include electronic structure theory, chemical reactivity theory, theoretical surface science and quantum machine learning.

O. Anatole von Lilienfeld (Rochester, Minnesota, USA, 1976) is a full university professor of computational materials discovery at the Faculty of Physics at the University of Vienna. Research in his laboratory deals with the development of improved methods for a first principles based understanding of chemical compound space using perturbation theory, machine learning, and high-performance computing. Previously, he was an associate and assistant professor at the University of Basel, Switzerland, and at the Free University of Brussels, Belgium. From 2007 to 2013, he worked for Argonne and Sandia National Laboratories after postdoctoral studies with Mark Tuckerman at New York University and at the Institute for Pure and Applied Mathematics at the University of California Los Angeles. In 2005, he was awarded a Ph.D. in computational chemistry from EPF Lausanne under the guidance of Ursula Röthlisberger. His diploma thesis work was done at ETH Zürich with Martin Quack and the University of Cambridge with Nicholas Handy. He studied chemistry at ETH Zürich, the Ecole de Chimie Polymers et Matériaux in Strasbourg, and the University of Leipzig. He serves as Editor-in-chief of the IOP journal *Machine Learning: Science and Technology* and on the editorial board of *Science Advances*. He has been on the editorial

board of *Nature's Scientific Data* from 2014 to 2019. He was the chair of the long IPAM UCLA program "Navigating Chemical Compound Space for Materials and Bio Design", which took place in 2011. He is the recipient of multiple awards, including a Swiss National Science Foundation postdoctoral grant (2005), a Harry S. Truman postdoctoral fellowship (2007), a Thomas Kuhn Paradigm Shift award (2013), a Swiss National Science professor fellowship (2013), the Odysseus grant from the Flemish Science Foundation (2016), an ERC consolidator grant (2017), and the Feynman Prize in Nanotechnology (2018).

## ACKNOWLEDGMENTS

O.A.v.L. acknowledges support from the Swiss National Science Foundation (407540\_167186 NFP 75 Big Data) and from the European Research Council (ERC-CoG grant QML and H2020 projects BIG-MAP and TREX). This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreements no. 952165 and no. 957189. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 772834). This result only reflects the author's view, and the EU is not responsible for any use that may be made of the information it contains. This work was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. We thank J. Wagner and F. A. Faber for helping with the design of Figures <sup>1</sup> and <sup>2</sup>.

## REFERENCES

- (1) Rupp, M. Special issue on machine learning and quantum mechanics. *Int. J. Quantum Chem.* **2015**, *115*, 1003–1004.
- (2) Rupp, M.; von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *J. Chem. Phys.* **2018**, *148*, 241401.
- (3) Schneider, W. F.; Guo, H. Machine Learning. *J. Phys. Chem. A* **2018**, *122*, 879.
- (4) Prezhdo, O. V. Advancing Physical Chemistry with Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 9656–9658.
- (5) Tkatchenko, A. Machine learning for chemical discovery. *Nat. Commun.* **2020**, *11*, 4125.
- (6) Schütt, K.; Chmiela, S.; von Lilienfeld, O.; Tkatchenko, A.; Tsuda, K.; Müller, K. *Machine Learning Meets Quantum Physics*; Lecture Notes in Physics; Springer International, 2020.
- (7) Ramakrishnan, R.; von Lilienfeld, O. A. *Reviews in Computational Chemistry*; John Wiley & Sons, 2017; Vol. 30; pp 225–256.
- (8) von Lilienfeld, O. A. Quantum machine learning in chemical compound space. *Angew. Chem., Int. Ed.* **2018**, *57*, 4164–4169.
- (9) Kitchin, J. R. Machine learning in catalysis. *Nat. Catal.* **2018**, *1*, 230–232.
- (10) Huang, B.; Symonds, N. O.; Lilienfeld, O. A. v. Quantum machine learning in chemistry and materials. *Handbook of Materials Modeling: Methods: Theory and Modeling* **2018**, 1–27.
- (11) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.
- (12) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The matter simulation ( $\tau$ ) evolution. *ACS Cent. Sci.* **2018**, *4*, 144–152.
- (13) Faber, F. A.; Anatole von Lilienfeld, O. Modeling Materials Quantum Properties with Machine Learning. *Materials Informatics: Methods, Tools and Applications* **2019**, 171–179.
- (14) Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019**, *119*, 6595.
- (15) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **2020**, *4*, 347.
- (16) von Lilienfeld, O. A.; Burke, K. Retrospective on a decade of machine learning for chemical discovery. *Nat. Commun.* **2020**, *11*, 4895.
- (17) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390.
- (18) Faber, F. A.; Christensen, A. S.; von Lilienfeld, O. A. *Machine Learning Meets Quantum Physics*; Springer, 2020; pp 155–169.
- (19) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *arXiv* **2020**, arXiv:2010.07067
- (20) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525.
- (21) Chibani, S.; Coudert, F.-X. Machine learning approaches for the prediction of materials properties. *APL Mater.* **2020**, *8*, 080701.
- (22) Dral, P. O. Quantum chemistry in the age of machine learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336–2347.
- (23) Haghhighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation. *Data, and Machine Learning Methods* **2020**, *6*, 1527–1542.
- (24) Hoffmann, R.; Malrieu, J.-P. Simulation vs. Understanding: A Tension, in Quantum Chemistry and Beyond. Part A. Stage Setting. *Angew. Chem., Int. Ed.* **2020**, *59*, 12590–12610.
- (25) Hoffmann, R.; Malrieu, J.-P. Simulation vs. Understanding: A Tension, in Quantum Chemistry and Beyond. Part B. The March of Simulation, for Better or Worse. *Angew. Chem., Int. Ed.* **2020**, *59*, 13156–13178.
- (26) Hoffmann, R.; Malrieu, J.-P. Simulation vs. Understanding: A Tension, in Quantum Chemistry and Beyond. Part C. Toward Consilience. *Angew. Chem., Int. Ed.* **2020**, *59*, 13694–13710.
- (27) von Lilienfeld, O. A. Introducing Machine Learning: Science and Technology. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 010201.
- (28) Pyzer-Knapp, E. O.; Cuff, J.; Patterson, J.; Isayev, O.; Maskell, S. Welcome to the first issue of Applied AI Letters. *Appl. AI Lett.* **2020**, *1*, e8.
- (29) Buckingham, A.; Utting, B. Intermolecular forces. *Annu. Rev. Phys. Chem.* **1970**, *21*, 287–316.
- (30) von Lilienfeld, O. A. First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties. *Int. J. Quantum Chem.* **2013**, *113*, 1676–1689.
- (31) Faulon, J. L. Stochastic generator of chemical structure: 1. Application to the structure elucidation of large molecules. *J. Chem. Inf. Comp. Sci.* **1994**, *34*, 1204–1218.
- (32) Braun, J.; Gugisch, R.; Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. MOLGEN-CID — A canonizer for molecules and graphs accessible through the internet. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 542–548.
- (33) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504.
- (34) von Rudorff, G. F.; von Lilienfeld, O. A. Simplifying inverse materials design problems for fixed lattices with alchemical chirality. *Sci. Adv.* **2021**, *7*, No. eabf1173.
- (35) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.
- (36) Kato, T. On the eigenfunctions of many-particle systems in quantum mechanics. *Communications on Pure and Applied Mathematics* **1957**, *10*, 151–177.
- (37) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
- (38) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.



- (39) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A., III ReaxFF: A reactive force field for hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396–9409.
- (40) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; et al. The ReaxFF reactive force-field: development, applications and future directions. *NPJ Comput. Mater.* **2016**, *2*, 15011.
- (41) Ischtwan, J.; Collins, M. A. Molecular potential energy surfaces by interpolation. *J. Chem. Phys.* **1994**, *100*, 8080–8088.
- (42) Wagner, A. F.; Schatz, G. C.; Bowman, J. M. The evaluation of fitting functions for the representation of an O(3P)+H<sub>2</sub> potential energy surface. *I. J. Chem. Phys.* **1981**, *74*, 4960–4983.
- (43) Schatz, G. C. The analytical representation of electronic potential-energy surfaces. *Rev. Mod. Phys.* **1989**, *61*, 669–688.
- (44) Sumpter, B. G.; Noid, D. W. Potential energy surfaces for macromolecules. A neural network technique. *Chem. Phys. Lett.* **1992**, *192*, 455–462.
- (45) Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural network models of potential energy surfaces. *J. Chem. Phys.* **1995**, *103*, 4129–4137.
- (46) Brown, D. F.; Gibbs, M. N.; Clary, D. C. Combining ab initio computations, neural networks, and diffusion Monte Carlo: An efficient method to treat weakly bound molecules. *J. Chem. Phys.* **1996**, *105*, 7597–7604.
- (47) Lorenz, S.; Gross, A.; Scheffler, M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chem. Phys. Lett.* **2004**, *395*, 210.
- (48) Manzhos, S.; Carrington, T., Jr. A random-sampling high dimensional model representation neural network for building potential energy surfaces. *J. Chem. Phys.* **2006**, *125*, 084109–084123.
- (49) Handley, C. M.; Popelier, P. L. A. Dynamically polarizable water potential based on multipole moments trained by machine learning. *J. Chem. Theory Comput.* **2009**, *5*, 1474.
- (50) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
- (51) Behler, J.; Martonak, R.; Donadio, D.; Parrinello, M. Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential. *Phys. Rev. Lett.* **2008**, *100*, 185501.
- (52) Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem., Int. Ed.* **2017**, *56*, 12828–12840.
- (53) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.
- (54) Faraji, S.; Ghasemi, S. A.; Rostami, S.; Rasoulkhani, R.; Schaefer, B.; Goedecker, S.; Amsler, M. High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 104105.
- (55) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet-A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
- (56) Unke, O. T.; Meuwly, M. A reactive, scalable, and transferable model for molecular energies from a neural network approach based on local information. *J. Chem. Phys.* **2018**, *148*, 241708.
- (57) Manzhos, S.; Carrington, T., Jr. Neural network potential energy surfaces for small molecules and reactions. *Chem. Rev.* **2020**, DOI: 10.1021/acs.chemrev.0c00665
- (58) Ho, T.; Rabitz, H. A general method for constructing multidimensional molecular potential energy surfaces from abinitio calculations. *J. Chem. Phys.* **1996**, *104*, 2584–2597.
- (59) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
- (60) Pozun, Z. D.; Hansen, K.; Sheppard, D.; Rupp, M.; Müller, K.-R.; Henkelman, G. Optimizing transition states via kernel-based machine learning. *J. Chem. Phys.* **2012**, *136*, 174101–174109.
- (61) Thompson, A.; Swiler, L.; Trott, C.; Foiles, S.; Tucker, G. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2015**, *285*, 316–330.
- (62) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.
- (63) Botu, V.; Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.
- (64) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309.
- (65) Soloviov, M.; Meuwly, M. Reproducing kernel potential energy surfaces in biomolecular simulations: Nitric oxide binding to myoglobin. *J. Chem. Phys.* **2015**, *143*, 105103.
- (66) Glielmo, A.; Sollich, P.; De Vita, A. Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 214302.
- (67) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, No. e1603015.
- (68) Unke, O. T.; Meuwly, M. Toolkit for the construction of reproducing kernel-based representations of data: Application to multidimensional potential energy surfaces. *J. Chem. Inf. Model.* **2017**, *57*, 1923–1931.
- (69) Dragoni, D.; Daff, T. D.; Csányi, G.; Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. Materials* **2018**, *2*, 013808.
- (70) Bereau, T.; DiStasio Jr, R. A.; Tkatchenko, A.; Von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*, 241706.
- (71) Deringer, V. L.; Caro, M. A.; Jana, R.; Aarva, A.; Elliott, S. R.; Laurila, T.; Csányi, G.; Pastewka, L. Computational Surface Chemistry of Tetrahedral Amorphous Carbon by Combining Machine Learning and Density Functional Theory. *Chem. Mater.* **2018**, *30*, 7438–7445.
- (72) Caro, M. A.; Aarva, A.; Deringer, V. L.; Csányi, G.; Laurila, T. Reactivity of amorphous carbon surfaces: rationalizing the role of structural motifs in functionalization using machine learning. *Chem. Mater.* **2018**, *30*, 7446–7455.
- (73) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.
- (74) Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Comput. Phys. Commun.* **2019**, *240*, 38.
- (75) Sauceda, H. E.; Gastegger, M.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. Molecular force fields with gradient-domain machine learning (GDML): Comparison and synergies with classical force fields. *J. Chem. Phys.* **2020**, *153*, 124109.
- (76) Kamath, A.; Vargas-Hernández, R. A.; Krems, R. V.; Carrington Jr, T.; Manzhos, S. Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy. *J. Chem. Phys.* **2018**, *148*, 241702.
- (77) Käser, S.; Koner, D.; Christensen, A. S.; von Lilienfeld, O. A.; Meuwly, M. Machine Learning Models of Vibrating H<sub>2</sub>CO: Comparing Reproducing Kernels, FCHL, and PhysNet. *J. Phys. Chem. A* **2020**, *124*, 8853–8865.
- (78) Csányi, G.; Albaret, T.; Payne, M. C.; De Vita, A. D. Learn on the Fly: A Hybrid Classical and Quantum-Mechanical Molecular Dynamics Simulation. *Phys. Rev. Lett.* **2004**, *93*, 175503.

- (79) Podryabinkin, E. V.; Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **2017**, *140*, 171–180.
- (80) Jacobsen, T.; Jørgensen, M.; Hammer, B. On-the-Fly Machine Learning of Atomic Potential in Density Functional Theory Structure Optimization. *Phys. Rev. Lett.* **2018**, *120*, 026102.
- (81) Jørgensen, M. S.; Larsen, U. F.; Jacobsen, K. W.; Hammer, B. Exploration Versus Exploitation in Global Atomistic Structure Optimization. *J. Phys. Chem. A* **2018**, *122*, 1504–1509.
- (82) Zhang, L.; Lin, D.-Y.; Wang, H.; Car, R.; E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **2019**, *3*, 023804.
- (83) Westermayr, J.; Gastegger, M.; Menger, M. F.; Mai, S.; González, L.; Marquetand, P. Machine learning enables long time scale molecular photodynamics simulations. *Chem. Sci.* **2019**, *10*, 8100–8107.
- (84) Jia, W.; Wang, H.; Chen, M.; Lu, D.; Lin, L.; Car, R.; Weinan, E.; Zhang, L., et al. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA 2020*; p 1
- (85) Marder, S. R.; Beratan, D. N.; Cheng, L.-T. Approaches for Optimizing the First Electronic Hyperpolarizability of Conjugated Organic Molecules. *Science* **1991**, *252*, 103–106.
- (86) Kuhn, C.; Beratan, D. N. Inverse Strategies for Molecular Design. *J. Phys. Chem.* **1996**, *100*, 10595–10599.
- (87) Ceder, G. Predicting properties from scratch. *Science* **1998**, *280*, 1099–1100.
- (88) Franceschetti, A.; Zunger, A. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature* **1999**, *402*, 60.
- (89) Jóhannesson, G. H.; Bligaard, T.; Ruban, A. V.; Skriver, H. L.; Jacobsen, K. W.; Nørskov, J. K. Combined Electronic Structure and Evolutionary Search Approach to Materials Design. *Phys. Rev. Lett.* **2002**, *88*, 255506.
- (90) von Lilienfeld, O. A.; Lins, R.; Rothlisberger, U. Variational particle number approach for rational compound design. *Phys. Rev. Lett.* **2005**, *95*, 153002.
- (91) Mounet, N.; Gibertini, M.; Schwaller, P.; Campi, D.; Merkys, A.; Marrazzo, A.; Sohier, T.; Castelli, I. E.; Cepellotti, A.; Pizzi, G.; et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **2018**, *13*, 246–252.
- (92) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **2010**, *22*, 3762.
- (93) George, J.; Hautier, G. Chemist versus Machine: Traditional Knowledge versus Machine Learning Techniques. *Trends Chem.* **2021**, *3*, 86.
- (94) Wilson, E. B., Jr. Four Dimensional Electron Density Function. *J. Chem. Phys.* **1962**, *36*, 2232.
- (95) Politzer, P.; Parr, R. G. Some new energy formulas for atoms and molecules. *J. Chem. Phys.* **1974**, *61*, 4258.
- (96) Mezey, P. G. Electronic energy inequalities for isoelectronic molecular systems. *Theor. Chim. Acta* **1980**, *59*, 321–332.
- (97) Mezey, P. G. New global constraints on electronic energy hypersurfaces. *Int. J. Quantum Chem.* **1986**, *29*, 85–99.
- (98) von Lilienfeld, O. A.; Tuckerman, M. E. Molecular grand-canonical ensemble density functional theory and exploration of chemical space. *J. Chem. Phys.* **2006**, *125*, 154104.
- (99) Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. Designing molecules by optimizing potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228.
- (100) Lesiuk, M.; Balawender, R.; Zachara, J. Higher order alchemical derivatives from coupled perturbed self-consistent field theory. *J. Chem. Phys.* **2012**, *136*, 034104.
- (101) Munoz, M.; Cardenas, C. How predictive could alchemical derivatives be? *Phys. Chem. Chem. Phys.* **2017**, *19*, 16003–16012.
- (102) Fias, S.; Chang, K. S.; von Lilienfeld, O. A. Alchemical normal modes unify chemical space. *J. Phys. Chem. Lett.* **2019**, *10*, 30–39.
- (103) von Rudorff, G. F.; von Lilienfeld, O. A. Alchemical perturbation density functional theory. *Phys. Rev. Research* **2020**, *2*, 023220.
- (104) Barkoutsos, P. K.; Gkritis, F.; Ollitrault, P. J.; Sokolov, I. O.; Woerner, S.; Tavernelli, I. Quantum algorithm for alchemical optimization in material design. *arXiv 2020*.arXiv:2008.06449
- (105) Marzari, N.; de Gironcoli, S.; Baroni, S. Structure and Phase Stability of Ga<sub>x</sub>In<sub>1-x</sub>P solid solutions from computational Alchemy. *Phys. Rev. Lett.* **1994**, *72*, 4001.
- (106) Beste, A.; Harrison, R. J.; Yanai, T. Direct computation of general chemical energy differences: Application to ionization potentials, excitation, and bond energies. *J. Chem. Phys.* **2006**, *125*, 074101.
- (107) Weigend, F.; Schrod, C.; Ahlrichs, R. Atom distributions in binary atom clusters: A perturbational approach and its validation in a case study. *J. Chem. Phys.* **2004**, *121*, 10380.
- (108) Weigend, F. Extending DFT-based genetic algorithms by atom-to-place re-assignment via perturbation theory: a systematic and unbiased approach to structures of mixed-metallic clusters. *J. Chem. Phys.* **2014**, *141*, 134103.
- (109) Rinderspacher, B. C.; Andzelm, J.; Rawlett, A.; Dougherty, J.; Beratan, D. N.; Yang, W. Discrete Optimization of Electronic Hyperpolarizabilities in a Chemical Subspace. *J. Chem. Theory Comput.* **2009**, *5*, 3321.
- (110) Sheppard, D.; Henkelman, G.; von Lilienfeld, O. A. Alchemical derivatives of reaction energetics. *J. Chem. Phys.* **2010**, *133*, 084104.
- (111) Balawender, R.; Welearegay, M. A.; Lesiuk, M.; De Proft, F.; Geerlings, P. Exploring Chemical Space with the Alchemical Derivatives. *J. Chem. Theory Comput.* **2013**, *9*, 5327–5340.
- (112) Chang, K. Y. S.; Fias, S.; Ramakrishnan, R.; von Lilienfeld, O. A. Fast and accurate predictions of covalent bonds in chemical space. *J. Chem. Phys.* **2016**, *144*, 174110.
- (113) Al-Hamdani, Y. S.; Michaelides, A.; von Lilienfeld, O. A. Exploring dissociative water adsorption on isoelectronically BN doped graphene using alchemical derivatives. *J. Chem. Phys.* **2017**, *147*, 164113.
- (114) Fias, S.; Heidar-Zadeh, F.; Geerlings, P.; Ayers, P. W. Chemical transferability of functional groups follows from the nearsightedness of electronic matter. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 11633–11638.
- (115) Balawender, R.; Lesiuk, M.; De Proft, F.; Geerlings, P. Exploring Chemical Space with Alchemical Derivatives: BN-Simultaneous Substitution Patterns in C60. *J. Chem. Theory Comput.* **2018**, *14*, 1154.
- (116) von Rudorff, G. F.; von Lilienfeld, O. A. Atoms in molecules from alchemical perturbation density functional theory. *J. Phys. Chem. B* **2019**, *123*, 10073–10082.
- (117) Saravanan, K.; Kitchin, J. R.; von Lilienfeld, O. A.; Keith, J. A. Alchemical Predictions for Computational Catalysis: Potential and Limitations. *J. Phys. Chem. Lett.* **2017**, *8*, 5002–5007.
- (118) Griego, C. D.; Saravanan, K.; Keith, J. A. Benchmarking computational alchemy for carbide, nitride, and oxide catalysts. *Adv. Theor. Simul.* **2019**, *2*, 1800142.
- (119) Griego, C. D.; Kitchin, J. R.; Keith, J. A. Acceleration of catalyst discovery with easy, fast, and reproducible computational alchemy. *Int. J. Quantum Chem.* **2021**, *121*, No. e26380.
- (120) von Rudorff, G. F.; von Lilienfeld, O. A. Rapid and accurate molecular deprotonation energies from quantum alchemy. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10519–10525.
- (121) Muñoz, M.; Robles-Navarro, A.; Fuentealba, P.; Cárdenas, C. Predicting Deprotonation Sites Using Alchemical Derivatives. *J. Phys. Chem. A* **2020**, *124*, 3754–3760.
- (122) Pérez, A.; von Lilienfeld, O. A. Path integral computation of quantum free energy differences due to alchemical transformations involving mass and potential. *J. Chem. Theory Comput.* **2011**, *7*, 2358.

- (123) Ceriotti, M.; Markland, T. E. Efficient methods and practical guidelines for simulating isotope effects. *J. Chem. Phys.* **2013**, *138*, 014112.
- (124) Geerlings, P.; De Proft, F. D.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793.
- (125) Yang, W.; Zhang, Y.; Ayers, P. W. Degenerate ground states and fractional number of electrons in density and density reduced matrix functional theory. *Phys. Rev. Lett.* **2000**, *84*, 5172.
- (126) Zeng, X.; Hu, H.; Hu, X.; Cohen, A. J.; Yang, W. Ab initio quantum mechanical/molecular simulation of electron transfer process: Fractional electron approach. *J. Chem. Phys.* **2008**, *128*, 124510.
- (127) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Discontinuous nature of the exchange-correlation functional in strongly correlated systems. *Phys. Rev. Lett.* **2009**, *102*, 066403.
- (128) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273.
- (129) Bultinck, P.; Gironés, X.; Carbo-Dorcaz, R. Molecular quantum similarity: theory and applications. *Rev. Comput. Chem.* **2005**, *21*, 127.
- (130) Kaji, M. Mendeleev's Discovery of the Periodic Law: The Origin and the Reception. *Found. Chem.* **2003**, *5*, 189.
- (131) Pauling, L.; Yost, D. M. The additivity of the energies of normal covalent bonds. *Proc. Natl. Acad. Sci. U. S. A.* **1932**, *18*, 414.
- (132) Pettifor, D. G. A chemical scale for crystal-structure maps. *Solid State Commun.* **1984**, *51*, 31–34.
- (133) Glawe, H.; Sanna, A.; Gross, E. K. U.; Marques, M. A. L. The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining. *New J. Phys.* **2016**, *18*, 093011.
- (134) Glawe, H.; Sanna, A.; Gross, E.; Marques, M. A. The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining. *New J. Phys.* **2016**, *18*, 093011.
- (135) Allahyari, Z.; Oganov, A. R. Nonempirical definition of the Mendeleev numbers: Organizing the chemical space. *J. Phys. Chem. C* **2020**, *124*, 23867.
- (136) Glushkovsky, A. AI Discovering a Coordinate System of Chemical Elements: Dual Representation by Variational Autoencoders. *arXiv* **2020**, arXiv:2011.12090
- (137) Huang, B.; Zhuang, L.; Xiao, L.; Lu, J. Bond-energy decoupling: principle and application to heterogeneous catalysis. *Chem. Sci.* **2013**, *4*, 606–611.
- (138) Wells, P. R. Linear Free Energy Relationships. *Chem. Rev.* **1963**, *63*, 171–219.
- (139) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.
- (140) Hansch, C.; Leo, A.; Taft, R. W. A survey of Hammett substituent constants and resonance and field parameters. *Chem. Rev.* **1991**, *91*, 165–195.
- (141) Hammett, L. P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935**, *17*, 125–136.
- (142) Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **1955**, *77*, 334–338.
- (143) Bell, R. P. The theory of reactions involving proton transfers. *Proc. R. Soc. Lond., A* **1936**, *154*, 414–429.
- (144) Evans, M. G.; Polanyi, M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Trans. Faraday Soc.* **1936**, *32*, 1333–1360.
- (145) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133.
- (146) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864.
- (147) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (148) Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew-Burke-Ernzerhof exchange-correlation functional. *J. Chem. Phys.* **1999**, *110*, 5029.
- (149) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (150) Hammer, B.; Nørskov, J. K. Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.* **1995**, *343*, 211–220.
- (151) Hammer, B.; Nørskov, J. K. Why gold is the noblest of all the metals. *Nature* **1995**, *376*, 238–240.
- (152) Calle-Vallejo, F.; Martínez, J. I.; García-Lastra, J. M.; Sautet, P.; Loffreda, D. Fast Prediction of Adsorption Properties for Platinum Nanocatalysts with Generalized Coordination Numbers. *Angew. Chem., Int. Ed.* **2014**, *53*, 8316–8319 eprint.
- (153) Calle-Vallejo, F.; Tymoczko, J.; Colic, V.; Vu, Q. H.; Pohl, M. D.; Morgenstern, K.; Loffreda, D.; Sautet, P.; Schuhmann, W.; Bandarenka, A. S. Finding optimal surface sites on heterogeneous catalysts by counting nearest neighbors. *Science* **2015**, *350*, 185–189.
- (154) Huang, B.; Xiao, L.; Lu, J.; Zhuang, L. Spatially Resolved Quantification of the Surface Reactivity of Solid Catalysts. *Angew. Chem.* **2016**, *128*, 6347–6351 eprint.
- (155) van Santen, R. A.; Neurock, M.; Shetty, S. G. Reactivity Theory of Transition Metal Surfaces: A Bronsted-Evans-Polanyi Linear Activation Energy Free Energy Analysis. *Chem. Rev.* **2010**, *110*, 2005–2048.
- (156) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. Scaling Properties of Adsorption Energies for Hydrogen-Containing Molecules on Transition-Metal Surfaces. *Phys. Rev. Lett.* **2007**, *99*, 016105.
- (157) Fernández, E. M.; Moses, P. G.; Toftelund, A.; Hansen, H. A.; Martínez, J. I.; Abild-Pedersen, F.; Kleis, J.; Hinnemann, B.; Rossmeisl, J.; Bligaard, T.; Nørskov, J. K. Scaling Relationships for Adsorption Energies on Transition Metal Oxide, Sulfide, and Nitride Surfaces. *Angew. Chem., Int. Ed.* **2008**, *47*, 4683–4686 eprint.
- (158) Calle-Vallejo, F.; Martínez, J. I.; García-Lastra, J. M.; Rossmeisl, J.; Koper, M. T. M. Physical and Chemical Nature of the Scaling Relations between Adsorption Energies of Atoms on Metal Surfaces. *Phys. Rev. Lett.* **2012**, *108*, 116103.
- (159) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **2009**, *1*, 37.
- (160) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford Science Publications, 1989.
- (161) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793–1874.
- (162) Geerlings, P.; Fias, S.; Boisdenghien, Z.; DE Proft, F. Conceptual DFT: chemistry from the linear response function. *Chem. Soc. Rev.* **2014**, *43*, 4989–5008.
- (163) Huang, B.; von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, 161102.
- (164) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite ( $ABC_2D_6$ ) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.
- (165) Ong, S. P.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Bailey, D.; Skinner, D.; Persson, K. A.; Ceder, G. The Materials Project. 2011; <http://materialsproject.org/>.
- (166) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.
- (167) Ye, W.; Chen, C.; Wang, Z.; Chu, I.-H.; Ong, S. P. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* **2018**, *9*, 3800.
- (168) Schmidt, J.; Shi, J.; Borlido, P.; Chen, L.; Botti, S.; Marques, M. A. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **2017**, *29*, 5090–5103.
- (169) Legrain, F.; Carrete, J.; van Roekeghem, A.; Curtarolo, S.; Mingo, N. How Chemical Composition Alone Can Predict Vibra-



tional Free Energies and Entropies of Solids. *Chem. Mater.* **2017**, *29*, 6220–6227.

(170) Hansen, K.; Biegler, F.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A.; et al. Interaction potentials in molecules and non-local information in chemical space. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(171) Ramakrishnan, R.; Dral, P. D.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

(172) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH, 2002.

(173) Huang, B.; von Lilienfeld, O. A. The “DNA” of chemistry: Scalable quantum machine learning with “amons”. *arXiv* **2017**, arXiv:1707.04146, submitted to *Nature*.

(174) Braun, J.; Kerber, A.; Meringer, M.; Rücker, C. Similarity of molecular descriptors: The equivalence of Zagreb indices and walk counts. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 163–176.

(175) Visco, J.; Pophale, R. S.; Rintoul, M. D.; Faulon, J. L. Developing a methodology for an inverse quantitative structure activity relationship using the signature molecular descriptor. *J. Mol. Graphics Modell.* **2002**, *20*, 429–438.

(176) Faulon, J.-L.; Visco, D. P., Jr.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 707.

(177) Martin, S.; Roe, D.; Faulon, J.-L. Predicting protein-protein interactions using signature products. *Bioinformatics* **2005**, *21*, 218–226.

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

(179) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Côté, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361.

(180) Besnard, J.; et al. Automated design of ligands to polypharmacological profiles. *Nature* **2012**, *492*, 215.

(181) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(182) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504–1508.

(183) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

(184) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091–6098.

(185) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(186) Unsleber, J. P.; Reiher, M. The Exploration of Chemical Reaction Networks. *Annu. Rev. Phys. Chem.* **2020**, *71*, 121–142 eprint.

(187) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

(188) Yang, Q.; Bassyouni, A.; Butler, C. R.; Hou, X.; Jenkinson, S.; Price, D. A.; et al. Ligand biological activity predicted by cleaning positive and negative chemical correlations. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3373–3378.

(189) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% robust

molecular string representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

(190) Hoffmann, C.; Menichetti, R.; Kanekal, K. H.; Bereau, T. Controlled exploration of chemical space by machine learning of coarse-grained representations. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2019**, *100*, 033302.

(191) John, S. T.; Csányi, G. Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials. *J. Phys. Chem. B* **2017**, *121*, 10934–10949.

(192) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **2019**, *5*, 755–767.

(193) Karthikeyan, M.; Glen, R. C.; Bender, A. General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks. *J. Chem. Inf. Model.* **2005**, *45*, 581–590.

(194) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.

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

(196) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; Dietterich, T., Ed.; MIT Press: Cambridge, 2006; [www.GaussianProcess.org](http://www.GaussianProcess.org).

(197) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater.* **2019**, *5*, 83.

(198) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media, 2013.

(199) Brown, A.; Braams, B. J.; Christoffel, K.; Jin, Z.; Bowman, J. M. Classical and quasiclassical spectral analysis of CH<sub>5</sub><sup>+</sup> using an ab initio potential energy surface. *J. Chem. Phys.* **2003**, *119*, 8790.

(200) Neal, R. M. *Bayesian Learning for Neural Networks*; Springer, 1996; pp 29–53.

(201) çaylak, O.; von Lilienfeld, O. A.; Baumeier, B. Wasserstein metric for improved quantum machine learning with adjacency matrix representations. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 03LT01.

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

(203) Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.

(204) Cortes, C.; Jackel, L. D.; Solla, S. A.; Vapnik, V.; Denker, J. S. Learning curves: Asymptotic values and rate of convergence. *Adv. Neur. Inform. Process. Syst.* **1994**, *6*, 327–334.

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

(206) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. Operators in quantum machine learning: Response properties in chemical space. *J. Chem. Phys.* **2019**, *150*, 064105.

(207) Christensen, A. S.; von Lilienfeld, O. A. Operator quantum machine learning: Navigating the chemical space of response properties. *Chimia* **2019**, *73*, 1028–1031.

(208) Christensen, A. S.; Bratholm, L.; Faber, F. A.; von Lilienfeld, O. A. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152*, 044107.

(209) Gastegger, M.; Schütt, K. T.; Müller, K.-R. Machine learning of solvent effects on molecular spectra and reactions. *arXiv* **2020**, arXiv:2010.14942

(210) Christensen, A. S.; von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045018.

(211) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96.

- (212) Bragato, M.; von Rudorff, G. F.; von Lilienfeld, O. A. Data enhanced Hammett Equation: Reaction Barriers in Chemical Space. *Chem. Sci.* **2020**, *11*, 11859.
- (213) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2009.
- (214) Moussa, J. E. Comment on "Fast and Accurate Modeling of Molecular Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *109*, 059801.
- (215) von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quantum Chem.* **2015**, *115*, 1084.
- (216) Pozdnyakov, S. N.; Willatt, M. J.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Incompleteness of Atomic Structure Representations. *Phys. Rev. Lett.* **2020**, *125*, 166001.
- (217) Parsaeifard, B.; De, D. S.; Christensen, A. S.; Faber, F. A.; Kocer, E.; De, S.; Behler, J.; von Lilienfeld, A.; Goedecker, S. An assessment of the structural resolution of various fingerprints commonly used in machine learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015018.
- (218) Ramakrishnan, R.; von Lilienfeld, O. A. Many Molecular Properties from One Kernel in Chemical Space. *Chimia* **2015**, *69*, 182.
- (219) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.
- (220) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning, ICML, 2017*.
- (221) Eickenberg, M.; Exarchakis, G.; Hirn, M.; Mallat, S.; Thiry, L. Solid harmonic wavelet scattering for predictions of molecule properties. *J. Chem. Phys.* **2018**, *148*, 241732.
- (222) Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine learning of molecular properties: Locality and active learning. *J. Chem. Phys.* **2018**, *148*, 241727.
- (223) Welborn, M.; Cheng, L.; Miller III, T. F. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.
- (224) Langer, M. F.; Goßmann, A.; Rupp, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *arXiv2020*, arXiv:2003.12081v2.
- (225) Englert, B.-G. *Semiclassical Theory of Atoms*; Springer, 1988.
- (226) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Reply to Comment on "Fast and Accurate Modeling of Molecular Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *109*, 059802.
- (227) Paliana, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3*, 2810.
- (228) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.
- (229) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
- (230) Pronobis, W.; Tkatchenko, A.; Müller, K.-R. Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules. *J. Chem. Theory Comput.* **2018**, *14*, 2991.
- (231) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094.
- (232) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic Spectra from TDDFT and Machine Learning in Chemical Space. *J. Chem. Phys.* **2015**, *143*, 084111.
- (233) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **2017**, *13*, S255–S264.
- (234) Stuke, A.; Todorović, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *J. Chem. Phys.* **2019**, *150*, 204121.
- (235) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Deep learning spectroscopy: neural networks for molecular excitation spectra. *Adv. Sci.* **2019**, *6*, 1801367.
- (236) Zhu, L.; Amsler, M.; Fuhrer, T.; Schaefer, B.; Faraji, S.; Rostami, S.; Ghasemi, S. A.; Sadeghi, A.; Grauzinyte, M.; Wolverton, C.; et al. A fingerprint based metric for measuring similarities of crystalline structures. *J. Chem. Phys.* **2016**, *144*, 034203.
- (237) Schütt, O.; VandeVondele, J. Machine learning adaptive basis sets for efficient large scale density functional theory simulation. *J. Chem. Theory Comput.* **2018**, *14*, 4168–4175.
- (238) Babaei, M.; Azar, Y. T.; Sadeghi, A. Locality meets machine learning: Excited and ground-state energy surfaces of large systems at the cost of small ones. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2020**, *101*, 115132.
- (239) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148*, 241718.
- (240) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *99*, 014104.
- (241) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural networks potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
- (242) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **2020**, *12*, 945–951.
- (243) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **2017**, *3*, e1701816.
- (244) Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine learning of molecular properties: Locality and active learning. *J. Chem. Phys.* **2018**, *148*, 241727.
- (245) Willatt, M. J.; Musil, F.; Ceriotti, M. Atom-density representations for machine learning. *J. Chem. Phys.* **2019**, *150*, 154110.
- (246) Taylor, M. E.; Stone, P. Transfer learning for reinforcement learning domains: A survey. *J. Mach. Learn. Res.* **2009**, *10*, 1633.
- (247) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.
- (248) Hobday, S.; Smith, R.; Belbruno, J. Applications of neural networks to fitting interatomic potential functions. *Modell. Simul. Mater. Sci. Eng.* **1999**, *7*, 397.
- (249) Bereau, T.; Andrienko, D.; von Lilienfeld, O. A. Transferable atomic multipole machine learning models for small organic molecules. *J. Chem. Theory Comput.* **2015**, *11*, 3225.
- (250) Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 045131.
- (251) Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Likhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Discovering a transferable charge assignment model using machine learning. *J. Phys. Chem. Lett.* **2018**, *9*, 4495–4501.
- (252) Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A. E.; Likhov, A.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks. *J. Chem. Theory Comput.* **2018**, *14*, 4687–4698.

- (253) Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 2120–2125 PMID: 26146493.
- (254) Kranz, J. J.; Kubillus, M.; Ramakrishnan, R.; von Lilienfeld, O. A.; Elstner, M. Generalized density-functional tight-binding repulsive potentials from unsupervised machine learning. *J. Chem. Theory Comput.* **2018**, *14*, 2341–2352.
- (255) Stöhr, M.; Medrano Sandonas, L.; Tkatchenko, A. Accurate Many-Body Repulsive Potentials for Density-Functional Tight Binding from Deep Tensor Neural Networks. *J. Phys. Chem. Lett.* **2020**, *11*, 6835–6843.
- (256) Zubatyuk, T.; Nebgen, B.; Lubbers, N.; Smith, J. S.; Zubatyuk, R.; Zhou, G.; Koh, C.; Barros, K.; Isayev, O.; Tretiak, S. Machine Learned Hückel Theory: Interfacing Physics and Deep Neural Networks. *arXiv* **2019**, arXiv:1909.12963
- (257) Ramakrishnan, R.; Dral, P.; Rupp, M.; von Lilienfeld, O. A. Big Data meets Quantum Chemistry Approximations: The A-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.
- (258) Mezei, P. D.; von Lilienfeld, O. A. Noncovalent Quantum Machine Learning Corrections to Density Functionals. *J. Chem. Theory Comput.* **2020**, *16*, 2647–2653.
- (259) Griego, C. D.; Zhao, L.; Saravanan, K.; Keith, J. A. Machine Learning Corrected Alchemical Perturbation Density Functional Theory for Catalysis Applications. *AIChE J.* **2020**, *66*, No. e17041.
- (260) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223.
- (261) Townsend, J.; Vogiatzis, K. D. Transferable MP2-Based Machine Learning for Accurate Coupled-Cluster Energies. *J. Chem. Theory Comput.* **2020**, *16*, 7453.
- (262) Nandi, A.; Qu, C.; Houston, P.; Conte, R.; Bowman, J. M. Delta-Machine Learning for Potential Energy Surfaces: A PIP approach to bring a DFT-based PES to CCSD (T) Level of Theory. *J. Chem. Phys.* **2021**, *154*, 051102.
- (263) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 theory for molecular energies of first- and second-row compounds. *J. Chem. Phys.* **1991**, *94*, 7221–7230.
- (264) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. Gaussian-3 theory using reduced Møller-Plesset order. *J. Chem. Phys.* **1999**, *110*, 4703–4709.
- (265) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.
- (266) Zaspel, P.; Huang, B.; Harbrecht, H.; von Lilienfeld, O. A. Boosting quantum machine learning models with multi-level combination technique: Pople diagrams revisited. *J. Chem. Theory Comput.* **2019**, *15*, 1546.
- (267) Le Gratiet, L.; Garnier, J. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *Int. J. Uncertain. Quan.* **2014**, *4*, 365.
- (268) Kennedy, M. C.; O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **2000**, *87*, 1–13.
- (269) Cui, J.; Krems, R. V. Gaussian Process Model for Collision Dynamics of Complex Molecules. *Phys. Rev. Lett.* **2015**, *115*, 073202.
- (270) Pilia, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156–163.
- (271) Batra, R.; Pilia, G.; Uberuaga, B. P.; Ramprasad, R. Multifidelity Information Fusion with Machine Learning: A Case Study of Dopant Formation Energies in Hafnia. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24906.
- (272) Wiens, A. E.; Copan, A. V.; Schaefer, H. F. Multi-fidelity Gaussian process modeling for chemical energy surfaces. *Chem. Phys. Lett.: X* **2019**, *3*, 100022.
- (273) Egorova, O.; Hafizi, R.; Woods, D. C.; Day, G. M. Multifidelity Statistical Machine Learning for Molecular Crystal Structure Prediction. *J. Phys. Chem. A* **2020**, *124*, 8065–8078.
- (274) Garcke, J.; Griebel, M.; Thess, M. Data Mining with Sparse Grids. *Computing* **2001**, *67*, 225–253.
- (275) Delvos, F. J. d-Variate Boolean interpolation. *J. Approximation Theory* **1982**, *34*, 99–114.
- (276) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359 Conference Name: IEEE Transactions on Knowledge and Data Engineering.
- (277) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.
- (278) Olivas, E. S.; Guerrero, J. D. M.; Sober, M. M.; Bedito, J. R. M.; Lopez, A. J. S. Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques; IGI Publishing: Hershey, PA, 2009.
- (279) Iovanac, N. C.; Savoie, B. M. Simpler is Better: How Linear Prediction Tasks Improve Transfer Learning in Chemical Autoencoders. *J. Phys. Chem. A* **2020**, *124*, 3679–3685 PMID: 32267698.
- (280) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 8683–8694.
- (281) Browning, N. J.; Ramakrishnan, R.; von Lilienfeld, O. A.; Roethlisberger, U. Genetic Optimization of Training Sets for Improved Machine Learning Models of Molecular Properties. *J. Phys. Chem. Lett.* **2017**, *8*, 1351.
- (282) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.
- (283) Wang, T.; Zhu, J.-Y.; Torralba, A.; Efros, A. A. Dataset Distillation. *arXiv* **2020**, arXiv:1811.10959.
- (284) Sucholutsky, I.; Schonlau, M. 'Less Than One'-Shot Learning: Learning N Classes From M < N Samples. *arXiv* **2020**, arXiv:2009.08449
- (285) Cerqueira, T. F.; Sarmiento-Pérez, R.; Amsler, M.; Nogueira, F.; Botti, S.; Marques, M. A. Materials design on-the-fly. *J. Chem. Theory Comput.* **2015**, *11*, 3955–3960.
- (286) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.
- (287) Vilhelmsen, L. B.; Hammer, B. Systematic Study of Au<sub>6</sub> to Au<sub>12</sub> Gold Clusters on MgO(100) F Centers Using Density-Functional Theory. *Phys. Rev. Lett.* **2012**, *108*, 126101.
- (288) Janet, J. P.; Kulik, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.* **2017**, *8*, 5137–5152.
- (289) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.
- (290) Simm, G. N.; Reiher, M. Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes. *J. Chem. Theory Comput.* **2018**, *14*, 5238–5248.
- (291) Proppe, J.; Gugler, S.; Reiher, M. Gaussian Process-Based Refinement of Dispersion Corrections. *J. Chem. Theory Comput.* **2019**, *15*, 6046–6060.
- (292) Goreinov, S. A.; Oseledets, I. V.; Savostyanov, D. V.; Tyrtyshnikov, E. E.; Zamarashkin, N. L. *Matrix Methods: Theory, Algorithms and Applications: Dedicated to the Memory of Gene Golub*; World Scientific, 2010; pp 247–256.
- (293) Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning Sydney, NSW, Australia, 2017*; Vol. 70, pp 1321–1330.
- (294) Skafté, N.; Jørgensen, M.; Hauberg, S. Reliable training and estimation of variance networks. *Adv. Neur. Inform. Process. Syst.* **2019**, 6326–6336.
- (295) Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978–10985.



- (296) Cortés-Ciriano, I.; Bender, A. Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2019**, *59*, 1269–1281.
- (297) Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 906–915.
- (298) Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 2016; pp 1050–1059.
- (299) Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* **2019**, *10*, 7913–7922.
- (300) Prodan, E.; Kohn, W. Nearsightedness of electronic matter. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 11635–11638.
- (301) Fias, S.; Heidar-Zadeh, F.; Geerlings, P.; Ayers, P. W. Chemical transferability of functional groups follows from the nearsightedness of electronic matter. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 11633–11638.
- (302) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
- (303) Misra, M.; Andrienko, D.; Baumeier, B.; Faulon, J.-L.; von Lilienfeld, O. A. Toward Quantitative Structure-Property Relationships for Charge Transfer Rates of Polycyclic Aromatic Hydrocarbons. *J. Chem. Theory Comput.* **2011**, *7*, 2549.
- (304) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
- (305) Heinen, S.; Schwilk, M.; von Rudorff, G. F.; von Lilienfeld, O. A. Machine learning the computational cost of quantum chemistry. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025002.
- (306) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693 PMID: 31042390.
- (307) McDonagh, J. L.; Silva, A. F.; Vincent, M. A.; Popelier, P. L. Machine learning of dynamic electron correlation energies from topological atoms. *J. Chem. Theory Comput.* **2018**, *14*, 216–224.
- (308) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
- (309) Chen, X.; Jørgensen, M. S.; Li, J.; Hammer, B. Atomic energies from a convolutional neural network. *J. Chem. Theory Comput.* **2018**, *14*, 3933.
- (310) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3401–3406.
- (311) Grisafi, A.; Nigam, J.; Ceriotti, M. Multi-scale approach for the prediction of atomic scale properties. *arXiv* **2021**, *12*, 2078.
- (312) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **2018**, *9*, 4501.
- (313) Engel, E. A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M. A Bayesian approach to NMR crystal structure determination. *Phys. Chem. Phys.* **2019**, *21*, 23385–23400.
- (314) Li, J.; Bennett, K. C.; Liu, Y.; Martin, M. V.; Head-Gordon, T. Accurate prediction of chemical shifts for aqueous protein structure on “Real World” data. *Chem. Sci.* **2020**, *11*, 3180–3191.
- (315) Navarro-Vázquez, A. A DFT/machine-learning hybrid method for the prediction of <sup>3</sup>JHCCH couplings. *Magn. Reson. Chem.* **2021**, *59*, 414.
- (316) Bratholm, L. A.; Gerrard, W.; Anderson, B.; Bai, S.; Choi, S.; Dang, L.; Hanchar, P.; Howard, A.; Huard, G.; Kim, S., et al. A community-powered search of machine learning strategy space to find NMR property prediction models. *arXiv* **2020**, arXiv:2008.05994
- (317) Gupta, A.; Chakraborty, S.; Ramakrishnan, R. Revving up <sup>13</sup>C NMR shielding predictions across chemical space: Benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 035010.
- (318) Dral, P. O.; Owens, A.; Yurchenko, S. N.; Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.* **2017**, *146*, 244108.
- (319) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924–6935.
- (320) Lopez-Bezanilla, A.; von Lilienfeld, O. A. Modeling electronic quantum transport with machine learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 235411.
- (321) Arsenault, L.-F.; Lopez-Bezanilla, A.; von Lilienfeld, O. A.; Millis, A. J. Machine learning for Many-Body Physics: the case of the Anderson impurity model. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *90*, 155136.
- (322) Arsenault, L.-F.; von Lilienfeld, O. A.; Millis, A. J. Machine learning for many-body physics: efficient solution of dynamical mean-field theory. *arXiv* **2015**, <http://arxiv.org/abs/1506.08858>.
- (323) Dral, P. O.; Barbatti, M.; Thiel, W. Nonadiabatic excited-state dynamics with machine learning. *J. Phys. Chem. Lett.* **2018**, *9*, 5660–5663.
- (324) Westermayr, J.; Faber, F. A.; Christensen, A. S.; von Lilienfeld, O. A.; Marquetand, P. Neural networks and kernel ridge regression for excited states dynamics of CH<sub>2</sub>NH: From single-state to multi-state representations and multi-property machine learning models. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025009.
- (325) Westermayr, J.; Gastegger, M.; Marquetand, P. Combining SchNet and SHARC: The SchNarc machine learning approach for excited-state dynamics. *J. Phys. Chem. Lett.* **2020**, *11*, 3828–3834.
- (326) Richter, M.; Marquetand, P.; González-Vázquez, J.; Sola, I.; González, L. SHARC: ab initio molecular dynamics with surface hopping in the adiabatic representation including arbitrary couplings. *J. Chem. Theory Comput.* **2011**, *7*, 1253–1258.
- (327) Westermayr, J.; Marquetand, P. Deep learning for UV absorption spectra with SchNarc: First steps toward transferability in chemical compound space. *J. Chem. Phys.* **2020**, *153*, 154112.
- (328) Westermayr, J.; Marquetand, P. Machine learning for electronically excited states of molecules. *Chem. Rev.* **2020**, DOI: 10.1021/acs.chemrev.0c00749
- (329) Westermayr, J.; Marquetand, P. Machine learning for nonadiabatic molecular dynamics. *Machine Learning in Chemistry* **2020**, *17*, 76.
- (330) Westermayr, J.; Marquetand, P. Machine learning and excited-state molecular dynamics. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 043001.
- (331) Zubatyuk, R.; Smith, J.; Nebgen, B. T.; Tretiak, S.; Isayev, O. Teaching a neural network to attach and detach electrons from molecules. *ChemRxiv* **2020**, DOI: 10.26434/chemrxiv.12725276.v2
- (332) Zhang, Y.; Ye, S.; Zhang, J.; Hu, C.; Jiang, J.; Jiang, B. Efficient and Accurate Simulations of Vibrational and Electronic Spectra with Symmetry-Preserving Neural Network Models for Tensorial Properties. *J. Phys. Chem. B* **2020**, *124*, 7284–7290.
- (333) Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *96*, 024104.
- (334) Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, *8*, 2689.
- (335) Janet, J. P.; Kulik, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.* **2017**, *8*, 5137.
- (336) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, 036002.
- (337) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

- (338) Rossi, K.; Juraskova, V.; Wischert, R.; Garel, L.; Corminboeuf, C.; Ceriotti, M. Simulating Solvation and Acidity in Complex Mixtures with First-Principles Accuracy: The Case of CH<sub>3</sub>SO<sub>3</sub>H and H<sub>2</sub>O<sub>2</sub> in Phenol. *J. Chem. Theory Comput.* **2020**, *16*, 5139.
- (339) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **2020**, *11*, 4068.
- (340) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120*, 8066.
- (341) Ulissi, Z. W.; Singh, A. R.; Tsai, C.; Nørskov, J. K. Automated Discovery and Construction of Surface Phase Diagrams Using Machine Learning. *J. Phys. Chem. Lett.* **2016**, *7*, 3931–3935.
- (342) Ulissi, Z. W.; Tang, M. T.; Xiao, J.; Liu, X.; Torelli, D. A.; Karamad, M.; Cummins, K.; Hahn, C.; Lewis, N. S.; Jaramillo, T. F.; et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO<sub>2</sub> reduction. *ACS Catal.* **2017**, *7*, 6600–6608.
- (343) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **2017**, *8*, 14621.
- (344) Singh, A. R.; Rohr, B. A.; Gauthier, J. A.; Nørskov, J. K. Predicting chemical reaction barriers with a machine learning model. *Catal. Lett.* **2019**, *149*, 2347–2354.
- (345) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem. Sci.* **2018**, *9*, 7069–7077.
- (346) Heinen, S.; von Rudorff, G. F.; von Lilienfeld, O. A. Quantum based machine learning of competing chemical reaction profiles. *arXiv* **2020**, arXiv:2009.13429
- (347) von Rudorff, G. F.; Heinen, S. N.; Bragato, M.; von Lilienfeld, O. A. Thousands of reactants and transition states for competing E<sub>2</sub> and S<sub>2</sub> reactions. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026.
- (348) Flores, R. A.; Paolucci, C.; Winther, K. T.; Jain, A.; Torres, J. A. G.; Aykol, M.; Montoya, J.; Nørskov, J. K.; Bajdich, M.; Bligaard, T. Active Learning Accelerated Discovery of Stable Iridium Oxide Polymorphs for the Oxygen Evolution Reaction. *Chem. Mater.* **2020**, *32*, 5854–5863.
- (349) Mamun, O.; Winther, K. T.; Boes, J. R.; Bligaard, T. A Bayesian framework for adsorption energy prediction on bimetallic alloy catalysts. *NPJ Comput. Mater.* **2020**, *6*, 177.
- (350) Garjito del Río, E.; Kaappa, S.; Garrido Torres, J. A.; Bligaard, T.; Jacobsen, K. W. Machine Learning with bond information for local structure optimizations in surface science. *J. Chem. Phys.* **2020**, *153*, 234116.
- (351) Groenenboom, M. C.; Anderson, R. M.; Wollmershauser, J. A.; Horton, D. J.; Policastro, S. A.; Keith, J. A. Combined Neural Network Potential and Density Functional Theory Study of TiAl<sub>2</sub>O<sub>5</sub> Surface Morphology and Oxygen Reduction Reaction Overpotentials. *J. Phys. Chem. C* **2020**, *124*, 15171–15179.
- (352) Schlexer Lamoureux, P.; Winther, K. T.; Garrido Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine learning for computational heterogeneous catalysis. *ChemCatChem* **2019**, *11*, 3581.
- (353) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.
- (354) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.
- (355) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.
- (356) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504–1508\_eprint.
- (357) Delarue Bizzini, L.; Müntener, T.; Häussinger, D.; Neuburger, M.; Mayor, M. Synthesis of trinorbornane. *Chem. Commun.* **2017**, *53*, 11399–11402.
- (358) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.
- (359) Hoja, J.; Medrano Sandonas, L.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio, R. A.; Tkatchenko, A. QM7-X: A comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **2021**, *8*, 43.
- (360) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732.
- (361) Narayanan, B.; Redfern, P. C.; Assary, R. S.; Curtiss, L. A. Accurate quantum chemical energies for 133000 organic molecules. *Chem. Sci.* **2019**, *10*, 7449–7455.
- (362) Chen, G.; Chen, P.; Hsieh, C.-Y.; Lee, C.-K.; Liao, B.; Liao, R.; Liu, W.; Qiu, J.; Sun, Q.; Tang, J.; Zemel, R.; Zhang, S. Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models. *arXiv* **2019**, arXiv:1906.09427
- (363) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **2015**, *143*, 084111.
- (364) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- (365) Huang, B.; von Lilienfeld, O. A. Dictionary of 140k GDB and ZINC derived AMONs. *arXiv* **2020**, arXiv:2008.05260
- (366) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, 170193.
- (367) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 134.
- (368) Tkatchenko, A.; Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- (369) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, e1603015.
- (370) Dandu, N. K.; Ward, L.; Assary, R. S.; Redfern, P. C.; Narayanan, B.; Foster, I. T.; Curtiss, L. A. Quantum Chemically Informed Machine Learning: Prediction of Energies of Organic Molecules with 10 to 14 Non-Hydrogen Atoms. *J. Phys. Chem. A* **2020**, *124*, 5804.
- (371) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109.
- (372) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300–1308.
- (373) Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J. Cheminf.* **2019**, *11*, 69.
- (374) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.
- (375) Stuke, A.; Kunkel, C.; Golze, D.; Todorović, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **2020**, *7*, 58.

- (376) Schober, C.; Reuter, K.; Oberhofer, H. Virtual Screening for High Carrier Mobility in Organic Semiconductors. *J. Phys. Chem. Lett.* **2016**, *7*, 3973–3977.
- (377) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Sci. Data* **2020**, *7*, 137.
- (378) Schwilk, M.; Tahchieva, D. N.; von Lilienfeld, O. A. Large yet bounded: Spin gap ranges in carbenes. *arXiv* **2020**, arXiv:2004.10600
- (379) Heinen, S.; Schwilk, M.; von Rudorff, G. F.; von Lilienfeld, O. A. Machine learning the computational cost of quantum chemistry. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025002.
- (380) Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K.; Bergdorf, M.; Klepeis, J. L.; Shaw, D. E. Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Sci. Data* **2021**, *8*, 55.
- (381) Korth, M.; Grimme, S. Mindless<sup>™</sup> DFT Benchmarking. *J. Chem. Theory Comput.* **2009**, *5*, 993–1003 PMID: 26609608.
- (382) Balcells, D.; Skjelstad, B. B. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135.
- (383) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513–524.
- (384) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter* **2021**, *4*, 1578.
- (385) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
- (386) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939–8954.
- (387) Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **2012**, *58*, 218–226.
- (388) Saal, J. E.; Kirklín, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **2013**, *65*, 1501–1509.
- (389) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W. The Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **2021**, *11*, 6059.
- (390) Talirz, L.; Kumbhar, S.; Passaro, E.; Yakutovich, A. V.; Granata, V.; Gargiulo, F.; Borelli, M.; Uhrin, M.; Huber, S. P.; Zoupanos, S.; et al. Materials Cloud, a platform for open computational science. *Sci. Data* **2020**, *7*, 299.
- (391) Jinnouchi, R.; Miwa, K.; Karsai, F.; Kresse, G.; Asahi, R. On-the-Fly Active Learning of Interatomic Potentials for Large-Scale Atomistic Simulations. *J. Phys. Chem. Lett.* **2020**, *11*, 6946–6955.
- (392) Garijo del Río, E.; Mortensen, J. J.; Jacobsen, K. W. Local Bayesian optimizer for atomic structures. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *100*, 104103.
- (393) Thompson, A.; Swiler, L.; Trott, C.; Foiles, S.; Tucker, G. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2015**, *285*, 316–330.
- (394) Christensen, A. S.; Faber, F. A.; Huang, B.; Bratholm, L. A.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. QML: A Python Toolkit for Quantum Machine Learning, 2017; <https://github.com/qmlcode/qml>.
- (395) Huang, B.; von Lilienfeld, O. A. AQML: Amons-based Quantum Machine Learning Code for Quantum Chemistry. 2020; <https://github.com/binghuang2018/aqml>.
- (396) Bonomi, M. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16* (8), 670–673.
- (397) Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **2018**, *9*, 2261–2269.
- (398) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J.; Roitberg, A. TorchANI: A Free and Open Source PyTorch Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2020**, *60*, 3408.
- (399) Huber, S. P.; Zoupanos, S.; Uhrin, M.; Talirz, L.; Kahle, L.; Hauselmann, R.; Gresch, D.; Müller, T.; Yakutovich, A. V.; Andersen, C. W.; et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* **2020**, *7* (1), 300.
- (400) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.
- (401) Carrasquilla, J.; Torlai, G.; Melko, R. G.; Aolita, L. Reconstructing quantum states with generative models. *Nat. Mach. Intell.* **2019**, *1*, 155–161.
- (402) Nesterov, V.; Wieser, M.; Roth, V. 3DMolNet: A Generative Network for Molecular Structures. *arXiv* **2020**, arXiv:2010.06477
- (403) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604.
- (404) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. Found in Translation<sup>™</sup>: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091–6098.
- (405) Nair, V. H.; Schwaller, P.; Laino, T. Data-driven Chemical Reaction Prediction and Retrosynthesis. *Chimia* **2019**, *73*, 997–1000.
- (406) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.
- (407) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **2020**, *11*, 3316–3325.
- (408) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **2020**, *11*, 4874.
- (409) Carrasquilla, J.; Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **2017**, *13*, 431–434.
- (410) Ch'Ng, K.; Carrasquilla, J.; Melko, R. G.; Khatami, E. Machine learning phases of strongly correlated fermions. *Phys. Rev. X* **2017**, *7*, 031038.
- (411) Broecker, P.; Carrasquilla, J.; Melko, R. G.; Trebst, S. Machine learning quantum phases of matter beyond the fermion sign problem. *Sci. Rep.* **2017**, *7*, 8823.
- (412) Vargas-Hernández, R. A.; Sous, J.; Berciu, M.; Krems, R. V. Extrapolating quantum observables with machine learning: inferring multiple phase transitions from properties of a single phase. *Phys. Rev. Lett.* **2018**, *121*, 255702.
- (413) Cheng, B.; Engel, E. A.; Behler, J.; Dellago, C.; Ceriotti, M. Ab initio thermodynamics of liquid and solid water. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 1110–1115.
- (414) Cheng, B.; Mazzola, G.; Pickard, C. J.; Ceriotti, M. Evidence for supercritical behaviour of high-pressure liquid hydrogen. *Nature* **2020**, *585*, 217–220.
- (415) Carleo, G.; Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science* **2017**, *355*, 602–606.
- (416) Schütt, K.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. Unifying machine learning and quantum chemistry with



a deep neural network for molecular wavefunctions. *Nat. Commun.* **2019**, *10*, 5024.

(417) Hermann, J.; Schätzle, Z.; Noé, F. Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.* **2020**, *12*, 891–897.

(418) Carrasquilla, J. Machine learning for quantum matter. *Adv. Phys. X* **2020**, *5*, 1797528.

(419) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73–76.